A deep learning approach to image quality assessment

Feng, Yeli

2019

Feng, Y. (2019). A deep learning approach to image quality assessment. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/85493>

<https://doi.org/10.32657/10356/85493>

**A DEEP LEARNING APPROACH
TO IMAGE QUALITY ASSESSMENT**

YELI FENG

SCHOOL OF MECHANICAL AND AEROSPACE ENGINEERING

A thesis submitted to Nanyang Technological University in partial fulfillment of
the requirement for the degree of Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

17 March 2019

.....
Date

.....
Yeli FENG

School of Mechanical and Aerospace Engineering (MAE)

This certificate is awarded to

FENG YELI

**on the completion of Research Integrity Course Module in
Engineering and Technology Track**

18 Oct 2013
Date

Prof Chua Chee Kai
Chair, School of Mechanical and Aerospace Engineering

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

27 Mar 2019

.....
Date

A handwritten signature in black ink, appearing to read 'Cai Yiyu', is written over a light gray rectangular background.

.....
A/Prof Cai Yiyu

Authorship Attribution Statement

This thesis contains materials from four papers published or to be published in the following peer-reviewed journal(s) where I was the first author.

Chapter 5 is published as Feng, Y. and Cai, Y., 2017, August. No-reference image quality assessment through transfer learning. In *Signal and Image Processing (ICSIP), 2017 IEEE 2nd International Conference on* (pp. 90-94). IEEE.

The contributions of the co-authors are as follows:

- I prepared the manuscript drafts. A/Prof Cai revised the manuscript draft.
- I designed and developed the study. And I conducted the experiment and analysis.

Chapter 6 is under review by journal IEEE Transactions on Cybernetics, 2019.

The contributions of the co-authors are as follows:

- I prepared the manuscript drafts. A/Prof Cai revised the manuscript draft.
- I designed and developed the study. And I conducted the experiments and analysis.

Chapter 7 is published as Feng, Y., Teh, H.S. & Cai, Y. Curr Radiol Rep (2019) 7: 24. <https://doi.org/10.1007/s40134-019-0333-9>.

The contributions of the co-authors are as follows:

- I prepared the manuscript drafts. Dr. Teh Hui Seong and A/Prof Cai Yiyu revised the manuscript draft.
- I conducted the study.

Appendix C is published as Feng, Y. and Cai, Y., 2017. A gaze tracking system for children with autism spectrum disorders. In *Simulation and Serious Games for Education* (pp. 137-145). Springer, Singapore.

The contributions of the co-authors are as follows:

- A/Prof Cai provided the initial project direction.
- I prepared the manuscript drafts. A/Prof Cai revised the manuscript draft.

- I designed and developed the study.

21 March 2019

.....
Date

.....
Yeli FENG

Acknowledgement

I want to express sincere gratitude to my supervisor A/Prof Cai Yiyu for providing the opportunity for me to embark on this journey and his patience throughout the way.

I want to thank my thesis advisory committee for spending time providing comments and suggestions and helping me to improve this thesis.

My special thanks to friends for inspiration and encouragement that helped me stay strong.

Abstract

Deep learning, more specifically deep convolutional neural network, has an unchallengeable performance advantage over traditional machine learning in solving visual problems. The latter relies heavily on feature extractors and descriptors that need to be carefully designed by human experts. In deep learning, the feature crafting process is replaced by automatically learning from input signals. Supported by affordable computing hardware, the performance of many existing computer vision systems is boosted through deep learning, resulting in the creation of many novel applications. However, like all networked digital systems, the deep convolutional neural networks are vulnerable when the input signals are distorted or manipulated maliciously. In this thesis, we explore the computational modeling of image quality assessment, investigate the vulnerability of deep convolutional neural networks to input manipulation, and utilize image quality assessment methods to mitigate the vulnerability of applications based on deep learning.

Firstly, the image quality assessment problem and methodologies are reviewed, unsolved research problems are identified, and the latest development in deep learning in network architectures and training techniques is studied.

Secondly, within the existing image quality assessment framework, we try to answer questions like, whether the model performance could be pushed up on par with the deep learning approach. Multiple image quality assessment models have been designed, and new quality awareness feature extraction and description methods are proposed. The feasibility and advantage of applying deep learning to image quality assessment are investigated. Our experiments show that the proposed no-reference quality assessment method based on the convolutional neural network outperforms existing counterparts.

Thirdly, a novel no-reference image quality assessment framework is proposed. Using a transfer learning method, off-the-shelf visual features are re-purposed for quality prediction. These features are learned by deep convolutional network model trained with over a million images for the object classification task. The recurrent neural network is introduced to distill global features. Experiments show that this method outperforms our previous proposals.

Fourthly, we investigate the visual characteristics of maliciously manipulated images (i.e. adversarial examples) and experimentally show the correlation between these manipulations and the quality degradation in the traditional image quality assessment field. Visual quality manipulation is relatively cheap to implement hence could easily be exploited to trigger a malfunction in machine vision systems. Treating adversarial examples as a type of quality problem, a non-intrusive detection framework is proposed to protect deep convolutional networks from benign quality degradation and malicious attack. Experimentally, we show that by combining image quality assessment and knowledge distillation technique, the proposed detection framework can effectively increase the resilience of deep convolutional networks when deployed in the real world environment.

Lastly, in this thesis, the very recent development of deep learning applications in chest radiology is studied and shared in the radiology practice community. Similar to the deep learning applications in other domains, poor quality radiographs and computed tomographic images could lead to false diagnosis by image analysis systems. The image quality in chest radiology includes many more aspects beyond the pixel level for diagnosis. These include the projection, necessary anatomy inclusion, and x-ray penetration. From the perspective of AI for diagnosis, a method for detecting projection

view of chest radiographs is designed and trained. When preparing a large-scale dataset for training a disease detection model, this model helps to quickly identify outliers. This work builds a pathway towards the author's wish to expand future work into radiology imaging analysis.

Table of Contents

Acknowledgement	i
Abstract.....	ii
Table of Contents	v
List of Figures.....	ix
List of Tables	xi
List of Abbreviations	xii
Chapter 1 Introduction.....	1
1.1 Background and Motivation	1
1.2 Objective and Scope of this Work	3
1.3 Organisation of the Thesis	4
Chapter 2 Literature Review	6
2.1 Background	6
2.2 Full-reference Image Quality Assessment	10
2.3 No-Reference Image Quality Assessment	14
2.4 Summary	18
Chapter 3 Deep Learning Approach.....	19
3.1 Background	19
3.2 Learning Methods.....	21
3.2.1 Backpropagation	21
3.2.2 Convolutional Neural Network.....	23
3.2.3 Recurrent Neural Network.....	29
3.2.4 Autoencoders	31
3.3 Model Training	34
3.3.1 Gradient Descent Optimization and Hyperparameters	34
3.3.2 Transfer Learning.....	36

3.4 Summary	38
Chapter 4 A Comparison Study of Image Quality Assessment	39
4.1 Full-Reference IQA	39
4.1.1 Enhancement Proposal	39
4.1.2 Experiment Setup	41
4.1.3 Performance and Analysis	42
4.2 No-Reference IQA	46
4.2.1 Enhancement Proposals	46
4.2.2 CNN-based Method	49
4.2.3 Experiment Setup	51
4.2.4 Performance and Analysis	51
4.3 Summary	57
Chapter 5 No-Reference Image Quality Assessment Through Transfer Learning	58
5.1 Background	58
5.2 Related Work	60
5.3 A Transfer Learning Design	62
5.3.1 Network Architecture	63
5.3.2 Heterogeneous Global Distortion	64
5.3.3 Learning Method	65
5.4 Experiments And Analysis	66
5.4.1 Data	66
5.4.2 Evaluation	67
5.4.3 Cross Dataset Evaluation	69
5.4.4 Computational Cost	71
5.5 Summary	71
Chapter 6 Towards Resilient Classification Through Image Quality Assessment	72

6.1 Background	72
6.2 Related Work	76
6.3 Image Quality And Classification Performance	79
6.3.1 Methodology	80
6.3.2 Image Distortions.....	83
6.3.3 Adversarial Examples	84
6.4 Fit-For-Task Approach.....	86
6.5 Experiments And Analysis.....	88
6.5.1 Data	88
6.5.2 Evaluation	90
6.5.3 Baselines	91
6.5.4 Performance of Fit-for-Task Detector	92
6.6 Comparison and Discussion.....	95
6.7 Summary	96
Chapter 7 Applications and Image Quality in Chest Radiology	98
7.1 Deep Learning in Chest Radiology	99
7.1.1 Background	99
7.1.2 Chest Radiography.....	104
7.1.3 Chest CT	109
7.1.4 Other Applications	112
7.2 Image Quality in Chest Radiography	114
7.2.1 A Projection View Classifier for Chest Radiographs	115
7.2.2 Discussion	117
7.3 Summary	118
Chapter 8 Conclusion	119
8.1 Contributions	119
8.2 Limitation and Future Work.....	121
8.2.1 Improving Fit-For-Task Preventive Framework.....	121
8.2.2 Radiological Imaging Quality Management.....	122

References	124
Appendices	140
A. Connection Between Linear Autoencoder and PCA	140
B. Distortions Types and Semantic Similarity	142
C. A Gaze Intervention Framework For Children With Autism.....	144

List of Figures

Fig 1.1	The Major Contents and Organization of the Thesis	4
Fig 2.1	A Schematic Diagram of Objective Image Quality Assessment Approaches.....	7
Fig 2.2	Examples of Distorted Image from Image Quality Datasets.....	9
Fig 2.3	An Example of Visual Codewords [71] Constructed From CSIQ Dataset	16
Fig 3.1	Forward and Backward Propagation in Multilayer Neural Networks.....	22
Fig 3.2	Architecture of a Typical CNN	23
Fig 3.3	Filters and Activations Inside a CNN.....	26
Fig 3.4	Image Reconstruction From Activation Layer.....	26
Fig 3.5	A Schematic Diagram of the Inception Module [18]	27
Fig 3.6	A Schematic Diagram of Residual Learning Module [3].....	28
Fig 3.7	A Schematic Diagram of Recurrent Neural Network [144].....	30
Fig 3.8	A Schematic Diagram of a Typical LSTM Block [34]	31
Fig 4.1	Performance Comparison of Full-reference IQA Methods	44
Fig 4.2	Performance in Median SROCC of Cross Database Generalization Tests	44
Fig 4.3	Computational Costs of the Proposed Full-reference IQA Models	45
Fig 4.4	Image and Histogram of Corresponding Signals after MSCN Process.....	47
Fig 4.5	A Block Diagram of Feature Design Schema	48
Fig 5.1	A schematic diagram of the architecture of DTNIQ-f IQA model	63
Fig 5.2	Examples of Heterogeneous Global Distortion.....	64
Fig 5.3	A Semantic Diagram of DTNIQ-r IQA Model	65
Fig 6.1	Sample Images That Fail to Fool Proposed Resilient Image Classification	75
Fig 6.2	Correlation Between Image Distortion and Semantic Dissimilarity	84
Fig 6.3	Fit-For-Task Approaches - Analysis	88
Fig 6.4	Predicted Image Quality of Adversarial and Distorted Images.....	90
Fig 6.5	Adversarial Examples.....	94
Fig 6.6	Discussion on Perceptual Quality and Object Recognition.....	96
Fig 7.1	A Paradigm of Multi-labels Abnormality Detection with Whole Radiograph	102
Fig 7.2	An Example of Edge Detection in a Lung CT Slice	103
Fig 7.3	Descriptive Paragraphs Generation with CNN+LSTM Framework	103

Fig 7.4	Detection and Localization of Common Thorax Diseases.....	106
Fig 7.5	Towards Automatic Text Report Generation	109
Fig 7.6	3D U-Net and Segmentation Task.....	111
Fig 7.7	Other Applications with Lung CT images	112
Fig 7.8	The Concept of Dilated Convolutions.....	113
Fig 7.9	Common Projection and View of Chest Radiography	115
Fig 7.10	Examples of frontal Views Miss-classified as Lateral View	117

List of Tables

Table 2.1	Overview of Datasets in the IQA literature.....	8
Table 4.1	Design Specification of Full-reference IQA Proposals.....	40
Table 4.2	Performance of Proposed Full-reference IQA Models on Single Dataset	43
Table 4.3	Performance Comparison of Full-reference IQA Methods.....	43
Table 4.4	Design Specification of No-reference IQA Enhancement Proposals	49
Table 4.5	Performance of Cross Datasets Generalization.....	53
Table 4.6	Performance of Deep Learning-based No-Reference IQA Methods	54
Table 4.7	Computational Cost of Training NRCNN Models.....	56
Table 4.8	Testing Time of Learning-based No-Reference IQA Models.....	57
Table 5.1	Performances On Single Dataset.....	68
Table 5.2	Cross Dataset Performance	70
Table 6.1	Performances of Networks Over IQA Datasets	82
Table 6.2	Visual Quality of Adversarial and Original Images.....	85
Table 6.3	Classification Confidence Score	87
Table 6.4	Baselines of Fit-for-Task Detection	91
Table 6.5	Performance of Detection Classifier	93

List of Abbreviations

IQA	Image Quality Assessment
HVS	Human Visual System
SROCC	Spearman Rank Order Correlation Coefficient
SVM	Support Vector Machine
SVR	Support Vector Regression
RBF	Radial Basis Function
CIELAB	Commission Internationale de l'Eclairage L*a*b* color space
YIQ	Luminance (Y), In-phase Quadrature (NTSC) color space
MSCN	Mean Subtracted Contrast Normalization
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GRN	General Regression Neural Network
SGD	Stochastic Gradient Descent
MIQA	Medical Image Quality Assessment

Chapter 1

Introduction

1.1 Background and Motivation

Machine learning is ubiquitous in computational modeling, especially in solving visual problems where statistical information is extracted from patterns and used to detect, classify, or predict particular feature configurations in digital images. Since the perceptron algorithm was invented sixty years ago [176] to recognize simple shapes from pictures, the machine learning field has evolved continuously with the discovery of new algorithms and the expansion of applications. Neocognition [10], a multilayered artificial neural network created in the 1980s inspired the convolutional neural networks. Today the latter has become the default building block for many computer vision applications, such as object classification, face recognition, medical imaging diagnosis, and many more. In parallel, the support vector machine [177] invented in the 1990s is still widely used today to solve classification and regression problems when the dimension of the input signal is moderate, for example, lesion characteristic analysis in the radiomics field, composition analysis in biological and material sciences, signal quality assessment in image processing.

In the last decade, ImageNet [12], graphics processing units and deep convolutional neural networks catalyze a new machine learning boom. A new subset of machine learning, deep learning, is established in academia and industry. With deep learning, unprecedented prediction power can be achieved in traditional research problems, and

commercialization becomes possible. Examples include DeepFace [97] from Facebook, object recognition from YouTube video, lung nodule characterization from computed tomography scans. More challenging tasks can be tackled, including semantic segmentation for autonomous driving, image description, multiple pathologies detection from radiography and more.

Digital images could be distorted during the process of capture, storage, and transmission. Quality degradation could affect the viewing experience of human users as well as the function of machine learning systems. Studying and modeling the impact of signal distortion on the human visual system is coined as image quality assessment. It has been a topic of research for several decades. Many feature extraction and description techniques and algorithms have been proposed in the literature. They have been proven useful in the multimedia era, including applications in image restoration, denoising, compression, image synthesis, digital watermarking to name a few.

The deep learning era creates a new class of users machine vision systems. One of the challenges in building such systems is that it is impossible to train the system with the entire data population of the past, present and future. When deployed into the real world, the performance of these systems will be affected by various kinds and magnitudes of degradations in input images. Robots and autonomous vehicles could act detrimentally in the real world if they made decision based on poor quality images.

In the meantime, studies [116-127] discover that deep neural networks can give surprisingly egregious outputs given artificially crafted fooling images or images injected with malicious changes that are hardly perceivable by the human visual system. Such maliciously manipulated images are called adversarial examples. Active research around

adversarial examples is ongoing in recent years, for the purpose of developing robust and resilient machine vision systems.

Today, artificial intelligence dominates the news headlines, promises enormous potential benefits to businesses and society. When artificial intelligence solutions are massively deployed into the real world, robust and safe would be one critical factor, among others. Images quality and related issues could impact the performance of deep learning based solutions, in a negative or even harmful way. Therefore I was motivated to dedicate this thesis to studying image quality assessment in the context of deep learning, in methodologies and applications.

1.2 Objective and Scope of this Work

The objective of this thesis is to conduct an in-depth experimental study about image quality assessment, design new computational models that would be beneficial to machine vision systems and explore their applications in the deep learning era. In particular, the following aspects are addressed in this thesis:

- How to design new feature extraction and description methods that can better characterize images for quality assessment within existing framework, whether they can perform on par with deep learning based methods.
- How to design new image quality assessment algorithms with a deep learning approach.
- How image quality and adversarial samples impact the performance of deep convolutional neural networks.

- How to utilize image quality assessment to protect convolutional neural networks from adversarial and poor quality images.
- Study the applications of deep learning in chest radiology and image quality related-issues in this domain.

1.3 Organisation of the Thesis

Fig. 1.1 outlines the major contents and organization of this thesis. The thesis is divided into eight chapters:

Fig 1.1 The Major Contents and Organization of the Thesis

Chapter 1 gives an introduction about the thesis, including the background and motivation, objective and scope, and thesis organization.

Chapter 2 reviews the core research problem, image quality assessment, major existing works in the literature related to this study, and state-of-the-art image quality assessment algorithms. Unsolved research problems are also identified.

Chapter 3 studies the deep learning techniques with a focus on influential architectures and visual applications that include convolutional neural networks, recurrent neural networks and autoencoders.

Chapter 4 investigates possible improvements on quality assessment modeling within the existing framework and the possibility of achieving better model performance with a method based on deep learning. Multiple full-reference and no-reference assessment algorithms are proposed and comparison are made through experiments.

Chapter 5 describes a transfer learning-based no-reference quality assessment framework. Based on the proposal, models are trained for experiments, performance analysis and comparison with methods in Chapter 4 and peers in the literature.

Chapter 6 presents a systematic study on how poor image quality and adversarial examples will drag down the object recognition capacities of well-known deep convolutional neural models trained on ImageNet. Subsequently, we propose a fit-for-task framework that improves the resilience of these object recognition models against poor quality images and adversarial examples. This fit-for-task framework is an application of the method proposed in Chapter 5.

Chapter 7 introduces the applications of deep learning in chest radiology and image quality-related issues. Method that can find visual outliers in large-scale datasets is proposed. Deep learning is very dependent on clean data and accurate labels. With this method, we can quickly find poor images where their projection or position is inadequate for training a disease detection algorithm.

Chapter 8 concludes the thesis with a summary of the main contributions of this thesis work, the limitation and the possible directions of further studies.

Chapter 2

Literature Review

Computationally assessed image quality is an important component behind various digital imaging applications including encoding and compression standards, video broadcasting and computer vision. We can find many different methods have been proposed in the past. In the literature, Wang et al. [139] categorize these methods into: full-reference methods where the quality of a test images is assessed by comparing it with a reference image; reduced-reference methods where the quality of a test image is assessed by comparing the features extracted from both test and reference images; and no-reference methods where the assessment is carried out without any reference. A reference image is assumed to have perfect quality to human viewers.

This chapter introduces the notion of image quality assessment, research methods and resource and terminologies commonly used for analysis and comparison studies. Subsequently, existing full-reference and no-reference methods that are influential or closely related to our work are reviewed. It doesn't mean to be exhaustive but it provides a background for chapters 4, 5 and 6.

2.1 Background

In the field of image and video applications, visual signals are subject to many kinds of distortions before being presented to human viewers. The distortions could be introduced during signal transforming, compression, transmitting, restoring, and other image processing. Today more and more digital images and videos are being distributed and

consumed over heterogeneous networks and devices. Methods that gauge perceptual quality play an increasingly important role in various applications, including selecting an image compression algorithm that provides the best perceptual quality for a given network bandwidth criteria or database storage quota and selecting optimal model parameters when designing an image de-noising or super-resolution algorithm.

The method that determines the quality of a digital image is termed as image quality assessment (IQA). In the history of IQA research, there are subjective and objective methods. Subjective methods rely on the perceptual assessment from human viewers. These require tremendous human resource and time. This has driven the search for more efficient objective methods together with the advancement of image processing and computational modeling techniques. Today the outputs of subjective quality assessment provide ground truth to evaluate objective methods. Objective methods are based on computational models that can predict perceptual quality. There are many approaches to designing objective methods and a low-resolution overview of it is outlined in Fig. 2.1.

Fig 2.1 A Schematic Diagram of Objective Image Quality Assessment Approaches

The blue lines depict pipelines of learning-based no-reference methods. Solid black lines represent the process of none learning-based no-reference method. And dashed lines describe steps that are unique to full-reference and reduced reference methods.

Most learning-based objective methods use quality assessment results from subject studies, including the methods proposed in this thesis. Since the subjective assessment results serve as ground truth, the rest of this section introduces concepts and terminologies that will be used extensively in the analysis, discussion and comparison in future chapters.

In subjective quality studies, human observers estimate the quality of distorted images under a controlled environment and then give an opinion score. For example, in [49] a pair of distorted and reference images are simultaneously presented to a human observer. These scores are then averaged into one mean opinion score (MOS) for each image, which serves as a ground truth to evaluate various objective methods proposed in the literature. Several research teams [81,83,82,100] made their experimental results available to the public. Table 2.1 gives an overview of the most widely used datasets in the literature to evaluate computational IQA models.

Table 2.1 Overview of Datasets in the IQA literature

Database	Available Since	Reference Images	Distorted Images	Distortion Types	Human Observers
TID2013 [82]	2013	25	3000	24	971
CSIQ [83]	2010	30	866	6	35
TID2008 [49]	2008	25	1700	17	838
LIVE [81]	2005	29	779	5	161
LIVE in the Wild [100]	2015	NA	1162	NA	8100

Among them, the four datasets categorize distorted images by types and use software programs to insert distortions into reference images. The mostly studied distortion types include Gaussian blur, white noise, compression error, and others. A complete list of distortion types can be found in Appendix B. Perceptual quality of images in the LIVE in the Wild are naturally degraded while snapped from smartphone or camera, and no distortion type is assigned. For a quick visual understanding, a few examples of distorted

images and their corresponding quality scores from these datasets are illustrated in Fig. 2.2.

Fig 2.2 Examples of Distorted Image from Image Quality Datasets

D in DMOS stands for the quality difference of a distorted image from its corresponding reference. The range of DMOS is 0 (high visual quality) to 100 in LIVE and 0 to 1 in CSIQ. The range of MOS is 0 (poor visual quality) to 9 in TID2013. The range of MOS is 0 to 100 in LIVE in the Wild.

To analyze various objective methods and how accurate their predictions are, performance measurement metrics that are mostly adopted in the literature include root mean squared error (RMSE) which measures the difference between the predicted quality value and the MOS score. The Pearson linear correlation coefficient (PLCC) evaluates the linear relationship between two continuous variables, for example, between the MOS values from human observers and predictions from an objective method. The relationship between MOS and prediction is not necessarily linear but needs to be monotonic for the quality scores predicted to be useful. Therefore, Spearman rank order correlation coefficient (SROCC) [185] is more frequently used to compare the performance of objective methods.

2.2 Full-reference Image Quality Assessment

For over a decade, many full-reference IQA methods have been proposed; a significant number of these methods focused on computationally modeling a deterministic mapping function between the image signal and its perceptual quality. Other methods [51, 59, 67-69, 71-72, 84, 101] apply support vector machines to learn such mapping functions. Based on the understanding of the human visual system (HVS) and image signal processing techniques, both quality assessment metrics and models seek features that hopefully capture some characteristics of perceived visual distortion well.

The HVS is a complex system yet to be fully understood. Extensive research show that the visual information distributed in a scene is not equally and simultaneously perceived. For example, the triadic architecture [52] for understanding scene consists of three sub-systems: a volatile low-level vision where ‘quick and clean’ measurements of pixels and edges are carried out; a nonvolatile nexus system where attention on object is formed and

released by grasping a small number of proto-objects from the low-level; and a non-attentional setting system of scene gist and spatial arrangement which guides eye movements and attentional shifts.

Nevertheless, knowledge about HVS inspires researchers to seek quality-aware features from potentially useful perspectives. We review a set of assessment methods here based on uniqueness, influence or relevance to our work.

Given a pair of reference and distorted images of m -by- n pixels, peak signal to noise (PSNR) is computed over the pixel intensity using (2.1), where MAX is the maximum intensity value. Today PSNR serves as a baseline quality metric.

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (2.1)$$

$$MSE = \frac{1}{m \cdot n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2$$

SSIM is a structural similarity quality metric proposed by Wang et al. [50]. Emphasizing the hypothesis that the HVS is highly adapted for extracting structural information from a scene, SSIM models the perceptual quality as a function of perceived changes of structural information in the spatial domain, (see 2.2) where x and y are reference and distorted images, μ is mean, σ is standard deviation and C is an arbitrary constant. Wang et al. [54] later enhance the metric performance by pooling quality scores from multiple scales of an image.

$$SSIM(x, y) = \frac{(2 \mu_x \mu_y + C_1) (2 \sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1) (\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2.2)$$

Based on the assumption that the random fields from subbands of a distorted image can be formulated as a function of random fields extracted from the corresponding reference

image, Sheikh et al. [55] propose VIF where statistics of distortion between a reference and distorted images are modeled in an information fidelity setting.

VSNR is a wavelet-based signal-to-noise metric proposed by Chandler et al. [56]. Firstly, a multi-level wavelet transformation is applied to pixel intensity, and then distortion contrast is measured in the frequency domain. If the distortion contrast is lower than a threshold of detectable distortion the quality score will be ∞ , otherwise another stage of computing is applied to the perceivable distortion contrast.

Arguing that low-level features play an important role in image understanding, Zhang et al. [57] propose a feature similarity index FSIM, where the phase congruency feature (PC) is computed from pixel intensity to capture local structural difference, and the gradient magnitude feature (G) is computed from pixel intensity to capture contrast variance. FSIMc extends the similarity measure to chrominance. The final quality score is pooled from feature similarity maps as shown in (2.3), where S_{pc} and S_G are similarity measures of pixel intensities between a distorted image and its reference, S_I and S_Q are similarity measures of pixels in Y and Q color channels, λ controls the importance of the chromatic, PC_m serves as a weighting function, and Ω is a set represents pixels in a distorted and reference image pair.

$$FSIM_c = \frac{\sum_{\Omega} S_{pc}(\mathbf{x}) \cdot S_G(\mathbf{x}) \cdot [S_I(\mathbf{x}) \cdot S_Q(\mathbf{x})]^\lambda \cdot PC_m(\mathbf{x})}{\sum_{\Omega} PC_m(\mathbf{x})} \quad (2.3)$$

Based on the assumption that highly visible distortion can significantly affect the saliency map of an image, Zhang et al. [58] propose VSI, where image saliency is used as one of the features as well as a weighting map to reflect the impact of a local region to the final quality score. At its time, VSI obtains the best overall performance across all datasets, compared with peers.

Narwaria et al. [59] propose a model that uses support vector regression (SVR) with radial basis function (RBF) kernel to train features extracted from singular value and vector that are computed from singular value decomposition (SVD) of 128-by-128 image blocks. Their method predicts quality scores of all blocks sampled from an image and then average them into one final quality score.

Liu et al. [60] propose a multi-metric fusion model that produces a quality score from synthesizing scores from a set of existing best-performing metrics. It is a two-stage model. Firstly, a support vector classifier discriminates distorted images into five groups. In the second stage, the image pair is evaluated by one to four existing metrics based on the output from the first stage and corresponding fusion rules. The fusion rules and parameters are learned through SVR.

Charrier et al. [61] propose another two-stage model. The features are MS-SSIM-like in spatial domain and components from steerable pyramid transformation. In the first stage, ten binary support vector classifiers discriminate an image by distortion type. In the second stage, output from the binary classifiers are deterministically mapped into five groups that serve as inputs to five SVRs. Finally, a scalar mapping function aggregates the regressors' output into one quality score.

In recent years, deep learning-based methods have been emerging. Kim et al. [137] designs a 6-layers convolutional neural network to train a predictor, DeepQA, directly from patches of distorted images and corresponding error maps. An error map is a normalized log difference between a distorted and corresponding reference image. The patches are 112x112 pixels cropped from an image using a sliding step of 80x80.

Bosse et al. [136] design a 14-layers CNN-based model, WaDIQaM-FR, to extract features from 32x32 pixel patches from a pair of distorted and reference image and learn a

regressor. Arguing that not every region in an image has an equal impact to the global image quality, a patch-based weight map is also learned to pool the quality prediction on patches into one score. Kim et al. [137] design a 6-layers CNN-based model, deepQA, to learn a quality regressor. In deepQA, reference images are used to compute error maps for each distorted image. Input of deepQA is a pair of distorted image and its corresponding error map.

2.3 No-Reference Image Quality Assessment

Most no-reference (NR) IQA metrics proposed before 2010 are distortion specific, that is they attempt to address a single or particular type of distortion such as blurring, blocking, ringing or noise. Applying fast Fourier transform (FFT) to a distorted image in horizontal and vertical directions separately and then computing the 2nd and 3rd order statistics of the FFT power spectrum, Wang et al. [62] measure the blocking distortion introduced by lossy image coding standards such as JPEG. Li et al. [63] analyze distinct mathematical properties that blur, random noise, blocking and ringing carry. Marziliano et al. [64] assess blur by measuring the spreading of the edges in the spatial domain along the vertical direction. Caviedes et al. [65] design a sharpness metric for video quality control. It combines kurtosis of non-DC components of discrete cosine transform (DCT) from 8-by-8 pixel blocks with the edge profile acutance of the same block but in the spatial domain. Zhu et al. [66] construct a signal to blur and noise indicator from pooling the singular values of the gradient matrix from 16-by-16 pixel blocks.

Research in developing non-distortion specific no-reference IQA models has been active since 2010. The general approach remains learning based but the features extraction methods are shifting to natural scene statistics [70] or unsupervised learning. In

the rest of this section, learning-based methods from 2010 to 2016 are reviewed.

Moorthy et al. [67] propose a two-step framework BIQI. The underlying features are shape and variance parameters of general Gaussian distributions (GGD). These parameters are fitted from the subband coefficients obtained by transforming the distorted image with Daubechies wavelet at three scales and in three orientations. Firstly, distortion type of an image is classified. Subsequently, a distortion type specific SVR model with RBF kernel produces a quality score. DIIVINE is also proposed by [68] Moorthy et al.. It is a framework similar to BIQI but the features are components from steerable pyramid decomposition.

[69] proposes BRISQUE, which is distinct from DIIVINE and BIQI in two aspects. Firstly, it is a one-step framework; distortion classification step is abandoned. Secondly, the features are extracted from the spatial domain; they are shape and variance parameters of GGDs and asymmetric GGDs fitted from the histogram of mean-subtracted contrast normalized (MSCN) pixel intensity. The feature selection is inspired by the statistical properties observed in natural scenes by Ruderman et al. [70]. The MSCN process is described in (2.4), where ω is a kernel used to convolve over a region of $2K$ -by- $2L$ pixels.

$$\begin{aligned}
 \bar{I}(i, j) &= \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + 1} \\
 \mu(i, j) &= \sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} I(i+k, j+l) \\
 \sigma(i, j) &= \sqrt{\sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} [I(i+k, j+l) - \mu(i, j)]^2}
 \end{aligned} \tag{2.4}$$

Ye et al. [71] apply k-mean clustering to construct a visual codebook from patches of ZCA whitened distorted images. This codebook is a sparse-coding dictionary comprising of codewords and corresponding ground truth, i.e., the MOS values. Some of the learned

visual codewords learned resemble Gabor filters as shown in Fig. 2.3. Given a test image, first the distance between its patches and the visual codewords are computed, then an occurrence distribution of the nearest codewords is obtained. Finally, a quality score is derived from the distribution using SVR with a linear kernel.

Xue et al. [72] propose BIQA. Its learning pipeline that is similar to BRISQUE. The uniqueness is that the features are marginalized joint distributions of responses from gradient magnitude and Laplacian of Gaussian operators, and joint adaptive normalized GM and LOG. The underpinnings of their feature selection are rooted in findings from Geusebroek et al. [73], Wainwright et al, [74] and Lyu et al. [75].

Fig 2.3 An Example of Visual Codewords [71] Constructed From CSIQ Dataset

The above reviewed NR-IQA models are opinion-aware, i.e., MOS is required during the model training. Next, models that are opinion-unaware are reviewed. Instead of relying on MOS, these models construct a reference to high visual quality by learning from natural images in databases used for other visual tasks, for example, object recognition. Mittal et al. [76] propose NIQE, a model that constructs a reference multivariate Gaussian (MVG) model from image patches selected from Flickr and the Berkeley image segmentation databases [178]. The distance between a test image's MVG model and the reference MVG model indicates the quality of the test image. The distance is computed using (2.5), where v_1 , v_2 and Σ_1 , Σ_2 are the mean vectors and covariance matrices of the MVG models of reference and test images. Using a maximum likelihood

estimation procedure, the MVG parameters are obtained from the features as proposed in the BRISQUE [69].

$$D(\nu_1, \nu_2, \Sigma_1, \Sigma_2) = \sqrt{\left((\nu_1 - \nu_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\nu_1 - \nu_2) \right)} \quad (2.5)$$

Zhang et al. [77] enhance NIQE by enriching the framework with more features; they are Weibull statistics from the gradient and Gaussian statistics of log-transformed chrominance in the color-opponent space (CIELAB).

Xue et al. [78] propose QAC, a quality-aware clustering model. In order to construct a quality reference, pristine images are selected from the Berkeley image segmentation database and distorted by four types in three quality levels. Then the FSIM metric is applied to pairs of original and distorted image patches to establish a mapping between a distorted image and quality score. From this mapping, a codebook is constructed in which images of the same quality level (total 10) are clustered into 30 groups based on their local structure. Provided a test image, patches are sampled and the nearest cluster centroid to each patch is found, then quality scores of the found clusters are averaged to derive a global quality score.

Ghadiyaram et al. [101] propose a two-phases framework to classify the quality of a test image into one of the five grades. Firstly, wavelet transformation is applied to images to extract magnitude and entropy from sub-bands to form 12-dimensional feature vectors. In the second phase, a deep believe network (DBN) is used to train a classifier. Finally, the output of the classification layer is pooled into a quality score using Bayes' rule.

Removing the input branch of reference image, Bosse et al. [136] re-purpose their convolutional neural network-based full-reference model to train a no-reference model.

Ma et al. [138] also propose a framework, MEON. It comprises four convolutional neural network layers and two fully connected neural layers. The uniqueness of this design is that the activations from each feature extraction layer are locally normalized before sent down to a next layer.

Kang et al. [51] proposes a dedicatedly designed 5-layer CNN architecture to learn a quality model from 32-by-32 pixel blocks sampled from distorted images. This model outperforms its predecessors on the LIVE database. Bianco et al. [102] propose DeepBIQ, a 2-phases framework that uses CNN to extract feature and then trains an SVR to obtain a quality score. Both transfer learning and fine-tuning techniques are exploited in their work.

2.4 Summary

In this chapter, we provide a background for the research problem of image quality assessment, and review some of the influential or top performing full-reference and no-reference IQA methods. There are yet unsolved problems in the IQA research.

Existing no-reference IQA methods don't perform equally well across quality distortion or degradation conditions investigated so far. Their prediction accuracy is on par with full-reference methods over simple single distortion types such as Gaussian blur and white noise, but drops significantly under certain conditions, such as multiple types of co-existing distortions and geometric changes. The non-eccentricity pattern noise distorted image shown in Fig. 2.2 is an example of geometric changes. Other unsolved problems include IQA for enhanced natural images and synthetic images, for example, those produced by digital photo editing software and virtual reality system, remains to be researched.

Chapter 3

Deep Learning Approach

Since the progress made in handwritten digit recognition by Hinton et al. [1], deep learning has dramatically improved application performance in image classification and many other applications such as speech recognition, machine translation, and drug discovery. The increasingly faster graphics processing unit (GPU) and big data empower applications that rival human intelligence, for example, AlphaGo [2] defeated the world champion of the Go game. The IQA research field is also picking up the power of deep learning. In this chapter, the origin of deep learning and key breakthrough development are provided in section 3.1. Followed by reviews on learning methods and model training techniques that are related to our works.

3.1 Background

Deep learning is a subfield of machine learning where multiple levels of representations useful for a specific task are learned. Higher-level representations (e.g., visual concepts) are learned from lower-levels (features) and so on, as a result form a hierarchy of features. Therefore a deep learning architecture contrasts a conventional machine learning architecture where there are typically one or two hidden layers. The work by He et al. [3] demonstrates it is practical to train more than a thousand layers deep network with current hardware and training techniques.

Unlike in the conventional machine learning paradigm where careful engineering is

required to design feature extractors that transform raw data into feature vectors before learning can be applied, deep learning techniques automatically learn the feature representations from raw data or with a little process. For example, mean subtraction and division is a common data pre-processing step when handling natural images.

Deep learning has its root in artificial neural networks. In 1957 Rosenblatt invented the perceptron algorithm to recognize patterns from 20-by-20 pixels images. The perceptron algorithm learns a function that maps the inputs to 0 or 1 which indicates the presence or absence of a class of pattern, i.e., a binary classifier. After the perceptron machine built it was quickly proven that the perceptron could not recognize many patterns, which stagnated the neural network research for many years.

During the 1980s research in neural network was popularized again by the work from Rumelhart et al. [6], where a new learning procedure, backpropagation, was introduced. In this learning procedure, firstly, the total error (a difference between desired output and output of the forwarding path) is computed at the output layer. Applying the chain rule, the derivative of this error is propagated back all the way to the first hidden layer. Along the backward path, the partial derivative of the error with respect to each neuron's weight is derived; this is used to optimize the weights using gradient descent until convergence. However, in practice, the backpropagation technique alone only works for networks with a small number of hidden layers due to the vanishing gradient problem found later by Hochreiter et al. [7]. The gradient of error decreases exponentially with the number of hidden layers.

Since 2006 several works that successfully tackled the difficulty of optimization revives deep neural network research. New network architectures and efficient training techniques have been developed. Hilton et al. [1] introduced a new architecture, deep

belief network, and training strategy. A deep belief network is a stack of restricted Boltzmann machines where the neuron weights are first learned layer-by-layer in an unsupervised way. This stage is called greedy layer-wise pre-training. The pre-trained weights are then used to initialize the entire deep model that is then trained end-to-end using backpropagation. This stage is called fine-tuning. Adopting this training strategy, Bengio et al. [8] show that autoencoders can also be used to construct an effective deep model. These progresses ignited rapid growth of deep learning research in subsequent years and collaboration between academia and industries, which is described as the "enlightenment" phase of deep learning [9] according to the Gartner hyper-cycle model on technology development.

3.2 Learning Methods

In this section a group of deep learning methods that are used in later chapters are investigated. They are the convolutional neural network (CNN), Recurrent neural network (RNN), and autoencoders. We start with a training algorithm that is essential to deep learning.

3.2.1 Backpropagation

Backpropagation refers to a family of algorithms that can efficiently train an artificial neural network. It evolved from a group of research in the 60s. In the 80s, LeCun proposed the modern version of backpropagation that successfully trained the CNN architecture. Today backpropagation is the default algorithm in deep learning.

Here we use a classification task to illustrate how it works. During the training phase, the machine is shown millions of data samples and the corresponding manually labeled

classes, an objective that measures the error between the machine output and the 'correct' class labels are computed, and weights of neurons in the machine are adjusted to minimize the 'error'. After hundreds to thousands of such learning iterations (see Fig. 3.1) the average 'error' of all data samples is small enough for the machine to produce sensible prediction given new samples. However, arriving at a small enough 'error' (i.e., convergence) is not promised here.

Fig 3.1 Forward and Backward Propagation in Multilayer Neural Networks

In Fig 3.1, dash lines illustrate the flows of forwarding activation and backward error propagation. σ is an activation function, w and b are weights and bias respectively. $\partial E / \partial z$ is the partial derivative of error to neurons' output before the activation in each layer. If the activation function is sigmoid (s), $\partial y / \partial z = s(z)(1 - s(z))$, where z is obtained from the forwarding path computation. Repeatedly applying the chain rule to the derivatives, error E is propagated back all the way to the first hidden layer.

If the network outputs a numerical value, it is a regression task. In the following sections, we review the convolutional neural network (CNN) and recurrent neural network (RNN). Since 2012 CNN has been the winning network architecture in the imageNet large-scale visual recognition challenges (ILSVRC) for classification tasks. RNN can handle tasks with sequential inputs, for example, sentences in natural language processing. In recent years, RNN has been actively deployed in vision applications such as image captioning task.

3.2.2 Convolutional Neural Network

Inspired by the structure of the visual system in vertebrate, in 1980, Fukushima proposed a learning architecture, neocognitron [10]. From here LeCun and other researchers later introduced gradient-based training and other techniques leading to a state-of-the-art performance in various visual applications. The literature attributes the excellent performance of CNN to its unique architecture, i.e., the local connectivity and subsampling. Fig. 3.2 shows the architecture of a typical CNN, several stacks of convolutional and pooling layers, followed by more convolutional and fully connected layers.

Fig 3.2 Architecture of a Typical CNN

In each convolution-pooling block, the convolutional computations are applied on local patches (white boxes) of input channels at different locations. The outputs are then passed to rectified linear units (ReLU) before being pooled across or within channels. The outputs from ReLU are often called activations or feature maps. Each location in a feature map is associated with a rectangle region from its previous layer through the same set of neurons. This local rectangle region and the interconnection between two layers resemble the mapping between receptive fields and complex cell in the visual system of the vertebrate. This architecture allows capturing distinct local features anywhere in an image, such as edges in particular orientations and color motif. The sharing of neurons among receptive fields dramatically reduces the network size at each convolutional layer, hence one could build a very deep model using the existing computing hardware.

The pooling layer merges local features detected by the convolutional layers. One common pooling function is the maximum operation over filtered neuron responses from neighboring regions, within a single channel or across multiple channels (i.e., feature maps). One obvious benefit of pooling is the dimensional reduction of feature space. Pooling also introduces invariance to a small amount of shift and distortion.

The typical CNN architecture discussed above originates from LeNet-5 designed to solve the handwritten digit recognition problem [11]. Proposed in 1998, LeNet-5 is the very first influential CNN-based application. In the 2012 ImageNet challenge [12] AlexNet [13] achieves spectacular performance improvement made in image classification. Since then many research activities have been evolving around CNN.

3.2.2.1 Feature Learning

Recent research [14] from neuroscience shows that the bio-inspired deep CNNs rival

the representational performance of inferior temporal (IT) cortex of primate on visual object recognition tasks. Their finding is drawn from a direct comparison of performances among AlexNet, Clarifai [15] and primates. However, whether the recognition mechanism in these CNN is similar to a primate's optical system (i.e., visual and IT cortexes) or not is not addressed in their work. So do these CNN 'see' as primates do? What features in an image do these CNN rely on to recognize there is a papillon not a poodle in the image?

A visualization technique is introduced by Zeiler et al. [15] to examine the activation of a trained CNN given a new image. The technique attaches a parallel De-Conv net to a trained CNN to probe what triggers activation in the trained CNN at each layer. Firstly, a test image is input to the trained CNN for a given activation at a particular layer, all other activations on the same layer are switched off. Then this processed activation map is passed to the De-Conv net to reconstruct the corresponding activation in the layer below. Repeating this procedure down to the data layer reveals the image patches that most excite a feature map at each layer. Fig. 3.3 shows an example of filters and activations generated using the visualization toolkit created by Yosinski et al. [16] where the De-Conv technique is applied. The top left image is the test image. We can see that the most 'excited' filters from convolutional layers 1 to 5 capture features from very local (edges) to motifs (texture and shape that are useful to 'recognize' objects in the test image) to the high-level semantic concept (faces of dog). Image patches in the dash line box are patches of training images that give top activation value to the filters displayed in the respective convolutional levels.

"What does a CNN see?" is addressed by Mahendran et al. [17] through reconstructing the original image from the last convolutional layer (before the softmax classification

layer) with prior regularization. Fig. 3.4 shows an example from their work, where we can observe that rich human comprehensible visual information is retained in the mpool5 (layer 10 of the CNN testing).

Fig 3.3 Filters and Activations Inside a CNN
Generated Using Deep Visualization Toolbox [16]

Fig 3.4 Image Reconstruction From Activation Layer.

Top row: original images. Bottom row: images reconstructed from activation of the mpool5 layer. Images are adopted from [18].

These visualization works shed light on the feature representation learned in each layer of a CNN. They also provide methods to diagnose or search for model architectures that would lead to a better performance.

3.2.2.2 Network Architecture

Both the network depth and performance of CNN for visual tasks have increased rapidly over the past decade. For example, in 2014 both VGGNet (19 layers) and GoogLeNet (22 layers) achieved 7.3% and 6.7% top-5 error rate respectively in ImageNet challenge classification category. In 2015, using a 152-layers CNN, ResNet achieved 3.57% top-5 error rate on the same task. Besides going deeper, we see network design has been deviating away from the typical architecture shown in Fig. 3.2, with the goal of maximizing the utilization of computing resources in a network, thereby having stronger representation power.

Szegedy et al. [18] introduce the Inception module as shown in Fig. 3.5, where 1x1 convolutions are used extensively to reduce the dimension of the input vectors before expensive 3x3 and 5x5 convolutions and after max pooling.

Fig 3.5 A Schematic Diagram of the Inception Module [18]

The expected benefit from big kernels is to capture spatially spread out local features. The output of an Inception module is the concatenation of convolutions covering regions of different sizes from its input, which bears some resemblance to conventional multi-scale signal processing. The design is theoretically inspired by a finding from Arora et al. [19] that an optimal network topology can be constructed layer-by-layer through analyzing the correlation statistics of a layer output and clustering them into groups by their correlation. Inception modules form the backbone of the GoogLeNet which has another architectural property where fully-connected layers are dropped. As a result, the GoogLeNet uses twelve times fewer model parameters than AlexNet.

An extreme deep design is introduced by He et al. [3], called ResNet. The ResNet architecture overcomes the problem that learning accuracy drops when a network goes very deep.

Fig 3.6 A Schematic Diagram of Residual Learning Module [3]

Using a typical CNN architecture, it is shown that the training accuracy on the CIFAR-10 [20] database degrades rapidly when the network layers increase from 20 to 56. The critical recipe of their solution is that instead of having each layer learning a mapping from its input directly, a mapping plus the origin input is learned across 2 to 3 convolution layers. Fig. 3.6. depicts the concept of this recipe, called residual learning. In

a residual module $F(x)$ denotes the residual between the input signal x and the learning objective. This design is inspired by residual representation approaches from shallow learning architecture [21] where the purpose is to achieve a compact signal encoding and a multigrid method architecture [22] which is used to synthesize a partial differential equation's solution from many sub-solutions of residuals projected onto multi-scales grid spaces. Many residual modules stacked together form the ResNet. Most of the residual modules have 3x3 convolution kernels, inspired by the VGGNet architecture [23].

Deep CNN architectures deliver exciting performance in image classification; however, designing and training such networks requires tremendous expertise as well as computing resources. In general, weeks are needed to train on millions of images with parallel training techniques either over GPUs or CPUs in the same node or clustering system. If these encoded representations are generally useful to visual tasks, could these representations be reused to solve other problems while saving computing resources and time? And how? This will be investigated in section 3.3.2.

3.2.3 Recurrent Neural Network

Recurrent neural network (RNN) is a class of learning machines where current and past neurons are connected to form a directed cycle, as shown in the left part of Fig. 3.7. When input elements are parsed one by one, e.g. words in a sentence or stock price over time, information of previous elements are encoded in the hidden gates forming a 'memory'. Such architecture makes RNN suitable for tasks with sequential inputs, whether for supervised or unsupervised learning. In this section, we provide a brief review of RNN in the supervised learning context.

General RNNs with different connection among input and hidden neurons were proposed back in the 1980s. However, training such networks was not practical, the gradients (of errors) tend to either explode or vanish in the 'memory' components. In the mid-1990s and recent years, progresses in architecture design and training methods have been made. Hochreiter et al. [34] propose the long short-term memory (LSTM) architecture. By using LSTM and a gradient-based training method, they are able to solve a problem with a sequence length of 1000 on artificial data. Fig. 3.8 shows a typical structure of an LSTM block, where circles are neurons and gates of different functions that store weights (i.e., memory here), and boxes are multiplication or summation operations. Outputs of the gates determine whether the information is allowed to flow into the block. If the output value of a forget gate is close to zero, the block omits values remembered.

Fig 3.7 A Schematic Diagram of Recurrent Neural Network [144].

Circle represents a neuron at a time and the solid diamond represents states of neurons in past time.

Together with the advances in addressing the vanishing and exploding gradient problem [35], LSTM finds success in mature learning applications such as speech, acoustics, and handwritten character recognitions, as well as new types of application, including question answering and language translation.

Fig 3.8 A Schematic Diagram of a Typical LSTM Block [34]

3.2.4 Autoencoders

In supervised learning, given observation x we directly model a conditional probability distribution $P(x|y)$ where y is a class label or continuous value. It is also called discriminative learning where $P(y|x)$ is used to predict y . In unsupervised learning, task specific information such as class label or target value is absent. The learning goal is to find features that represent inputs well. There are probabilistic-based methods such as Deep Belief Network. In this section, we look into autoencoder and its two variations sparse autoencoder and denoising autoencoder.

Autoencoder is also called auto-associator or Dialolo network, where the dimension of the network output is the same as its input. Its objective is to learn a representation (encoding) of input so that it can be reconstructed from the representation. Fig. 3.9 shows a typical autoencoder, it always has an encoder and decoder of single or multiple hidden layers, with the neurons at adjacent layers fully connected. Let f and g denote the encoder and decoder respectively, training an autoencoder with a single hidden layer is to minimize the reconstruction error $L(x, x')$, traditionally square errors $\|x - g(z)\|^2$, where $z = f(x)$ and $x' = g(z)$. When the dimension of the code (representation) z is less than

the input dimension, it is an under-complete autoencoder for dimensionality reduction purpose, for example in Hinton et al. [45]. If the hidden and output layers are linear, an autoencoder becomes principal component analysis (PCA). (See Appendix A for their connection.) With non-linear layers before and after the code, an autoencoder learns a nonlinear generalization of PCA that tends to be more powerful.

Fig 3.9 A Schematic Diagram of Autoencoder

On the other hand an overcomplete autoencoder learns a code with dimension greater than its input. An overcomplete representation has the potential to better approximate complex distribution properties in the data. However, a problem is that the network can simply copy the input without learning anything useful. To overcome the problem, different regularization methods were introduced to encourage an overcomplete autoencoder to learn abilities other than merely copying its input, for example sparse and contractive autoencoders. Another approach is to embed the learning ability into a reconstruction error function, for example, a denoising autoencoder.

Sparse autoencoder introduces a sparsity penalty to the training objective in addition to

the reconstruction error. The cost function is $L(x, g(f(x))) + \lambda\Omega(z)$, where z is output of the encoder. The penalty function Ω inhibits z being active everywhere, i.e., z is inactive (close to 0) most time in most dimensions, and λ controls the penalty force. Let ρ be a sparsity parameter (mean of a Bernoulli random variable), a common choice of the penalty function is Kullback-Leibler (KL) divergence which measures the difference between two probability distributions. As shown in (3.1), $\hat{\rho} = \frac{1}{m} \sum_{i=1}^m f(x_i)$ is average activation from m inputs. When $\hat{\rho}$ equals ρ the penalty is zero, and it increases monotonically when $\hat{\rho}$ diverges from ρ . Minimizing the loss function forces $\hat{\rho}$ closer to desired sparsity level.

$$\Omega(z) = D_{KL}(\rho || \hat{\rho}) = \rho \log \frac{\rho}{\hat{\rho}} + (1 - \rho) \log \frac{1-\rho}{1-\hat{\rho}} \quad (3.1)$$

In a situation where the direct supervised learning causes over-fitting due to a lack of labeled data, sparse autoencoder can be a choice to learn an overcomplete representation from the data. The overcomplete representation is then used as input features for supervised training. In [46] Ranzato et al. propose a sparse encode-decode architecture to learn shift invariant features from images.

Denoising autoencoder (DAE) changes the reconstruction error of the cost function instead of adding a penalty. A DAE minimizes $L(x, g(f(\tilde{x})))$, where $\tilde{x}$ is a copy of x with noise generated from a stochastic corruption process. Types of noise can be additive Gaussian noise, masking noise or salt-and-pepper noise. In the original implementation [47], masking noise is used. The stochastic corruption process sets v proportion components of each input to zero, and v varies from 10% to 50%. From manifold learning perspective corrupted data $\tilde{x}$ is likely off the manifold where clean data resides, as illustrated in Fig. 3.10. A DAE learns a stochastic operation that maps $\tilde{x}$ on or nearer to

the manifold while learning useful feature from uncorrupted part of the data. Experiments by Vincent et al. [48] on hardened MNIST dataset (with rotation, different types of background noise applied to original samples) show that a shallow structure SVM network can benefit from features learned by DAE in an unsupervised way.

Fig 3.10 Manifold Learning of Denoising Autoencoder [47]

Clean data (circles) align near a low-dimensional manifold (in dash line). Corrupted data (solid black dots) are generated from a stochastic process denoted by a big grey circle. When trained the DAE minimizes the reconstruction error $L(x, g(f(\tilde{x})))$, corrupted data $\tilde{x}$ is projected (indicated by the dash arrow line) on to the nearest point on the manifold.

3.3 Model Training

In this section, we describe techniques and terminologies about training neural networks that will be used in works in later chapters.

3.3.1 Gradient Descent Optimization and Hyperparameters

Gradient descent is an algorithm that is by far the most common way to train neural networks. It minimizes an objective function of a model by updating the model parameters in the opposite direction of the gradient of the objective function w.r.t. to the

parameters. The parameters are updated in batch, or each training sample, a.k.a. stochastic gradient descent (SGD). The vanilla gradient descent methods do not guarantee good convergence because it is very hard to choose a proper base learning rate and updating schedule. Several variants of optimization algorithms had been proposed to overcome the practical difficulties. They are momentum, Nesterov accelerated gradient, Adam, Adagrad, RMSprop and more [140]. Momentum is the most widely used method. We use it in later works.

Let the goal be to minimize the regularized error, L , between prediction value of training data and corresponding ground truth, (3.2) outlines how the mini-batch with momentum method updates model parameters W . At training point $i+1$, the model parameters W_{i+1} is updated with V_{i+1} , which is computed by adding a fraction μ of the update vector from point i to the negative gradient of L . The momentum term μ is usually set to 0.9 or a similar value. The base learning term α_0 varies per the training data. The learning rate is scheduled to drop every fixed numbers of training iterations by the term γ .

$$\begin{aligned}
 W_{t+1} &= W_t + V_{t+1} \\
 V_{t+1} &= \mu V_t - \gamma * \alpha_0 \nabla L(W_t) \\
 L(W_t) &\approx \frac{1}{m} \sum_{i=1}^m \text{cost}(f_w(x^{(i)}), y^{(i)}) + \lambda * r(W)
 \end{aligned} \tag{3.2}$$

Besides data and network design, these training variables have an impact on performance of a trained model. They are termed as hyperparameters. Hyperparameter searching is very time consuming. Common strategies include trial and error, grid search, random search and Bayesian optimization. A combination of trial and error and grid search is adopted in our work.

Another common hyperparameter is dropout rate. Dropout is a popular technique to prevent model from over-fitting to train data. It works by randomly dropping output from nodes of a neural layer during the training. It is first proposed by Srivastava et al. [104], where the authors apply dropout values ranging from 0.9 to 0.5 to input and different hidden layers to handwritten character recognition and object classification problems. A node's output has 50% probability of being masked out when the dropout values equals 0.5. Experimentally the authors show that this simple technique improves the performance of neural networks.

3.3.2 Transfer Learning

Re-purposing large scale deep CNN to solve similar visual tasks where labeled data is limited or to create an entirely different type of visual applications has been an active research area in recent years. In [24] Oquab et al. design a method to transfer representations (weights exclude last fully connected layer) from the AlexNet trained on the ImageNet for image classification task to object and action classification tasks on Pascal VOC [25] datasets. To adapt the difference of object classes between the ImageNet and Pascal VOC, layer eight of AlexNet is replaced by two fully connected layers. Their work shows a simple transfer learning creates a new start-of-the-art performance on target datasets for both object and action classification tasks. Experiments carried out by Razavian et al. [26] with OverFeat network [27] confirm this finding. OverFeat is trained for image classification on the ImageNet dataset. Using an off-the-shelf CNN representation with linear SVM classifier framework, the authors test a variety of target tasks including object classification, scene understanding, fine-grained object recognition, attribute detection and visual instance retrieval on datasets commonly used in the

respective task. Performance comparison between using CNN features and various traditional feature descriptors such as deformable part descriptor (DPD), scale-invariant feature transform (SIFT), histogram of gradients (HOG) are conducted for fine-grained recognition and attribute detection tasks. The conclusion is that off-the-shelf CNN representation is a strong competitor to more sophisticated and highly tuned state-of-the-art methods.

A lot of successful applications are built on this idea. The R-CNN proposed by Girshick et al. [28] is the choice for object detection task now. In R-CNN, AlexNet or VGGNet trained for the ImageNet challenge is used to extract a feature vector from each region proposal (the potential object of interest). These feature vectors then served as inputs to SVM classifiers for detection purposes. Karayev et al. [29] introduce an approach to predict styles of images and paintings. Among a broad set of image features they used are layer 5 and 6 activations from DeCAF. DeCAF is a variant of AlexNet implemented by [30] to study the feasibility of transfer learning.

Another trend is re-using trained CNN to fine-tune a network for specific tasks. DeepFix [31] utilizes VGGNet to boost the performance of visual saliency prediction, which is among the leading models on the MIT saliency benchmark MIT300 dataset [32]. The architecture of DeepFix is a mixture of VGG-16 net and the Inception module. During training, the weights of VGG-16 are used to initialize the VGG layers.

This review about re-using deep CNNs is far from exhaustive. There are many innovative applications along this line. In recent years, applying VGGNets to art applications by Gatys et al. [33] attracts a lot of interest in deep neural networks from public media and app developers. Cross application domain transfer learning, for example, from natural image classification to pathology detection in medical radiology

image, is also on the rise.

3.4 Summary

In this chapter, we investigate the origins of deep learning, and learning methods and training techniques that will be applied to the IQA studies in later chapters.

When more and more deep learning networks, in particular, convolutional neural network-based machine vision systems are invented and transferred into commercial products, we anticipate an increasing need for good and fast IQA algorithms that can be conveniently integrated into the deep learning framework. IQA will continue play a vital role to digital imaging systems and applications in the machine learning era. This provides the author a motivation to research CNN-based IQA methods.

Chapter 4

A Comparison Study of Image Quality Assessment

There is an idiom that the devil is in the details. The motivation for the work in this chapter is to investigate the details in full-reference and no-reference IQA problems, from visual feature extraction and description to methods that learn a quality score. Through this study, we propose enhancements within the traditional IQA paradigm as well as experiment with deep learning approach. Experimentally we show that enhanced traditional methods can achieve a competitive performance compared with deep neural network based full-reference IQA in the literature. In no-reference IQA, we show that a small CNN network can achieve performance competitive with much larger CNN networks in the literature.

Firstly, full-reference IQA algorithms are proposed, and comparison experiments results are explored with leading proposals in the literature. Secondly, no-reference IQA algorithms are proposed, developed, and subsequently, experiments are carried out and analyzed. In the end, a conclusion is drawn, and potential future work is identified.

4.1 Full-Reference IQA

4.1.1 Enhancement Proposal

Two trends observed from the literature review inspire this work. The first is the simplicity and effectiveness of features in SSIM and FSIM. The other is that a metric's performance benefitting from complicated pooling methods is arguable. Pooling is a step to compute one (global) quality score from a local (or pixel level) quality map produced

by the quality analysis. Maximum, percentile averaging, visual saliency based weighting, normalizing, and many other methods can be found in the literature. For more details see Chandler et al. [79]. These works imply that pooling is a highly complex mapping function.

We like the simplicity of feature selection in SSIM and FSIM. We have also observed that several works [67-69] on no-reference IQA methods look into the statistics of the natural images [70]. Therefore we search for equal or better feature descriptors in the direction of statistics instead of the structural similarity theory that underpins SSIM and FSIM. Meanwhile, we have also observed that potential useful information in the color space is under appreciated. This leads us to include chromatic features for seeking enhancement. However, we are not very convinced by the benefit of complicated pooling methods proposed in the literature, and decide to adopt the learning-based trend for pooling purposes. Along this line of thinking, we design three sets of slightly different features descriptors, and train models to validate the result. Table 4.1 gives a summary of the design specification.

Table 4.1 Design Specification of Full-reference IQA Proposals

Proposed Method	Features	Feature Descriptor	Dimension of Feature Vector	Learning Method
1	Luminance, opponent color components (LAB color space)	Statistics (1 to 4 moments), Log transformed, 4 scales	48 (4*3*4)	nu-SVR (RBF kernel) g = 0.1 C = 1
2	Same as method 1	Czekanowski coefficient, 4 scales	12 (4*3)	nu-SVR (RBF kernel) g = 150 C = 1
3	Phase congruency, gradient magnitude, luminance and chrominance (in YIQ color space)	Statistics (1 to 4 moments), Log transformed, 4 scales	80	nu-SVR (RBF kernel) g = 0.1 C = 1

The dimension of images from the IQA datasets ranges from 192-by-256 to 384-by-

256 pixels after downsizing by a factor of 2. After downsizing, the dimension of a pixel level quality map is still too high to be trainable with support vector machines. Our goal here is to minimize the feature extraction process. So the process starts with adopting FSIMc [57] features, i.e., method 3. Unlike FSIMc that uses the SSIM style score summary function, we introduce simple statistics as feature descriptor to reduce the feature dimension and a machine learning approach. Inspired by the findings of statistical properties from natural images in the spatial domain (see section 2.2), simple statistics is employed for dimensional reduction which is denoted as feature descriptor in methods 1 and 3.

The Canberra distance and Czekanowski coefficient (4.1) are chosen as a feature descriptor in method 2. Experiments show that Czekanowski coefficient serves the dimensional reduction purpose and leads to a better model performance in this case.

$$d_z(i, j) = 1 - \frac{2 \sum_{k=1}^p \min(x_{ik}, x_{jk})}{\sum_{k=1}^p (x_{ik} + x_{jk})} \quad (4.1)$$

4.1.2 Experiment Setup

This section presents and discusses the experiments conducted to evaluate the performance of the proposed methods. The datasets used are obtained from authors' websites (see Table 2.1 of chapter 2). TID2008 has a lot of commonality with TID2013. TID2013 includes all the seventeen types of distortions of TID2008 and the reference images are the same in both datasets. The difference is that there are seven new types of distortion that are introduced in TID2013, and the level of distortion is finer, five instead of four levels. In this report, experiment results on TID2008 are not included as the information provided would be redundant.

Distorted images are uniformly sampled from each dataset with replacements to form the training and testing sets. The train-test ratio is 80%-20% unless otherwise stated. Using an 80%-20% ratio, the training sample sizes are 2400, 693, 623 for TID2013, CSIQ, and LIVE respectively.

For nu-SVR version 3.21 of LIBSVM [80] is used. nu-SVR is a variant that reduces the burden of searching for the error tolerance parameter epsilon suitable for the problem.

4.1.3 Performance and Analysis

In this section we compare and analyze performances of the proposed methods with leading full-reference IQA published in the literature.

The comparison study between the proposed methods and existing ones is carried out on a single test-bed. The test-bed is a computer equipped with an Intel i7-3770 CPU, 3.4 GHz and 8GB RAM. The software used is MATLAB on 64-bit Windows 7.

Source codes of SSIM, VIF, FSIMc, and VSI are obtained from corresponding authors' websites. MATLAB version of PSNR is used.

The source code for SVD learning [59] is not obtainable. We implement it according to the publication and name it SVD-ML. The performance in SROCC term from our implementation agrees with the value reported in the original work.

The overall performance in SROCC term of each proposed method is summarized in Table 4.2 and Fig 4.1. The weight in the weighted average is the number of testing samples from each dataset. We observe that, except method 2, the other two methods perform on par with each other. This outcome is not surprising considering the feature dimension in method 2 is twelve; only 15% of method 3's feature dimension.

Comparing the results between methods 1 and 3, the performance advance from extra

low-level feature operators (phase congruency and gradient magnitude) is not as significant as expected. With extra phase congruency and gradient magnitude features the weighted average is 0.009 better in SROCC term. Experiments in LAB and YIQ color spaces infer this margin could probably be increased to 0.013 if LAB color space were used in method 3.

Table 4.2 Performance of Proposed Full-reference IQA Models on Single Dataset Median SROCC of 1000 runs. Trained with 80% of a dataset.

Dataset	Proposed Method 1	Proposed Method 2	Proposed Method 3
LIVE	0.9597	0.9386	0.9743
CSIQ	0.9659	0.9285	0.9709
TID2013	0.9355	0.8649	0.9421

The comparison study with existing methods is summarized in Table 4.3. We observe that in SROCC term, the proposed methods 1, 2 and 3 outperform the rest by weighted average. Method 3 outperforms all existing none deep-learning based methods across three datasets.

Table 4.3 Performance Comparison of Full-reference IQA Methods

	LIVE	CSIQ	TID2013	Weighted average
PNSR	0.8729	0.8087	0.6869	0.741
SSIM	0.9225	0.8718	0.7214	0.783
VIF	0.9636	0.9219	0.6769	0.771
FSIMc	0.9645	0.9309	0.9422	0.944
VSI	0.9524	0.9422	0.8965	0.914
SVD-ML	0.9671	0.9115	0.8357	0.872
Proposed Method 1	0.9597	0.9659	0.9355	0.944
Proposed Method 2	0.9386	0.9285	0.8649	0.888
Proposed Method 3	0.9743	0.9709	0.9421	0.953
DeepQA [137]	0.981	0.961	0.939	0.949
WaDIQaM-FR [136]	0.970	-	0.940	-

Both Bosse et al. [136] and Kim et al. [137] take a neural network approach. Using a 80% training and 20% of testing schema, the best performance reported on WaDIQaM-FR [137], in the average of 10 runs, is 0.970 on LIVE and 0.940 on TID2013, in SROCC term. The performance of full-reference DeepQA [137] in SROCC term is 0.981 on LIVE, 0.961 on CSIQ and 0.939 on TID2013. Results from both works are close tied with the performance of our method-3 which doesn't use neural network. Therefore we would argue the benefit of applying neural network and deep learning to the full-reference IQA problem is limited.

Fig 4.1 Performance Comparison of Full-reference IQA Methods.

Fig 4.2 Performance in Median SROCC of Cross Database Generalization Tests

Cross dataset performance is studied on proposed method 1 and the learning-based model SVD-ML. From Fig. 4.2 we observe that for the proposed method-1, the average test result on the LIVE dataset using the models trained on the entire TID2013 and CSIQ datasets is 0.91 in SROCC term, only 0.05 lower than model trained and tested on LIVE.

The weighted average shows that the generalization capability of models trained using method 1 is better than that of SVD-ML, 0.09 higher in SROCC term.

The computational Costs are summarized in Fig. 4.3. Proposed methods 1 and 3 extract features from 4 different scales of the image. The significantly longer prediction time of proposed method 3 is due to the phase congruency and gradient magnitude operations, which increase the training time by around 4 to 5 folds (bottom left). On transforming pixel signal across color spaces, the impact of YIQ and CIELab on the performance is not significant, but the feature computation in the CIELab color space is about twice slower. However, the experiment shows that there is around 0.004 performance gain in SROCC term.

Fig 4.3 Computational Costs of the Proposed Full-reference IQA Models

4.2 No-Reference IQA

In the IQA modeling pipeline there are two indispensable components: 1) extracting pixel signals from image patches and transforming them into quality awareness features; 2) designing a function to map the quality awareness features into one image level quality score that would be perceived by human observers. The challenge lies in that this mapping relationship is nonlinear, highly complicated, and nondeterministic. In the early generation of IQA metrics, various deterministic pooling functions were designed to approximate this mapping relationship. As its name indicates a pooling function takes the quality awareness features extracted from patches and outputs a score that predicts the image quality. Later on, learning-based IQA models emerge in the literature as a pooling mechanism. Among them is support vector machine which is the widely preferred learning method until recent years.

In this section, we aim to answer two questions. Firstly, whether the model performance could be further improved within the traditional handcrafting feature paradigm. We present new feature designs that lead to performance enhancement. Secondly, whether CNN-based methods could directly learn a better quality estimator from pixel signals is investigated. We propose a lean CNN network design. Experimentally, we show that our method achieves a competitive performance compared with methods in the literature that use larger deep neural networks.

4.2.1 Enhancement Proposals

On feature extraction and description, studies [69,73,74] of natural images and texture statistics show that pristine natural images exhibit specific statistical properties that are absent in distorted images or weak by varying degrees. For example, in Fig. 4.4, the

distortion in the right side image is hardly perceivable. However, the histograms of luminance and chrominance after mean subtracted contrast normalization (MSCN) process reveal easily distinguishable distributions.

Inspired by these work and findings, BRISQUE [69] in particular, we seek a feature design schema in the spatial domain that would have better quality prediction power. Fig. 4.5 explains the detail of our multi color space and multi-scale feature design.

The global branch transforms pixel signals from the RGB color space into the HSV color space where H and S refer to hue and saturation information, and L refers to luminance information. The local branch transforms pixel signals into the LAB color space. The MSCN process is operated on the original and downsized image by a factor of two. L^1 that denotes the MSCN process is applied to the original image scale. Superscript (2) denotes the deviation of the local signal. For both scales, Gabor operators are applied to extract information in four orientations. D refers a feature descriptor that represents feature into a reduced dimension space that is trainable using a non-deep learning-based machine learning method.

Fig 4.4 Image and Histogram of Corresponding Signals after MSCN Process

Signals are luminance, and a and b channels in color-opponent space. Left is a reference image from the TID2013 dataset. Right is the corresponding distorted image with noise in color components.

Fig 4.5 A Block Diagram of Feature Design Schema

Compared to BRISQUE, our feature design has unique treatments: 1) information in color spaces is exploited to capture new types of quality distortion that is introduced in TID2013, for example, chromatic aberrations, change of color saturation and more; 2) a high-order statistic feature descriptor is proposed; 3) our feature descriptors represent distortions at both local (patch) and global (image) levels; 4) instead of using a product of MSCN processed signals in adjacent orientations as in BRISQUE, we apply MSCN process to the deviation of the local signal; 5) the shallow neural network is explored as a learning method.

The proposals are summarized in Table 4.4 and Fig. 4.6. In NRSTATS method we use the simple statistics (1st to 4th moments) as feature descriptors and nu-SVR (with RBF

kernel) to learn a quality model. In the NRGGD method, the network is trained with Levenberg-Marquardt algorithm. The transfer functions in the hidden layers are hyperbolic tangent.

Table 4.4 Design Specification of No-reference IQA Enhancement Proposals

Proposed Method	Features	Feature Descriptor	Feature Vector Dimension	Learning Method
NRGGD	See Fig. 4.5	Generalized Gaussian Distribution (GGD)	47	2-layer General Regression Neural Network (GRN)
NRSTATS	Same as NRGDD	Statistics (1 to 4 moments)	83	nu-SVR (RBF kernel) g varies from 0.6 to 1.2 per dataset C=1

Fig. 4.6 A Schematic Diagram of the General Regression Neural Network in NRGGD

4.2.2 CNN-based Method

Recent years, a few research works attempt to enhance or replace the entire IQA modeling pipeline through the deep learning approach. For example, DBN-based methods are proposed by Ghadiyaram et al. [101] and Kim et al. [137] and CNN-based methods are proposed by Kang et al. [51] and Bosse [136]. Inspired by the work in [51], we seek new CNN network designs.

Fig. 4.7 depicts the architecture of our design. A distorted color image is first transformed into YIQ color space and then it is cropped into patches. The luminance and

summation of color information of a patch is concatenated into a single input vector to the convolutional network. Merging two chrominance channels into one reduces the dimension of input vector from 3072 to 2048.

Fig 4.7 A Schematic Diagram of NRCNN Model

The kernel size of both convolution layers is 5 pixels, and their output feature maps are 32 and 64 respectively. The first pooling layer reduces the feature maps by a factor of 2. The kernel size and sampling stride are both 10 in the Max and Min pooling layers, which reduces each feature map to 1 pixel. The uniqueness of this design lies in the signal pre-processing: the color images are transformed from RGB to YIQ color space, where Y encodes luminance information, I and Q represents the chrominance information.

The model design involves a range of hyperparameters, patch size, crop stride, base learning rate, momentum, dropout rate and etc. Through parameter searching, we find that the model performance is not sensitive to the crop stride. A schema for a patch size of 32-by-32 pixels with a stride of 24 pixels is chosen. We train the model with SGD method and the momentum is set to 0.9.

4.2.3 Experiment Setup

The datasets used in the experiment include the three as explained in section 4.1.3 and the LIVE in the Wild. Since LIVE in the Wild doesn't include reference images, full-reference models cannot be evaluated with it. Train and test sets are prepared in the same way as in 4.1.3.

4.2.4 Performance and Analysis

In this section we compare deep learning-based and other learning-based no-reference IQA methods separately. The experiment data of other learning-based methods and NRCNN are collected from a Linux box with 3.4GHz CPU and GTX1070 GPU. The source codes of BIQI, DVIINE, BLIIND2 and BRISQUE are obtained from authors' websites.

4.2.4.1 Learning-based Methods

Firstly, we use the NRSTATS design to discuss the impact of global features and the MSCN treatment to the deviation of the local features on model performance. The experiment result is shown in Fig. 4.8. In NRSTATS-1 we exclude the global feature vector and in NRSTATS-2 we exclude both the global and the deviation of the local feature vectors. Comparing NRSTATS-1 and NRSTATS-2, we observe that the model obtains significant performance gain on both TID2013 and CSIQ datasets, and some gain on the least challenging dataset LIVE. Comparing NRSTATS and NRSTATS-1, a meaningful performance gain is observed on the CSIQ dataset.

Now we compare performance of NRGGD and NRSTATS with leading peers published in the literature. As shown in Fig 4.9, in terms of median SROCC of 100 runs, our proposals outperform those in the literature by weighted average. Because the most

challenging dataset TID2013 is more than twice as large compared than the LIVE, We use the weighted average to derive a less biased overall performance of each method. On single dataset, only BLIIND2 is 0.01 better than our proposals on the LIVE dataset.

Fig 4.8 Comparison of Feature Design and Model Performance in SROCC Term

Fig 4.9 Comparison of Non Deep learning based No-Reference IQA Methods

Cross dataset performance is summarized in Table 4.5. Top row is performance tested on entire dataset. We observe that models trained using NRGGD and NRSTATS features generalize better across datasets than BLIINDS2. The performance edge over BRISQUE is relatively weak. The bottom row is performance tested on the four common distortion types: Gaussian blur, additive white noise, JPEG and JP2K compression. These four distortion types are less challenging hence the model generalization capacity of all methods leaps when tested on TID2013.

Table 4.5 Performance of Cross Datasets Generalization
(Median Value of 100 Trials in SROCC Term)

	CSIQ	TID2013		LIVE	TID2013		LIVE	CSIQ
BLIINDS2	0.5353	0.3409	BLIINDS2	0.8373	0.2764	BLIINDS2	0.7824	0.5118
BRISQUE	0.6645	0.4134	BRISQUE	0.8692	0.4813	BRISQUE	0.7987	0.6231
NRGGD	0.7042	0.4232	NRGGD	0.8305	0.4354	NRGGD	0.7731	0.7344
NRSTATS	0.7964	0.4595	NRSTATS	0.7475	0.4799	NRSTATS	0.7684	0.6746
	CSIQ	TID2013		LIVE	TID2013		LIVE	CSIQ
BLIINDS2	0.5246	0.8251	BLIINDS2	0.8646	0.7304	BLIINDS2	0.8070	0.5154
BRISQUE	0.6614	0.8625	BRISQUE	0.8858	0.7997	BRISQUE	0.8187	0.6396
NRGGD	0.7371	0.8867	NRGGD	0.8509	0.8585	NRGGD	0.8175	0.7242
NRSTATS	0.7622	0.8377	NRSTATS	0.7723	0.8411	NRSTATS	0.7906	0.6351

4.2.4.2 Deep Learning-based Methods

Firstly, we discuss the performance of our proposal with peer methods in the literature. The performance of NRCNN is reported in median SROCC of 100 runs. Performance data of peer methods are adopted from respective authors' published work. We acknowledge that various experiment parameters might have a small impact on model performance. One important parameter is how the training and testing sets are prepared. All the peer works adopt an 80% for training and 20% schema as we do. Both Kang et al. [51] and Ghadiyaram et al. [101] report performance in median of 100 runs. DeepQA [137] and WaDIQaM [136] report on average of 20 runs and 10 runs respectively. MEON [138] reports median of 10 runs. From Table 4.6, three qualitative observations can be made:

1) The performance of NRCNN on LIVE dataset is among the upper range, slightly below DeepQA. All the three CNN-based methods here have two fully connected layers. As for convolutional layers, NR-IQA uses one, our network uses two, and DeepQA uses

six. WaDIQaM-NR takes a two-phase approach. In phase one, a VGG inspired 10 convolutional layers CNN is trained as a feature extractor. However, the depth of CNN seems not to deliver performance edge. This is also true on TID2013 dataset.

Table 4.6 Performance of Deep Learning-based No-Reference IQA Methods

	LIVE	CSIQ	TID2013	Live in the Wild
NRCNN	0.971	0.925	0.807	0.580
NR-IQA [51]	0.956	-	-	-
FRIQUEE [100,101]	0.950	-	-	0.672
DeepQA [137]	0.975	-	-	-
WaDIQaM-NR [136]	0.954	-	0.761	0.671
MEON [138]	-	-	0.808	-

2) Comparing the result on TID2013, the most challenging synthetic quality dataset, our NRCNN design is better than WaDIQaM-NR and is competitive with MEON.

3) On the Live in the Wild dataset where quality distortion is not synthetically inserted into pristine images, our proposal performs worse than WaDIQaM-NR and FRIQUEE. The latter is a DBN-based two-phase method with four hidden restricted layers of Boltzmann machines. It is not clear here whether the extra convolutional layers or DBN method lead to a better outcome on this dataset or due to the two-phase approach in which feature extraction and quality score regression are trained separately. However, comparing results of these two methods and their phase-two design, we could exclude that WaDIQaM-NR's approach to address the spatial variance of image quality with an extra branch gives no performance edge.

Comparing NRCNN with our other no-reference IQA proposals, we observe that the CNN-based method delivers a performance edge across the four datasets (See Fig. 4.10).

Fig 4.10 Performance of NRGGD, NRSTATS and NRCNN in Median SROCC

Next, the cross dataset generalization capability of NRCNN is discussed. As shown in Fig. 4.11, we report experiment results in a median SROCC of 100 runs, with error bars. Instead of retaining 100 models on each full dataset, we utilize those models trained on 80% of one dataset to conduct this generalization test on the other three datasets. In the top left sub-diagram, LIVE (train) refers to the performance on trained dataset. CSIQ-4 refers to the performance of model trained on LIVE and test on the common 4 distortion types of CSIQ.

We observe that models trained on the LIVE and CSIQ datasets exhibit a dramatic performance drop on the Live in the Wild dataset, and correspondingly, models trained on the LIVE in the Wild performs very poorly on the TID2013 dataset. This implies that the LIVE in the Wild dataset includes very unique quality degradation characteristics that are not captured in all the other three datasets where distortion are synthetically injected, although there are 24 types of synthetic distortion in TID2013.

Fig 4.11 Cross Dataset Performance of NRCNN in Median SROCC

4.2.4.3 Computational Cost

The experiments are carried out on a Linux box with 3.4GHz CPU and GTX1070 GPU. The computational cost of training NRCNN models on GPU is summarized in Table 4.7. NRCNN takes a 32-by-32-pixel patch as input, for each model a random 80% of a dataset is sampled and processed. The patches are extracted without resizing, and image in the datasets comes in different sizes except TID2013. One epoch is that the model is feed all training patches for one time.

Table 4.7 Computational Cost of Training NRCNN Models

	LIVE	CSIQ	TID2013	Live in the Wild
Num. of patches	247,436	183,500	464,640	209,568
Total size in GB	3	2.3	5.7	2.6
Minutes for data preparation	9	6.7	18	7.5
Minutes per epoch	1.6	0.6	2.2	1.4
Num. of epochs	25	25	40	40
Minutes per model	49	21.7	106	63.5
Days for 100 models	3.4	1.5	7.4	4.4

Given an image, the time of using our no-reference IQA methods to assess its quality is summarized in Table 4.8. The quality prediction of NRCNN is carried out on GPU, and

the rest computations reported here are executed on CPU. In NRGGD and NRSTATS, the data preparation time includes feature computation and patch extraction operations.

Table 4.8 Testing Time of Learning-based No-Reference IQA Models

	NRGGD	NRSTATS	NRCNN
Data Preparation	2.036	2.495	0.869
Quality Prediction	0.000	0.000	0.472
Total (Seconds)	2.036	2.495	1.341

4.3 Summary

This chapter reports a comprehensive study of the image quality assessment problem. The investigation into full-reference and no-reference learning-based IQA methods provides us a solid grasp of existing IQA techniques, as well helps us to identify the gap that is meaningful to address. During this study, we conduct more exploration than reported here, for example, experimenting with two training phases framework, applying an autoencoder to learn a quality representation of image patch in unsupervised and semi-supervised ways. Although we learn from these exercises, their results don't directly influence the thesis work further. We decide not to include them.

So far all the learning-based no-reference IQA methods access visual quality at image patch level. The final image level quality score is pooled together either using a simple average or training a pooling network. As we know today, many deep learning-based detection and localization applications can take a whole image as input. This motivates us to look into the feasibility of learning the final quality score directly from a whole image. We are also curious to find out whether learning a quality score directly from a whole image would lead to a better IQA model on the LIVE in the WILD dataset.

Chapter 5

No-Reference Image Quality Assessment Through Transfer Learning

Transfer learning has emerged from recent years' great success in applying convolutional neural networks to object recognition tasks. In this work, we describe a transfer learning framework that learns an image quality estimator end-to-end in classification or regression. Experiments show that this approach achieves the state-of-the-art performance on four publicly available image quality databases. Image quality is ignored in visual recognition challenges such as PASCAL VOC and ILSVRC. However, it poses a practical challenge to the design of robust machine vision applications. This work responds to such a challenge without impacting the performance of recognition tasks.

5.1 Background

Visual quality assessment has played an essential role in the design of traditional image and video applications where visual signals are subject to different kinds of distortions before being presented to human users. With or without a reference image or knowledge extracted from the reference, the task of a visual quality model is to evaluate the perceptual quality of an image as close as the human visual system would perceive it. Research in this field can be categorized into the early days' development of full [86] and reduced reference quality metrics to later distortion specific [62] and general-purpose no-

reference machine learning-based quality models where support vector machines are widely adopted.

Recent breakthroughs of the convolutional neural network (CNN) in object classification tasks [13] inspire many types of research. Among them, one way is using well-known CNN models [3, 18, 23] trained on millions of images as a generic visual feature extractor, i.e., transfer of knowledge learned from source tasks to similar or relevant target tasks in a beneficial way.

Weiss et al. [99] present a comprehensive survey about categories, strategies, and applications in transfer learning. The original motivation for transfer learning is in situations where obtaining sufficient training data and labels for target domains can be difficult and expensive; using datasets that are related to but not the same as the target domain to build learners is an attractive approach. At higher levels, there is homogeneous transfer learning where using a dataset from a source domain to solve a task in a different target domain, and heterogeneous transfer learning where both the dataset and task to solve are from a target domain that is different from the source domain. Our framework belongs to the latter.

In the context of CNN for fully supervised image applications, Razavian et al. [26] and Donahue et al. [30] demonstrate generic features learned by deep CNNs trained on object recognition tasks are better candidates than traditional hand-designed features for other visual recognition tasks such as scene classification, object localization, and fine-grained recognition. Unlike these works, we propose to use these generic features to solve a lower level image understanding task, namely, quality understanding.

Generic visual features learned from a source task may be fully or partially retained in the convolutional layers or fully forgotten. The last setting is often named as fine-tuning.

Using CNNs [7-9] trained on ImageNet, we study these different settings and reports that feature transferability exhibits across all these settings by varying degrees. In our framework, the generic features learned from a source task are fully retained. If a source task is object classification, our model predicts the visual quality of the image to be classified without degenerating performance on the source task. Experimentally we show that the proposed framework establishes a new state-of-the-art in no-reference image quality assessment on LIVE [81], LIVE in the Wild [100], TID2013 [82] and CSIQ [83] image quality datasets.

5.2 Related Work

In this section, we examine influential no-reference IQA models published after 2010 that are learning-based models using either traditional or new generation machine learning techniques. Most no-reference IQA models proposed before 2010 are distortion specific i.e., they attempt to address a single or specific type of distortion such as blurring, blocking, ringing or noise. Research in general-purpose no-reference IQA models has been active since 2010. One popular approach is using the support vector machine to learn a model from carefully designed perceptual quality-aware features. In BIQI [67], the features are shape and variance parameters of general Gaussian distribution that approximates the subband coefficients derived from transforming a distorted image with Daubechies wavelet. In DIIVINE [68], the features are components from steerable pyramid decomposition. In BLIINDS2 [84], the features are general Gaussian distribution statistics fitted from DCT coefficients. BRISQUE [69] is the first no-reference IQA model that processes the signal in the spatial domain. Mean subtracted contrast normalization (MSCN) is applied to pixel intensities. The MSCN outputs are compressed

into features, using the statistical method by Saad et al. [84].

Research in automatically learning quality-aware features is introduced for first time in Ye et al. [71]. CORNIA applies k-mean clustering to learn a sparse-coding visual codebook from ZCA whitened blocks sampled from distorted images. Subsequently, a linear support SVR is trained to pool quality of blocks into one global quality score.

More recent research start to introduce deep learning techniques into no-reference IQA. In FRIQUEE [101], Ghadiyaram et al. design a stacked-restricted Boltzmann machine to learn deep features in an unsupervised pre-training phase, followed by a supervised fine-tuning phase to classify quality into 100 grades. Finally, an SVR is trained to pool the classification scores into one image quality score. Kang et al. [51] proposes a dedicatedly designed 5-layer CNN architecture to learn a quality model from 32-by-32 pixel blocks sampled from distorted images. This model outperforms its predecessors on the LIVE database. Bianco et al. [102] propose DeepBIQ. It is a 2-phases framework that uses CNN to extract feature and then trains an SVR to obtain a quality score. Both transfer learning and fine-tuning are exploited in their work.

With regards to transfer learning we are along the same line with Bianco et al.. However, our work contributes uniquely in three aspects.

- Unlike Bianco et al. which use CNN as a coarse quality classifier in the pre-training phase, our model is a single repressor or classifier trained end-to-end.
- To the best of our knowledge, recurrent neural network (RNN) has not been applied to no-reference IQA. Replacing conventional fully connected neurons with RNN in the application layer, our model shows stronger quality prediction power on images where distortions are not homogenous globally.

- We present a thorough evaluation of model performance on one quality database and generalization ability across other quality databases. Bianco et al. don't study model generalization capability and Kang et al. evaluate their proposed model only on the LIVE quality database where all distortions are globally homogenous.

In the rest of the chapter, section 5.3 introduces the proposed framework; Section 5.4 describes the data, evaluation metrics, experiment setup and analysis of experimental results; Section 5.5 draws the conclusion.

5.3 A Transfer Learning Design

Image quality is ignored in visual recognition challenges such as PASCAL VOC and ILSVRC, where most if not all images are of average to good quality. CNN models produced in these challenges are widely used today as the architectural backbone for various visual recognition tasks. However, in the real world, vision applications could receive imagery inputs of far below average quality, for example, during autonomous driving in an extremely hazy day or when image acquisition lenses are made dirty maliciously. Such scenarios pose a practical challenge to the design of robust vision applications. The objective of this work is to build perceptual quality prediction models for these backbone CNN models through re-using their visual feature extraction capacity without alteration.

Given a color image, the proposed framework first takes ten samples following the sampling method in Krizhevsky et al. [20], i.e., one sample from the center and four samples from the four corners. Each of these samples is mirrored horizontally. This

sampling method is widely used in CNN-based object recognition applications to improve overall classification accuracy. Then a quality score is estimated either using a classifier or regressor. In parallel, the framework can predict object class in the image.

5.3.1 Network Architecture

The design abandons the manual feature extraction step and instead uses the convolutional layers of popular CNN object classification models [3, 13, 23] trained on ImageNet as a generic image feature extractor and trains a quality predictor with datasets from the IQA domain. As shown in Fig. 5.1, the network takes the last convolutional layer's outputs from pre-trained models [3,18,20] as inputs. It contains two fully connected layers of 2048 neural nodes and an output layer of 70 nodes.

Fig 5.1 A schematic diagram of the architecture of DTNIQ-f IQA model

The activations from the fully connected layers are filtered by Rectified Linear Units (ReLU). Krizhevsky et al. [13] and subsequent research show that, compared to the traditional tanh and sigmoid activation units, ReLUs converges faster and leads to a better solution. The prediction layer is a softmax operation. Let $\bar{v}$ be output from the prediction layer and p the number of samples taken from one image during testing, the predicted

MOS is calculated using
$$q = \frac{1}{p} \sum_{i=1}^p \sum_{j=1}^{70} j * v_j$$

During training, the quality score of every sample from one image is treated as an

independent sample. During testing, quality scores of all samples from one image are pooled into a scalar value by averaging. We name this architecture DTNIQ-f.

5.3.2 Heterogeneous Global Distortion

Most types of distortion in the image quality datasets are homogeneous everywhere in an image, i.e., globally. However, some images in the LIVE in the Wild dataset [100] exhibit distortion of a single type concentrated in a region. See Fig. 5.2 for such examples. Left is a JPEG transmission error from TID2013 [82]. Right is a mixture of underexposure and blur from LIVE in the Wild, where multiple types of authentic distortions may exist in one image. Therefore distortion is heterogeneous globally. In such cases, the quality of image samples extracted from different regions of an image can be different. But they are labeled identically because the quality label available is per image. Treating every sample as an independent image poses a challenge to learning and taking more samples from one image is not helpful.

Fig 5.2 Examples of Heterogeneous Global Distortion

Consider when a human observer evaluates the quality of an image, eyes first fixate on different regions of the image relevant to the task at hand, then the quality opinion of each

region is formed quickly and stored in the short term memory, and finally a global score is decided. This process is similar to the operation of RNN where information of previous elements of a sequential input is encoded in hidden gates that form a "memory". Therefore we consider organizing the samples from a heterogeneously distorted image into one sequential sample can be a more optimized approach. Fig.5.3 depicts the proposed architecture DTNIQ-r, where the "memory" has n hidden states. The output from the last hidden state is regressed to a scalar quality score.

Fig 5.3 A Semantic Diagram of DTNIQ-r IQA Model

5.3.3 Learning Method

The proposed quality assessment models are implemented in Python programming language and trained on Caffe [103] platform. For training, each sample is assigned a ground truth score from its source image. In DTNIQ-f, the training goal is to minimize the multinomial logistic loss of the softmax output from the prediction layer. For DTNIQ-r, the training objective function is the Euclidean loss between the regressor output and target value.

To solve the minimization problem, we use the stochastic gradient descent method with momentum-based weights updating. For both architectures the momentum μ is 0.9, the

learning rate α is updated every 20 epochs, and the mini-batch size m is 100. Optimal values for other hyperparameters are found through grid searching. In DTNIQ-f, the base learning rate α_0 is 0.0005 for TID2013 and LIVE in the Wild, 0.005 for LIVE and 0.007 for CSIQ. The learning rate updating factor γ is 0.8 and the regularization parameter λ is 0.0005. In DTNIQ-r, $\alpha_0 = 0.01$, $\gamma = 0.5$ and $\lambda = 0.0001$. A dropout value of 0.3 to the RNN layer of DTNIQ-r is optimal. The dropout is not helpful to DTNIQ-f.

5.4 Experiments And Analysis

5.4.1 Data

Four publicly available image quality datasets are used for evaluation. They are:

- LIVE [82] release 2 has 29 reference images and 779 distorted images in five different types. They are JPEG2000 and JPEG compressions, Gaussian blur, white noise, and fast fading.
- LIVE in the Wild [100] has 1162 images. Unlike LIVE where the distortions are injected into high-quality reference images by algorithms using software, this database contains authentic image distortions captured from the real world by popular camera devices.
- TID2013 [82] has 25 reference images and 3000 distorted images in 24 types and five levels for each type. See Appendix C for the types of distortions.
- CSIQ [83] has 30 reference images and 866 distorted images in six different types they are JPEG2000 and JPEG compressions, Gaussian blur, white noise, pink noise, and contrast decrements.

For evaluation, each dataset is divided into 80% training and 20% testing randomly by reference images. Such a schema ensures the same image of different quality types or levels will not appear in a train set and its corresponding test set twice.

Reference images i.e., high-quality images are included in both training and testing. The rationale is that a quality prediction model should be able to assess a full range of quality from high to extreme distortion.

In DTNIQ-f, 30 samples are taken from each image for both training and testing. Through experiments we find sampling more patches from one image doesn't help. In DTNIQ-r, 100 samples are taken from one training image and organized into 10 sequential inputs of length ten. During testing ten samples are taken from one image.

5.4.2 Evaluation

Table 5.1 summarizes the performance of our algorithms in the median value of 100 trials and compares them with Kang et al. [51], Mittal et al. [69], Saad et al. [84] and Ghadiyaram et al. [101]. Numbers in italics are adopted from the authors' papers. We don't compare the results with that reported in [102] that uses the fine-tuning method. It re-trains all the convolutional layers of pre-trained object recognition CNN model, and as such, wipes out the knowledge about object recognition learned from ImageNet. A dash indicates the experiment result is not reported by the author.

We observe that both DTNIQ methods show the strongest correlation to the ground truth across four datasets in both SROCC and PLCC terms, except that on LIVE in the Wild, DTNIQ-f is a little behind DeepBIQ. Note that DeepBIQ uses a different source model from Zhou et al. [105], which is trained on a merged database of ImageNet and

Places205 using AlexNet [20] architecture. DTNIQ methods also outperform four existing deep-learning based methods that are compared in section 4.2.4.2.

Table 5.1 Performances On Single Dataset
(Median Value of 100 Trials in SROCC Term)

SROCC	LIVE	CSIQ	TID2013	LIVE in the Wild
DTNIQ-r	0.986*	0.901*	0.673	0.804
DTNIQ-f	0.982	0.909	0.604	0.772
<i>DeepBIQ</i> ^[102]	0.97	-	-	0.783
<i>CNN</i> ^[51]	0.956	-	-	-
<i>FRIQUEE</i> ^[101]	-	-	-	0.672
BRISQUE ^[69]	0.929	0.768	0.550	0.578
BLIINDS2 ^[84]	0.944	0.760	0.570	

PLCC	LIVE	CSIQ	TID2013	LIVE in the Wild
DTNIQ-r	0.987	0.894	0.679	0.829
DTNIQ-f	0.985	0.926	0.648	0.808
<i>DeepBIQ</i> ^[102]	0.98			0.794
<i>CNN</i> ^[51]	0.953	-	-	-
<i>FRIQUEE</i> ^[101]				0.705
BRISQUE ^[69]	0.934	0.820	0.598	0.600
BLIINDS2 ^[84]	0.948	0.811	0.642	

We also observe that using RNN DTNIQ-r correlates significantly better with the ground truth on TID2013 and LIVE in the Wild. Both have heterogeneous distorted images. This performance enhancement is not observed on LIVE and CISQ datasets where all distortions are homogeneous globally. This result supports the theoretical analysis in section 5.3.2. We take a sequence length of ten samples, which is optimal for

DTNIQ-r per our experiments. Note that the results of DTNIQ-r on LIVE and CSIQ (numbers with an asterisk in table 5.1) are from 10 trials.

5.4.3 Cross Dataset Evaluation

In this subsection, we compare the generalization capability of our algorithms with BRISQUE and BLIINDS2, where cross-database experiment results are reported. Table 5.2 summarizes the generalization ability across the entire testing dataset and the four common types of distortion. They are JPEG2000 and JPEG compressions, Gaussian blur and white noise. We use the same 100 models in section 5.4.2 for this experiment. Datasets in bold are used for training the model to be tested here. Each row in the table reports testing results on the target database and the underlined numbers are for the entire target dataset. For example, when trained on LIVE and tested on TID2013, DTNIQ-f gives SROCC 0.530 over entire TID2013 and SROC 0.921 over the four common distortion types.

We observe that the generalization capability of our models overall is significantly better than BRISQUE and BLIINDS2 where quality-aware features are carefully designed instead of automatically learned. The same observation is made for the four common distortion types, except for models trained on TID2013 and tested on LIVE as a 1-tail 2-sample t-test indicates DTNIQ-f is not significantly better than BRISQUE.

We also observe that models trained on LIVE generalize better than the other two datasets over LIVE in the Wild. It implies distortions in LIVE share more similarity to the authentic distortions in the LIVE in the Wild, although TID2013 includes 24 different types of synthetic distortion. This aligns with the analysis in Bianco et al. [102] that distortions in real-world images are challenging to no-reference IQA modeling because of

the complex mixture of noise, blurriness, and exposure instead of the diversity of distortion types.

Table 5.2 Cross Dataset Performance
(Median Value of 100 Trials in SROCC Term)

LIVE	CSIQ	TID2013	LIVE in the Wild
DTNIQ-f	<u>0.689</u> 0.823	<u>0.53</u> 0.921	<u>0.587</u>
BRISQUE	<u>0.665</u> 0.661	<u>0.413</u> 0.863	<u>0.333</u>
BLIINDS2	<u>0.535</u> 0.525	<u>0.341</u> 0.825	<u>0.168</u>

TID2013	LIVE	CSIQ	LIVE in the Wild
DTNIQ-f	<u>0.818</u> 0.821	<u>0.638</u> 0.674	<u>0.307</u>
BRISQUE	<u>0.797</u> 0.819	<u>0.623</u> 0.64	<u>0.207</u>
BLIINDS2	<u>0.782</u> 0.807	<u>0.512</u> 0.515	<u>0.156</u>

CSIQ	LIVE	TID2013	LIVE in the Wild
DTNIQ-f	<u>0.909</u> 0.914	<u>0.526</u> 0.886	<u>0.413</u>
BRISQUE	<u>0.869</u> 0.886	<u>0.481</u> 0.8	<u>0.291</u>
BLIINDS2	<u>0.837</u> 0.865	<u>0.276</u> 0.73	<u>0.087</u>

LIVE in the Wild	LIVE	CSIQ	TID2013
DTNIQ-f	<u>0.681</u> 0.641	<u>0.576</u> 0.648	<u>0.426</u> 0.769
BRISQUE	<u>0.367</u> 0.32	<u>0.282</u> 0.267	<u>0.179</u>
BLIINDS2	<u>0.573</u> 0.607	<u>0.163</u> 0.219	<u>0.242</u> 0.528

Using models trained on LIVE, Kang et al. [51] report the generalization ability of their CNN-based no-reference IQA method. Their model performance is 0.920 in SROCC term when tested on the common four types of distortion on TID2008. When tested on the common four types of distortion on TID2013, Our DTNIQ-f model trained on LIVE gives 0.921 in SROCC term. TID2008 was revamped into TID2013 with seven more distortion types. The common four distortion types were the same on both TIDs except that the quality levels are increased from four to five. Based on the data available and the commonality of the TIDs, it is reasonable to expect that generalization capability over

four common distortion types is not differentiable between these two CNN-based methods.

5.4.4 Computational Cost

The experiments are carried out on a Linux box with 3.4GHz CPU and GTX1070 GPU. Taking 10 samples of size 256-by-256 pixels from each image, we measure the testing process time. For DTNIQ-f model, the average quality evaluation time per image is 0.233 seconds. And the average time is 0.341 seconds for DTNIQ-r model. This time is faster than three other NR IQA methods we propose in Chapter 4. See table 4.8, NRCNN takes 1.341 seconds and NRSTATS and NRGDD take more than 2 seconds.

5.5 Summary

In this chapter, we introduce the design and implementation of a no-reference image quality assessment framework. We show that the knowledge learned from high-level image-understanding task, object recognition specifically, can be effectively transferred to tackle low-level image quality assessment. Furthermore, we show that recurrent neural network leads to better model performance. Our algorithms achieve state-of-the-art quality prediction and generalization capabilities in terms of correlation with the human visual system.

Chapter 6

Towards Resilient Classification Through Image Quality Assessment

The deep convolutional neural network is capable of solving complex image problems beyond the reach of traditional machine learning methods. Its superiority inspires many academic works and industrial products that push the performance of computer vision applications closer to the capacity of the human visual system. One common challenge in developing these learning-based applications is that the data used during model training does not cover all conditions and the entire population of the task domain. For example, various kinds and magnitudes of input signal degradation could exist in cases of real-world use. The potential for adversarial attack is another emerging issue. In order to tackle these challenges at the input point, data augmented techniques such as benign noise injection and adversarial training have been investigated.

In this chapter, we propose a method that combines image quality assessment with knowledge distillation to detect input images that are not fit for object classification either because of poor image quality or adversarial problems. Using the ResNet [3] pre-trained model as an example, we demonstrate that the proposed method can effectively detect and discard distorted images and adversarial images that otherwise would lead to a wrong classification result by the convolutional neural network.

6.1 Background

Following the groundbreaking leap of deep convolution neural network (CNN) in

object classification tasks [20], numerous advances have been made in the network architecture [18, 23], training and transfer learning techniques [108, 109]. Many researchers are motivated to apply CNN to tackle various visual pattern recognition problems. For example, object detection and localization [28], image and video caption [41], facial recognition [97], and more. CNN-based applications have flourished in recent years and in various areas including medical imaging diagnosis [110], autonomous driving [111] as well as many other industry and consumer sectors. Even crossover work in computational art generating novel visual patterns even exists [33].

CNN's near human-level performance in terms of accuracy naturally raises questions such as why it performs so well; what is precisely learned inside each neural layer; and whether it is as robust as human vision when quality degradation is present in images.

Image quality assessment (IQA) [79] as a research topic has existed since the beginning of the image processing discipline. The types and magnitude of image quality distortion, as well as their impact on visual quality perceived by the human visual system have been extensively studied. Numerous IQA algorithms have been developed and have found their way in many image processing and multimedia applications such as denoising and super-resolution. In very recent years, studies have emerged around the effect of image quality on CNN performance. CNNs trained on ImageNet suffer non-trivial performance degradation [112, 113] when images are distorted with a non-trivial amount of blur or noise. Significant performance degradation also occurs in deep CNN-based face recognition model [114] trained on popular face datasets. Hosseini et al. [115] report that Google's Cloud Vision API is also vulnerable when fed with noisy images.

The investigation on CNN reveals its vulnerabilities and shows many methods that can generate adversarial images to elicit utterly wrong results from CNN. Szegedy et al. [116]

is the first in the literature to demonstrate that when given a clean image, iteratively exploiting computed gradient and perturbing pixel values in a range that is indistinguishable to the human eye can fool CNN to misclassify a school bus as an ostrich. Using evolutionary algorithms, Nguyen et al. [106] show that CNN can be easily fooled into making highly confident predictions on images that are unrecognizable to humans. Very recently, Moosavi et al. [107] propose a method that can produce a single perturbation map from a trained model, dubbing universal perturbation map. Using several popular deep CNN networks, the team shows that CNN is vulnerable to fooling images perturbed with such maps. According to Lu et al. [117], this universal perturbation method is one of the strongest adversarial methods so far in the literature. Others look into the physical world for potential threats to CNN-based application. Karahan et al. [114] find that the ImageNet Inception classifier [118] misclassifies a significant percentage of images obtained from cell-phone cameras. In [119], Sitawarin et al. take a black-box approach and creates adversarial samples by distorting road sign images with lenticular printing and background injection techniques. Their adversarial samples successfully deceive auto-driving cars.

We find that a significant percentage of adversarial images generated with methods in the literature are noticeable to the human eye, when viewed under conditions that have been adopted in the IQA research field. Most of these alternations bear some resemblance to various kinds of distortion in the color image quality dataset TID2013 used to assess and compare visual quality metrics and algorithms. The visual qualities of some of these images are not fit for carrying out any object detection task.

Fig 6.1 Sample Images That Fail to Fool Proposed Resilient Image Classification

Images with orange color border are detected by the proposed method as not fit-for object recognition task. Images with green border are classified as fit-for-task. Original images in the top and bottom rows are from the validation set of ImageNet. Top row images are perturbed with the universal perturbation map [107] and classified by ResNet-152 in left to right order as goldfish, vending machine, sock, and bow tie. Middle row images are examples from [106] and classified by ResNet-152 in left to right order as honeycomb, traffic light, accordion and water snake. Bottom row includes images injected with blur or Gaussian noise at different distortion levels and classified by ResNet-152 in left to right order as whiskey jug, beacon (true label), tree frog and zebra.

Our study shows that the object recognition capacity of CNN drops when the quality of input images degraded. Image quality is often ignored in visual recognition challenges such as PASCAL VOC [25] and ILSVRC [12], where most images, if not all, used for

model training and validation are recognizable or around average quality. However, in reality, vision applications could receive images of far below average quality. For example, an auto-driving car operates in an extremely hazy day or when its image acquisition lenses are dirtied. Such scenarios pose practical challenges to the design of robust vision applications.

By gauging adversarial and poor quality images with a visual quality metric, we propose a method to detect and reject images that are not fit for the object classification task. One of the uniqueness of the proposed method is that it is non-intrusive as per the object classification system. We do not alter the structure of the object classification CNN and its trained weights. Instead, we conjugate the image quality score with the probability vector from the object classification system to form an auxiliary network that is capable of detecting input images that are not fit for the object classification task, i.e., the system would predict a wrong object class with high probability if not stopped.

In the rest of the chapter, related works are reviewed first, followed by a correlation study between object classification performance and image quality. The proposed fit-for-task detection methods are explained in section 6.4, followed by experiments and analysis in section 6.5. Finally, conclusion and future work are discussed in section 6.6.

6.2 Related Work

On protecting CNN against adversarial attacks, distinct approaches have been explored and demonstrated by many researchers. One direction is modifying the images by injecting small noise and minor adversarial perturbation into clean data and then training or re-training the CNN with both original and processed data. It is also shown training

with compressed images mitigates the adversarial impact to some degree. Another direction is modifying the CNN structure by adding extra layers to smooth out small adversarial noise or designing regularization formulas to increase the stability of neuron activations. This line of approach has been shown to be effective on smaller networks trained with MNIST and CIFAR datasets but hard to re-train deep networks trained on the entire ImageNet database. A third direction is attaching an auxiliary network to the intermediate layers or output layer of the CNN and using these layers' output as features to learn a detector that discriminates adversarial images from clean images. Our work is closer to the third direction in that the proposed method also utilizes auxiliary networks. Related works in this direction are discussed below.

Quantizing the output from the late stage ReLU layers of VGG-19 and ResNet-32 into discrete input features, Lu et al. [124] train an RBF-SVM as an adversary detector and reports performance on four types of adversarial attacks including DeepFool [125]. It also shows that adversarial manipulations in the depth map of RGBD images can be detected.

Collecting convolutional layers' output from VGG-16 and AlexNet and then processing them with PCA analysis, Li et al. [122] extract the statistics from normalized PCA coefficients as input features to a sequence of base classifiers which form a cascade SVM. The SVM is trained with 2000 L-BFGS adversarial images generated from the original images of the ImageNet validation set and a subset of the original images. The adversarial detection performance is tested with 2000 L-BFGS adversarial images and model generalization capability is tested using the fooling images from Nguyen et al..

Without requiring adversarial images, Meng et al. [126] use an autoencoder network to learn and encode representations of clean images. Their detector will reject a test image if

its reconstruction error is significant or the probability divergence between the target classifier and the encoded representation is large. The authors conclude the detector performs well over MNIST and CIFAR-10 images perturbed by DeepFool.

Metzen et al. [123] propose to extend a smaller subnetwork from ResNet-32 as an adversary detector. The detector itself is a CNN and trained with adversarial images in various degrees of perturbations. Using images from the CIFAR-10 and relevant ten object classes from the ImageNet, the authors conclude the detector has strong discriminative power over tiny perturbations even almost invisible even to humans.

Xu et al. [127] propose a detection framework by measuring the difference between object classification result of a given image and that of its reduced images using reducing color bits and smoothing processing. Detection performance of eleven attacks over six types of adversarial methods is reported over MNIST, CIFAR-10 and a small portion of images from ImageNet validation set.

Liang et al. [128] propose a detection strategy to compare the difference between the classification results of a given image and that of its reduced images is exploited as well. The reduced images are produced using scalar quantization and smoothing processing. The model performance is reported on the MNIST and three object classes from ImageNet only.

On the impact of image quality to the performance of CNN, Dodge et al. [113] investigate the performance degradation on object classification task over two types of distortion Gaussian blur and additive Gaussian noise. In [114], Karahan et al. study the decrease of performance in the face recognition task over six kinds of distortion. Hosseini et al. [115] find Google's Cloud Vision API produces different classification given the

same image but injected with noises of various degrees, and conclude denoising the image before feeding it to the API mitigates the problem.

To the best of our knowledge, there is no work in the literature that connects image quality factors to adversarial attacks and proposes a holistic solution tackling both at the same time. The main contributions include:

- From an image quality assessment perspective and using semantic distance, we thoroughly study the impacts of four common types of distortion injection and three types of adversarial manipulation to object classification outcome. We also consider that the semantic distance is a better metric in the classification context because the information on how wrong the result produced is encoded. Wu-Palmer [129] metric computes a catamaran is 0.88 similar to a trimaran and 0.50 similar to a safety pin. See appendix B for more examples.
- Experimentally we show the characteristic signal shared between distortion injections and adversarial perturbations and propose the image quality as a useful feature component for adversarial detection.
- Using IQA model, state-of-the-art object classification CNN trained on ImageNet, and experimental data covering 1000 object classes from ImageNet, we demonstrate that the proposed method can effectively detect poor quality images and protect the object classification CNN from the adversarial methods tested.

6.3 Image Quality And Classification Performance

Various factors could have an impact on the performance of a CNN when deployed in

applications. In this section, we present a comparison study that serves as a foundation for the subsequent proposed method. Firstly, the methodology of the investigation is described. Secondly, the experimental findings of the correlations between the performance degradation of CNN and quality distortions and malicious manipulations are discussed.

6.3.1 Methodology

Traditionally, image quality is systematically studied in types and degrees of distortion [81-83]. Many imaging processing techniques are applied to pristine images to produce distorted images. In a laboratory or semi-controlled environment, the distorted images are presented to human observers to evaluate the visual quality, with or without the pristine image aside. The quality scores collected from human observers are consolidated into a subjective mean opinion score (MOS) for each distorted image.

Designing computational models to estimate the perceptual quality of image and video is the core of image quality assessment. Based on whether the high-quality reference image is available or not, an IQA algorithm is either of full-reference, reduced-reference, or no-reference. In Chandler et al. [79]., well-known full and reduced reference IQA algorithms are discussed. Most no-reference IQA algorithms are machine learning-based. At first, image signal processing experts manually design a quality-aware feature extraction scheme. Subsequently, machine-learning networks are designed to train a quality prediction model using the features produced by the extraction schema.

Several no-reference IQA algorithms [131,132] design the feature extraction scheme by combining human visual system inspired features with free-energy based distortion metrics but deploy different learning methods to build the final quality evaluator. Gu et al.

[131] use support vector regression with non-linear kernel. Liu et al. [132] propose a 4-layer learning network with three hidden neural layers. Exploiting the statistics of natural scenes and their divergence due to quality degradation, Saad et al. [83] devise a scheme that first computes coefficients of a discrete cosine transform from pixel intensity in the spatial domain and then fits a general Gaussian distribution (GGD) statistics from the coefficients. Mittal et al. [69] take the same approach, but histograms of pixel intensity are the underlying signals. Both use the support a vector machine to train the final quality evaluator. Identifying the lack of the connection between chromatic information and quality distortion in precedent IQA development, Liu et al. [133] propose a scheme that extracts chromatic features from SCIELAB color space besides the classical human visual system inspired features and builds quality evaluation models with support vector regression, neural network, and Adaboosting approaches. Zhang et al. [134] propose a feature scheme that replaces the manual designing with sparse filtering-based learning. Firstly, a filtering dictionary that maps an image to quality-aware feature representation is learned. Then the random forest is deployed to acquire a quality model from the transformed feature.

In this study, we use a learning-based no-reference IQA algorithm DTNIQ-f by Feng et al. [120] to obtain the visual quality score of distorted and adversarial images. For investigation, we collect 4966 distorted or pristine images from the traditional image quality research literature, 5000 unrecognizable fooling images from [106]; using color images from ImageNet validation set and the universal perturbation [125] method produces 49,056 recognizable adversarial images, and using gradient ascent method produces 3000 unrecognizable adversarial images.

Firstly, the object recognition results of 4966 images from IQA datasets [27,36,37] are produced using DCNN models trained on the ImageNet. The semantic distances of results between distorted and reference images are calculated in Wu-Palmar term [35]. Using semantic similarity, we measure the impact of image quality to DCNN. From Table 1, we see AlexNet, VGG-16 and ResNet-152 are all affected. To CSIQ dataset, the mean similarity drops to 0.799. See appendix B for imagery examples about semantic similarity measured in the Wu-Palmar term. Nevertheless, the performance degradation of the three DCCN models is on par. Consider the ResNet-152 is more widely used in the recent literature we use ResNet-152 only for the subsequent analyses and experiments.

Table 6.1 Performances of Networks Over IQA Datasets

		Median	Mean	Std.
Images of TID2013	Subjective MOS	0.511	0.497	0.138
AlexNet	Wu-Palmar Similarity	1.000	0.825	0.260
VGG-16		1.000	0.874	0.191
ResNet-152		1.000	0.894	0.196
		Median	Mean	Std.
Images of CSIQ	Subjective MOS	0.677	0.649	0.263
AlexNet	Wu-Palmar Similarity	1.000	0.800	0.257
VGG-16		1.000	0.826	0.249
ResNet-152		1.000	0.799	0.263
		Median	Mean	Std.
Images of LIVE	Subjective MOS	0.426	0.490	0.300
AlexNet	Wu-Palmar Similarity	0.800	0.767	0.251
VGG-16		1.000	0.851	0.225
ResNet-152		1.000	0.835	0.238

The subjective quality scores of TID2013 are in MOS ranging from 0 to 9. Difference MOS (DMOS) is used in LIVE and CSIQ with ranges from 1 to 0 and from 100 to 0 respectively. For comparison, the original scores in these datasets are normalized to MOS

from 0 to 1 in Table 1. The MOS scores predicted by the IQA model are normalized to [0,1].

6.3.2 Image Distortions

The visual signal in an image and video applications are subject to many kinds of distortions before being presented to human or machine viewers. The distortions could be introduced during the signal acquisition, transforming, compression, transmitting and restoring phases. Inarguably the poorer quality the input image, the less accurate object classification will be [114]. We are curious as to 1) whether there is a simple linear coloration between the image quality and the CNN performance in terms of accuracy, and 2) given images of similar quality, whether there is a type of distortion impacting CNN performance less.

Analyzing the classification outputs from the ResNet-152 on 3000 distorted images from TID2013 by distortion types, several distinct patterns can be observed from Fig. 6.2. The object classifier performance is fairly robust to images distorted by masked noise (MN), none-eccentricity pattern noise (NEPN), the mean shift of intensity (MS) and a few more, as their kernel density estimation (KDE) plots indicate that the semantic distance of these distorted images to pristine image hold close to 0.9 while their visual quality varies from poor to OK quality, i.e., from 0.3 to 0.7 in predicted MOS. On the other hand, a CNN model is highly sensitive to local block-wise distortion (BLOCK) and change of color saturation (CCS), as their KDEs stretch down to 0.3 along the semantic similarity axis while the image quality varies within a smaller range.

From Fig. 6.2 we can also make a qualitative observation that the positive correlation between CNN performance and input image quality is not linear given that the performance degradation of CNN responds differently to different types of distortions. Remember that these distorted images are generated from the same set of pristine images.

Fig 6.2 Correlation Between Image Distortion and Semantic Dissimilarity

KDEs of Wu-Palmar similarity of TID2013 distorted images to their pristine images versus quality in MOS term scored by human observers. The subplot legend is the abbreviation of distortion type, e.g., AWN stands for additive Gaussian white noise, NEPN stands for none-eccentricity pattern noise.

6.3.3 Adversarial Examples

Szegedy et al. [116] show that hardly observable perturbation of image pixel values can cause a CNN to produce a ridiculously wrong classification result. Several studies have been done to fool CNNs. On the flip side, methods defending CNN against such potential attacks are emerging. These two research directions are forging an arms race dynamic. From the quality of the fooling images that are generated using gradient ascent method and the mean map in Nguyen et al. [106], an immediate intuitive observation arises: could image quality be a useful feature to block CNN from being fooled or attacked?

In [120], we lock down the weights of the ResNet-152 that is pre-trained on ImageNet. We then apply transfer learning to build no-reference IQA models. Using the DTNIQ-f IQA model trained with TID2013 dataset, we evaluate the quality of the unrecognizable and recognizable adversarial images in predicted MOS. Two types of unrecognizable adversarial images are assessed. We collect 5000 fooling images from Nguyen et al. [106] that are produced by the evolutionary algorithm and generate another 3000 fooling images using the gradient ascent method and a mean map. For recognizable adversarial images, we apply the universal perturbation map of ResNet from Moosavi et al. [107] to perturb the 50,000 original images of ImageNet validation set. Excluding gray images and failed to perturb, a total of 49,056 recognizable adversarial images are generated. Gray and images that are too small, e.g. 90 by 90 pixels are excluded. A few such recognizable adversarial images are shown in the top row of Fig. 6.1. The overall quality of these adversarial images is significantly poor as shown in Table 6.2, statistically lower than the three image quality datasets as shown in Table 6.1. The average quality of perturbed images is statistically lower than their original versions as well.

Table 6.2 Visual Quality of Adversarial and Original Images

Number of Images	Adversarial Method	Objective MOS		
		Median	Mean	Std.
5000	Evolutionary Algorithm ^[106]	0.271	0.289	0.122
3000	Gradient Ascent Method	0.271	0.271	0.018
49056	Universal Perturbation ^[107]	0.357	0.383	0.074
Origin ImageNet validation set		0.429	0.454	0.106

To this end, we propose to take the image quality into consideration as a useful feature component in seeking methods that defend CNN from performance degradation and mitigate the probability of being fooled or attacked.

6.4 Fit-For-Task Approach

When visibility is low, a human driver will instinctively brake or slow down under normal circumstances. It is logical to extend this behavioral pattern to machine vision systems for their own safety and their human users. The experimental data and analyses around CNN performance and input image quality lead us to hypothesize that visual quality is a useful feature to be exploited for CNN-based applications to detect input signals that are not fit for carrying out object recognition tasks, disregarding the intention behind the distortion of the images.

In supervised learning, a CNN-based vision model is trained and validated with a cohort of meaningful images. Given images utterly unrecognizable to human eye, there is no value for a machine vision system to strive to recognize an object from it. From the statistics in Table 6.2, a simple threshold-based baseline can be constructed. For the fooling images generated from a mean map, the mean and standard deviation are 0.271 and 0.018 respectively in predicted MOS. If set a quality threshold at 0.3 in predicted MOS, 94.6% of fooling images in this catalog can be detected and discarded, if the distribution of the visual quality scores was a perfectly normal distribution.

Applying this threshold on a CNN model trained on ImageNet, testing on the validation set shows that 2.44% images are below the threshold and would be rejected as not fit for the object classification task, including tiny images of 90 by 90 pixels. Examples of such images are shown in Fig. 6.3.

However, when the visual quality of input images is better than extremely poor or objects inside are recognizable to human eye, say around 0.429 in predicted MOS, the

statistics in Table 6.2 tells us that this fixed-threshold baseline detection method would reject about 50% of original images in ImageNet validation set.

Studying the prediction confidence scores from the ResNet-152, significant statistical differences can be observed between the original and adversarial images. The median value of the top-1 confidence prediction score of the original images is twice as high of the adversarial images but its standard deviation is smaller. See table 6.3.

Table 6.3 Classification Confidence Score

Category of Images	Top-1 Confidence Score		
	Median	Mean	Std.
Evolutionary Algorithm ^[21]	0.465	0.507	0.261
Universal Perturbation ^[22]	0.452	0.508	0.292
Original ImageNet Val. Set	0.929	0.800	0.244

Distilling knowledge from the confidence score is first proposed by Hinton et al. [130], in which prediction probabilities from an ensemble of neural network models are distilled into a single model, and the performance of this single model is improved significantly. Our approach is to combine the quality score from the IQA model with the classification probability score from ResNet-152 aiming to improve detection power. The prediction probability vector from ResNet-152 is in 1000 dimensions, but most are close to zero probability residual, thereby very limited knowledge to distill. We take the probability values of the top 5 predictions of an image together with its predicted quality score to form an input feature of six dimensions. Using the t-SNE technique from Van der Matten et al. [85], we visualize these input features in the left image of Fig. 6.3. The gray dots represent features learned from the original images in ImageNet validation. The red dots represent features learned from fooling images produced by the evolutionary method in

[106]. We can see most of the red dots are well separated from the gray dots, which infers it is hopeful to use SVM learning a useful detector.

Fig 6.3 Fit-For-Task Approaches - Analysis

The left column is a t-SNE visualization of input features for the fit-for-task classifier. Middle and right columns are images from the ImageNet validation set. The objective MOS values of both images are below 0.3.

The proposed methods can easily be extended to object detection and localization segmentation tasks where the input features for the fit-for-task classifier would be extracted from the candidate proposal regions instead of a whole image.

6.5 Experiments And Analysis

6.5.1 Data

A set of experiments is designed to measure how accurately the proposed approach would accept input images that ResNet-152 will correctly recognize objects in it and reject images that will trigger ResNet-152 to make a mistake. Four categories of experimental data are collected or generated:

- Original images: The top-1 and top-5 classification accuracy of the entire ImageNet validation set is 75.92% and 92.83%, respectively, by the ResNet-152 model trained on the ImageNet. Only original images that are correctly classified by ResNet-152 per the top-1 prediction probability are selected, which leaves 37,962 original images in total.
- Distorted images: four types of distortion in five levels are applied to the above selected original images. The distortion types include Gaussian blur, white noise, and JPEG and JPEG2000 compression noise. A total of 759,240 distorted images are generated from the above selected 37,962 original images. Their predicted quality scores are from 0.071 to 0.729 in MOS term. The semantic similarity of object recognition outcome of these distorted images to corresponding clean images are computed in Wu-Palmer term. The correlation between the predicted image quality and the Wu-Palmer similarity is shown in columns 2 to 5 of Fig. 6.4.
- Recognizable adversarial images: Using the universal perturbation map of ResNet-152 from [125], 37,962 adversarial images are generated using the above selected original images. The adversarial target of each sample is chosen randomly from the 999 classes excluding the true object class. The coloration between the semantic similarity of this group of images and their predicted quality is shown in the 1st column of Fig. 6.4.
- Unrecognizable adversarial images: 3000 images are generated from a mean map using the ascent gradient method, we denote this groups as Fool_GD. 5000 images are collected from [106] that are produced using the evolutionary algorithm, we denote this group as Fool_EV. See examples in the 4th row of Fig. 6.5.

Fig 6.4 Predicted Image Quality of Adversarial and Distorted Images

6.5.2 Evaluation

We use ten DTNIQ-f IQA models trained with the TID2013 image quality dataset to predict MOS scores of the distorted and adversarial images. The performance is reported in an average value of the ten detectors. For distorted images one IQA model is selected randomly from the ten IQA models to produce MOS scores, i.e., the ten detectors are trained with a single set of features. For recognizable distorted and adversarial images, a baseline of classification results from ResNet-152 is established first. The detection rate and the classification accuracy of accepted images are compared with the corresponding baseline.

The detectors are trained using a linear SVM classifier with the value of parameters C that controls the classification margin set to 0.5. All experiments and data generation are carried out on a computer equipped with a GeForce GTX 1080 Ti and using the Caffe platform, with exception to recognizable adversarial images which are generated on the TensorFlow platform.

6.5.3 Baselines

Not all attacks are successful, and not all distortions lead to incorrect object classification outcomes. We define the test result without quality awareness as baseline-1 and test result with the threshold-based quality filtering as baseline-2. The detection rate and the object classification accuracy of images accepted by the detector are calculated to analyze the performance.

Table 6.4 summarizes the test results. Note the corresponding original images that are used to produce these four image groups are all correctly classified by ResNet-152. In baseline-2, 96.7% of Fool_GD and 54.11% of Fool_EV images are detected and rejected. And 48.16% distorted images are detected, which leads to the object classification accuracy of accepted distorted images increases from 57.04% in baseline-1 to 73.04% in baseline-2.

Table 6.4 Baselines of Fit-for-Task Detection

Image Group	Baseline-1 (Without quality awareness)	Baseline-2 (With quality awareness)	
	Object Cl. Accuracy	Detection Rate	Object Cl. Accuracy
Adversarial	55.42%	4.21%	54.69%
Distorted	57.04%	48.16%	73.04%
Fool_EV	-	54.11%	-
Fool_GD	-	96.70%	-

Baseline-2 detects only 4.21% of the recognizable adversarial images. If increasing the quality threshold from predicted MOS 0.3 to 0.35, 47.25% of them can be detected. However, the accuracy of ResNet-152 over fit images doesn't increase, instead drops from 73.04% to 72.45%. This indicates that the discriminative power of quality feature alone is weak over the recognizable adversarial images.

6.5.4 Performance of Fit-for-Task Detector

Next, we show that by distilling object classification information from ResNet-152, the detection power increases significantly. With input features generated from four groups of images, many different training schemas are possible. However, our primary goal is to improve the detection capacity over the recognizable adversarial images in this experiment. After some trials, we decide to take only the recognizable adversarial images and their corresponding original images to learn a binary detection classifier.

50% of recognizable adversarial images and their corresponding original images are selected randomly to form a training set, which ensures an adversarial and its original image appear together in either the train or test set. The rest 50% is held out for testing. This sampling method is repeated ten times to train ten detection classifiers. Table 6.5 summarizes the test result.

We observe that 65.95% of recognizable adversarial images are detected, and the object classification accuracy of accepted images increases from 54.69% in baseline-2 to 85.22%. Also, more distorted images are detected, and as such the object classification accuracy of accepted images increases from 73.04% in baseline-2 to 92.36%, although the distorted images are not used to train this detection classifier. The result indicates that there is indeed useful information residing in the prediction probability vector of the object classifier. For the Fool_EV group, the detection rate increases from 54.11% in baseline-2 to 78.29%. Note the Fool_EV images are not used to train this detection classifier as well. The average detection rate of the detection classifier is 77.42%, and the average object classification accuracy of images accepted by the detector is 88.79%, and 0.932 in Wu-Palmar similarity.

Table 6.5 Performance of Detector

Image Group	Detection Rate	Object Classification	
		Accuracy	Wu-Palmar Similarity
Adversarial	65.95%	85.22%	0.912
Distorted	65.81%	92.36%	0.951
Fool_EV	78.29%	-	-
Fool_GD	99.64%	-	-
Average	77.42%	88.79%	0.932

Examples of adversarial images that the detection classifier succeeds or fails to detect are shown in Fig 6.6. Predicted image quality scores of adversarial examples are from 0.143 to 0.7 in MOS term. Adversarial examples in rows 1 and 2 are successfully detected and rejected, otherwise would be wrong objects would be classified. The left two adversarial examples in row 3 are falsely rejected otherwise correct objects would be recognized. The right two adversarial examples in row 3 are not detected and successfully fool the object classifier. The actual object labels of the 3rd and 4th images are banana and goldfish, but are recognized as pineapple and brain coral. The fooling images in the 4th row are successfully detected.

Will the detection classifier wrongly reject clean images? Tests with the original image group show that the baseline-2 accepts 97.17% of these clean images and the detection classifier accepts 91.43% of them.

Fig 6.5 Adversarial Examples

6.6 Comparison and Discussion

The defense against adversarial work in the literature look into adversarial examples only. Therefore, most of these works validate their methods against adversarial examples but don't test the false detection rate over clean images. For example, the cascade SVM in Li et al, [122] achieve an overall 97.34% detection accuracy over the Fool_EV images. However, the potential misclassification rate is not investigated. Using 1000 MNIST samples and 100 random samples from the ImageNet, Liang et al. [128] conclude that the cascade SVM produces a 92% false positive.

Most of these works use images from the MNIST and CIFAIR-10 to validate their defense methods. If images from ImageNet are involved, only a small portion is selected. For example, Xu et al. [127] test 5000 images from ten object classes from ImageNet that are relevant per CIFAIR-10. Liang et al. [128] validate 1489 images from ImageNet that belong to three object classes (zebra, panda, and cab). Why these particular three object classes are selected is not explained.

We validate proposed methods over 805,202 adversarial or distorted images, and 37,962 original images from the ImageNet validation set, covering 1000 object classes in the ImageNet. It is hard to run a direct performance comparison with other defense methods at this massive scale. Besides adversarial images yet perceivable to human eyes per IQA standard, we are interested in detecting poor quality images too. Alternatively, we construct two baselines and compare the performance of the proposed detection classifier with these baselines. Compared to the baselines, the detection classifier can fairly well protect the object classification CNN from making wrong predictions when input images are of poor quality or maliciously altered with the three attacking methods tested.

Examining the unrecognizable adversarial images, we observe that these images do not have a meaningful composition from a photography standard or distinctly different composition compared to natural images. The composition of an image impacts the visual distinctness of objects in an image, including the space and size of the foreground object, the complexity of background texture, contrast and many more. Analyzing the classification results of pristine and distorted images from the image quality datasets, we notice a couple of images that stand out in such a way that Wu-Palmar similarity of all their distorted images equals 1, i.e., ResNet-152 classifies all their distorted images as the same object class as corresponding pristine images. These pristine images are shown in Fig. 6.6 columns 1 and 2. Column 3 shows a sample of additive pink Gaussian distortion in MOS 0.449 per human observer, and column 4 is a sample of jpeg distortion in MOS 0.451 per human observer as well. This could be a potential direction to look into in the future.

Fig 6.6 Discussion on Perceptual Quality and Object Recognition

6.7 Summary

Although poor quality images are not malicious, they can drag the object classification performance down and pose potential harm in the same way as adversarial images. In this chapter, we identify the value from the traditional IQA field in tackling one emerging

challenge to rising deep CNN technologies. Experimentally we show that there is an intrinsic connection between the adversarial pixel perturbations and visual quality per the human visual system. We introduce the fit-for-task notation and propose non-intrusive methods to prevent the production of object classification results when the inputs are far below optimal, thereby increasing the robustness of CNN-based vision systems in real-world environments. Visual quality manipulation is relatively cheap to implement but it could also bring down the performance of a CNN network. We would like to conclude that the proposed methods could serve well as a first line security mechanism for CNN-based vision applications.

Chapter 7

Applications and Image Quality in Chest Radiology

Analyzing medical radiology images is a natural fit for deep learning. At the 2016 Machine learning and Market for Intelligence Conference, Geoff Hinton made a statement that we should "stop training radiologists now" [141]. Since then, the academia camp is making rapid progress with regards to the variety and performance of deep learning applications in radiology. Debates around whether AI should be seen as an opportunity or a threat have been intense among radiologists [142]. Every radiology conference in the world now holds a session on the AI topic.

Motivated by the great application potential in the radiology and its current status, the author wishes to expand work into this field, and beyond this research program. This chapter reports a body of work completed. Firstly, a literature review on the progress of machine learning applications in the chest radiology can be found in section 7.1. This work is published in a journal that reviews the most significant recent developments in the radiology. As the target readers are radiologists, this work could help them be better prepared when engaging in above mentioned debates in their professional community. Secondly, our explorative work on applying deep learning to detect chest x-ray images whose overall quality is not adequate for diagnosis purpose is proposed. Finally, potential future research work from here is discussed.

7.1 Deep Learning in Chest Radiology

The developing and applying of computer-aided detection and diagnosis (CAD) systems to analyze chest radiography and computed tomography (CT) images had been advancing rapidly over the past decades, in parallel with image signal processing and machine learning disciplines. These CAD systems heavily rely on filter designs to extract useful visual features from image samples. These filters usually have fixed parameter configurations except the neural filter, which is a three-layer neural network to learn parameters from image samples. Ginneken et al. [143] provide a comprehensive review that covers the literature of traditional CAD systems in the detection and classification of lung nodules and other tasks.

Backed by the massive-parallel, ultra-fast graphics processing unit (GPU) and advances in model training techniques, deep learning has distinct advantages over classical machine learning. For example, now it is feasible to implement convolutional computation over a whole radiograph and CT image in reduced size, Fig. 7.1 illustrates a whole-image deep learning paradigm. Now it is doable to train big networks (artificial intelligence models) with millions of parameters and layers beyond a hundred neural layers on advanced GPUs. Back in 2012 AlexNet [20], a watershed in deep learning, has eight layers and 62 million parameters. The AlexNet model recognizes 1000 different classes of objects in everyday life.

7.1.1 Background

Deep learning is a class of machine learning techniques. It inherits traditional learning forms and adopts new network architectures and training techniques. Given an application

problem and available data, there are several important design factors to consider for building an AI model: 1) supervised, unsupervised or other learning forms; 2) convolutional neural network, recurrent neural network or other network architecture; 3) Training the target AI model from scratch or using transfer learning. This section provides an overview of the mostly adopted choices in deep learning for chest radiology applications.

Supervised learning is the most common form of machine learning. In supervised learning, the data available are all fully labeled. For example, if the application problem is to build an AI model that picks out cases showing pulmonary abnormality from the chest radiography examinations, the data is a two-dimensional matrix of pixels representing the radiography images, and each case will be labeled as either normal or abnormal. Together with corresponding data, the labels are the ground truth that is fed to the AI model under training. The principal of a model training is that it is repeatedly adjusts the parameters (sometimes called weights) of all layers in the network through comparing the model prediction with the ground truth until the disagreement between the model prediction and ground truth cannot be further reduced. In deep learning, feedforward and backpropagation procedure is the state-of-the-art training technique. In [144], LeCun, the pioneer of backpropagation provides a historical overview and mathematical detail of this procedure.

The opposite is unsupervised learning, where labels are not available. In unsupervised learning, specific learning target, i.e. the ground truth is absent, the model attempts to discover useful structure from data only automatically. The learned structure is a representation of raw data in a transformed domain and can be used: to directly solve for example problems require a definite answer; or as input features for supervised learning.

Representation learning and clustering are unsupervised learning methods.

In between exists weakly-supervised learning in which the scope of prediction an AI model is tasked to produce is larger than that in the available labels, for example, training a model to classify disease and locate it in the image with labels including absent or presence of the disease only. Supervised learning and weakly-supervised are the dominant learning forms among the deep learning based solutions published in the literature on radiological image tasks.

In the image processing, convolution computation alters an image by adding the value of each pixel with its local neighbors weighted by a kernel. The convolved result is further computed with different filter functions for edge detection, blur, sharpen, etc. Fig. 7.2 shows an example of edge detection using Sobel operator. In traditional machine learning, task-specific filter operations will need to be carefully designed to extract useful features from raw images. In deep learning, same as CNN, this manual feature crafting process is replaced by automatically learn features from data. The convolutional filter operations are applied over multiple layers. In each layer, the kernel slides over the input data array and computes out a cubical feature map which serves as input to layer further down the forward path after some non-linearity and pooling treatment. The units between consecutive convolution layers are the filter banks. In a supervised form, the weights in these filter banks are learned from data and label fed to the network during the training phase. As shown in Fig. 7.1, at the bottom of a CNN network, is the application layer usually comprises of one or two fully connected neural layers and a prediction layer, for example, softmax function is a common choice for the classification task.

Fig 7.1 A Paradigm of Multi-labels Abnormality Detection with Whole Radiograph

In supervised learning, during the training error (the difference between label and prediction) is repeatedly computed and propagated back to correct the model parameters, i.e., learned features. The chest radiograph is from ChestX-ray14 [146].

CNN is the unchallengeable architecture for image understanding tasks so far, across application domains including radiology image. More technical details about CNN written for radiologists are explained in Chartrand et al. [145].

Today the recurrent neural network is a popular choice for tasks involving sequential input, for example, words and sentences. RNN was first introduced in 1986, but at the time there was no practical method to training it. In 1997 a variation of RNN long short term memory network (LSTM) showed that with a special hidden unit acting like a memory cell plus gradient-based backpropagation technique, it is capable to selectively retain relevant information from far previous step while the input sequence is being parsed element by element. Major technology companies have adopted LSTM in their speech recognition and language translation products. In the radiology field, we observe that RNN has been applied to mining text content from radiology reports.

Combining CNN and RNN, researchers in the computer vision field challenged problems that bridge visual and semantic understanding. In an influential work [147], Krause et al. train a multimodal RNN framework by feeding image and sentence pair to

the system. Given a new image the trained system generates description in a complete sentence. Later work from Krause et al. demonstrate that using CNN plus hierarchical RNN descriptive paragraph can be generated from images. See Fig. 7.3. Along with this line, we see that research work that can produce explanatory paragraph given a chest radiograph are emerging.

Fig 7.2 An Example of Edge Detection in a Lung CT Slice

Left is the kernel of a Sobel operator. The middle is a 512-by-512 lung CT image from the Kaggle DSB 2017 challenge [148]. Right is edge detection output, with contrast adjust for visualization. The Sobel operation convolves the kernel and its transpose with the image to approximate the derivatives horizontally and vertically, from which gradient magnitude and direction are calculated to derive edge.

Fig 7.3 Descriptive Paragraphs Generation with CNN+LSTM Framework

Examples are adopted from Razavian et al. [26]. The bounding boxes in the pictures are regions of visual features detected for sentence generation.

It is a popular approach in deep learning to start training a new model with a pre-trained model instead of initializing the parameters of the new model randomly. A pre-trained model is a model of same or similar architecture to the new model but trained with a distinct dataset for a related or different task. Through transfer learning the visual feature representations learned by a pre-trained model from a massive dataset including million samples is beneficial in several ways: reduced training time for the new model; potential performance enhancement in the new model and less training data may be prepared from the new task domain.

The benefit of transfer learning is consistently observed in situations with different learning tasks and diverse datasets. For example from object recognition to scene recognition in Razavian et al. [26], as well as across application domain in Shin et al. [149], from object recognition with natural images to interstitial lung disease classification with CT images.

Most works on chest radiography discussed in the next section adopt transfer learning and start training with a top performing CNN model from the past years ImageNet, the large-scale visual recognition challenges [12].

7.1.2 Chest Radiography

7.1.2.1 Common Thorax Diseases

ChestX-ray8 is a large-scale database made available to the public as part of the project work published by Wang et al. [146]. The database was soon expanded from eight common thorax diseases labels to fourteen labels and renamed to ChestX-ray14. It contains 112,120 frontal-view chest x-ray images extracted from the clinical PACS

database at the National Institutes of Health Clinical Center. The disease labels are mined from the radiological reports through natural language processing (NLP). The total quality of the disease labels in precision, recall and, F1-score terms are 0.90, 0.91 and 0.90 respectively. The non-perfect quality labels is a limitation that all studies using this dataset will inherit.

In [146], Wang et al. propose a deep CNN multi-label disease classification framework, and multiple models are built with four different pre-trained ImageNet models respectively. The model trained on a 50-layers ResNet network performs the best across all eight diseases in AUC-ROC curve (AUROC), except the mass class. Location of detected disease is estimated from global pooling maps.

Since the ChestX-ray14 dataset is made available applying deep CNN for chest x-ray analyses becomes active. Instead of using a pre-trained model, Yao et al. [150] design a two-stage framework for the disease detection task. Stage one is a deep CNN encoder to process the images, and stage two is an RNN decoder to exploit statistical dependencies between disease labels in the purpose to improve disease classification accuracy. The average AUROC of the fourteen diseases is 0.798 compared to 0.738 in [146]. Rajpurkar et al. [151] take a similar approach as [150] but use a pre-trained ImageNet model of 121 layers deep dense connection network (DenseNet) and an average AUROC 0.842 is achieved. Li et al. [152] propose an end-to-end jointly trained classifier and localizer. The backbone of the framework is a 200-layers deep ResNet pre-trained on ImageNet. Although the network is much deeper, the average AUROC reported in [152] is lower, at 0.81.

The detection power of AI models varies per disease class. For example, a detection power matching a group of 4 radiologists on pneumonia is observed in [151]. The

detection performance per disease class is summarized in Fig. 7.4 along with sample outputs of disease localization. Left is the performance of common thorax disease classification in AUROC term. Right top is a sample of localization of pneumothorax from [152]. Right bottom is a sample of localization of effusion from [146].

Fig 7.4 Detection and Localization of Common Thorax Diseases

7.1.2.2 Abnormality Detection

Deep learning based utilities that reach a clinic acceptable performance have been explored actively in the radiology community, with an overarching purpose of improving the entire clinical workflow beyond the node at radiologists.

In [153], Dunnmon et al. curate 216,431 unique frontal view chest radiographs from their institution and meticulously designs a training and testing schema: Each image is labeled manually as 1 of the six classes by attending radiologists with associated text report available. 20,000 such labeled images are randomly selected to train and validate the AI models. The normal, abnormal but not previously reported, and critical finding labels are consolidated into normal or abnormal to formulate a binary classification task

with a class balance of 79% abnormal. The test is carried out over 533 images drawn from the hold-out images and manually labeled by two expert radiologists with consensus. Several CNN models are trained either from random initialization or pre-trained ImageNet models. The best performance observed is 0.96 in AUROC and 0.91 in test accuracy. The authors conclude that the CNN based AI system potentially enables the triage of examinations in clinical practice.

Utilizing data and label available in public, Yates et al. [154] simulate a multi-sites scenario to study the normal or abnormal binary classification task. 51,760 frontal view radiographs with one or multiple disease labels from the ChestX-ray14 dataset and 1389 frontal view radiographs from the Indiana University hospital network dataset (IU X-ray) are assembled into a dataset with class balance 97.4% abnormal and 2.6% normal. 10% of this dataset is held-out for testing. The model, dubbed as red-dot, is trained from a pre-trained ImageNet model using Inception-v3 network. On test dataset, performance of the red-dot system reaches 0.98 in AUROC and 0.946 in accuracy, sensitivity and specificity are 0.946 and 0.934 respectively. The authors conclude that their work demonstrates that the application of such a binary normal or abnormal system to real-world datasets may be warranted in optimizing clinician workload.

In [155], Annarumma et al. build a four priority level classification system. Their purpose is to examine the feasibility of automated triaging of adult chest radiographs in real-time per urgency. 470,388 frontal view chest radiographs and reports are retrieved from three hospitals under the authors' institution, 10% is held-out for testing. The priority level labels are generated by a rule-based system that maps each of the 15 radiological tags into one of the four levels: critical, urgent, nonurgent or normal. The labels are mined from the radiological reports through an NLP system with a reported

accuracy from 0.81 to 0.99 in F1-score term. On the test set, the sensitivity and specificity for normality prediction are 0.71 and 0.95 respectively, for critical level prediction is 0.65 and 0.94 respectively. Through a simulation study, the team finds their AI triaging system could reduce reporting delay from 11.2 to 2.7 days for critical imaging findings and from 7.6 to 4.1 days for urgent imaging findings. The authors conclude that a real-time AI triaging system is feasible, with clinically acceptable performance.

7.1.2.3 Towards Automated Text Reporting

A chest radiograph report usually includes complete sentences that describe the impression and findings.

Deploying CNN-RNN architecture to generate descriptive text automatically is first attempted in Shin et al. [156]. However, the annotation produced is a sequence of discrete words or phrases, not a complete sentence. See an example in Fig. 7.5. Generating quality long sentences is first time shown feasible in Jing et al. [157]. The contribution of their work includes 1) a co-attention model that simultaneously captures the visual feature map and corresponding tags (critical findings) into one context vector; 2) a 2-level LSTM model, the first LSTM generates topics from the context vector and the second LSTM synthesizes paragraph from the topic vector. See Fig. 7.5 for examples. Wang et al. [158] propose a TieNet system that generates a text report and detects common thorax diseases simultaneously. Their contributions are introducing 1) a multi-level text-image attentions mechanism to provide more salient and meaningful embedding for the joint task; 2) an end-to-end trainable CNN-RNN model. Sample output of TieNet is shown in Fig. 7.5.

Automated report generation is promising and there is a big room to improve from here.

Fig 7.5 Towards Automatic Text Report Generation

In the top row, the left part is a chest x-ray image and its text annotations from [156], the right part is a chest x-ray image and its text reports from [157]. The bottom row is a chest x-ray image and its text reports from [158].

7.1.3 Chest CT

7.1.3.1 Pulmonary Nodule

Detecting and characterizing pulmonary nodules has been an active research area for the last two decades. A comprehensive review of this area up to 2016 is covered in sections 6 and 7 of Ginneken et al. [143]. In the LUNA16 challenge, it is observed that solutions employing 2D or 3D CNN networks for both detection and false positive reduction tasks achieve leading positions. In the Kaggle DSB 2017 challenge, the first place solution by Liao et al. [159] extend the fast object detection method (R-CNN) from natural image tasks into a 3D nodule candidate detector, subsequently extend the nodule detector with fully connected neural layers and a leaky noisy-or scoring method to predict the malignancy probability of detected nodule candidates.

Besides size and growth, nodule type is the key characteristics that the Lung-RADS guidelines adopt to define follow-up management in low-dose CT screening exams for lung cancer. In a 2017 Nature report, Ciompi et al. [160] propose a multi-stream multi-scale CNN system to detect and classify nodule into six classes that align with the Lung-RADS classification system. The team trains the deep learning system on data from one site and analyzes its performance on data from another hospital using Cohen κ statistics and accuracy per pair measures. It is observed that the performance of the proposed system is within the inter-observer variability among four experienced human observers.

7.1.3.2 Interstitial Lung Diseases (ILDs)

To accurately classify the pattern that manifested by ILDs most traditional CADx systems require firstly segment the lung then detect regions of interest before making a diagnosis. Without a segmentation stage, Anthimopoulos et al. [161] propose a 5-layer CNN network that classifies a 32x32 patch extracted from a CT slice without overlapping into one of the seven classes (6 common ILDs and healthy). The intensity values of CT images are cropped within the window [-1000, 200] in HU and then normalized to [0, 1]. Using the same datasets, they show that the specially designed network achieves better classification result compares to using AlexNet and VGGNet network. Gao et al. [162] propose a holistic approach that takes a whole slice image as input to the 7-class classifier. Each gray image is pre-processed into a 3-channel image based on the HU ranges that are selected in such a way that each channel focuses on the pattern of lower attenuation, higher attenuation or normal lung attenuation. An adaption of AlexNet is used to train a system model. Whole silica image classification is much more challenge then patch-based approach. The overall accuracy is 68.6%. In Wang et al. [146], the same

dataset with a close train/test split ratio is used to train several whole image classifiers, but without the HU attenuation rescale. A best accuracy 57% is observed on the GoogleLeNet pre-trained with ImageNet.

7.1.3.3 Segmentation

U-Net is a CNN based segmentation architecture first proposed by Ronneberger et al. [163] to segment the neuronal structures in electron microscopic stacks. As shown in the left column of Fig. 7.6 the network first contracts to encode context from input then symmetrically expands enabling precise localization. Juarez et al. [164] extend the architecture to 3D U-Net and proposes a simple end-to-end trainable network to segment airway tree from chest CT automatically. A dice coefficient of 0.8 is observed between the segmentation result and the ground truth. A sample segmentation result is presented in the right column of Fig. 7.6.

Fig 7.6 3D U-Net and Segmentation Task

Left column: A schematic diagram of the U-Net architecture [163]. Height and width of a solid column indicate the size and numbers of feature maps produced by convolution blocks. Left double line is input image and the right double line is output segmentation map. Right column: A visualization of airway segmentation output from [164]. True positive voxels are in yellow, false positive voxels are in red, and false negative voxels are in blue.

7.1.4 Other Applications

Utilizing the lung CT scan images from cancer screening exams, more types of applications are feasible through deep learning.

Using the chest CT scan images from the COPDGene study, González et al. [165] showcase an image-to-biomarker paradigm. As shown in Fig. 7.7, the framework has two stages: 1) CT images are pre-processed with object detectors to generate composite images suitable for bone mineral density (BMD) and lung percentage of emphysema estimations. The object detectors are designed with traditional machine learning approach. 2) The composite images are fed into a 4-layer CNN regression network to train the biomarker estimator. The performance of the estimators is measured against chosen reference standards in Pearson correlation coefficients. A correlation of 0.940 is observed for BMD estimation and 0.976 for the percentage of emphysema.

Fig 7.7 Other Applications with Lung CT images

Top row is a schematic diagram of the system proposed in [165] for the BMD and percentage of emphysema estimation. Images are adopted from [165, 167]. Outputs of the object detector are composite images. Bottom row are composite images from lung CT

scans for the percentage of emphysema estimation in [165]. From left to right the percentage estimated is 55%, 19% and 0.4% respectively. The corresponding ground truth is 55%, 25% and 1% respectively.

Using CT scans from the National Lung Screening Trial, Lessmann et al. [166] propose a two-stage CNN framework to automate cardiac calcium scoring: 1) the first CNN comprises three subnetworks to find candidate voxels of CAC, TAC, aortic valve and mitral valve from axial, sagittal and coronal views. Dilated convolutions (as explained in Fig. 7.8) are used in this stage with particularly large receptive fields; this enables the CNN to find candidates based on their anatomical location without typical segmentation and localization procedures. The second CNN is a binary classifier to detect calcifications from the voxels located in the previous stage. Using manual scores by one radiologist with extensive experience in calcium scoring as the reference, a best 0.89 F1-score is measured from the automated system for both CAC and CAT volume estimations during the test. Inter-observer agreement with two trained observers on CAC risk category assignment is high, a best 0.91 kappa coefficient is observed.

Fig 7.8 The Concept of Dilated Convolutions.

The grid represents pixels of the image fed to a CNN network. The red dot represents a 3-by-3 kernel. The green areas are the receptive fields that impacts parameters to be learned. From left to right are receipt fields of 1-dilated, 2-dilated and 4-dilated

convolutions. A big receptive field enables learning spatial context, but without dilation, it is computational infeasible today. The image is adopted from [168].

7.2 Image Quality in Chest Radiography

Good quality chest x-rays images allow human radiologists to extract maximum clinically useful information. On the other hand, poor quality images significantly increase false positive and false negative interpretations. The factors related to the quality of chest x-rays images include positioning, degree of inspiration, technical exposure factors and sharpness [179].

Designing medical IQA (MIQA) methods are studied, if not as extensive as for natural images. Hua et al. [180] propose a structural similarity based full-reference MIQA specifically for assessing quality of MR images. Jayageetha et al. [181] propose a CNN based no-reference MIQA and evaluate it with four publically available medical image databases. Instead of designing specific image modalities, Zhang et al. [182] explore whether off-the-shelf no-reference IQA methods designed for natural images can predict the quality of MR images, and conclude that BLIINDS-II [68] and BRISQUE [69] are full of potential to be re-purposed. Chow et al. [183] provide a comprehensive review of the development of MIQA in the past.

These IQA methods assess the perceptual quality degradation at the pixel level such as exposure, noise, blur, and contrast. They are useful to radiology professionals in quality assurance and clinical diagnosis. Factors such as positioning and degree of inspiration are very easy for trained radiographers to capture, but not algorithms. We argue that researching for computational methods that can assess all quality factors relevant to

diagnosis outcome is critically important for deep learning-based diagnosis applications to achieve high accurate and robust performance. And our explorative work in presented the next section.

7.2.1 A Projection View Classifier for Chest Radiographs

Using general chest x-rays as example, in a good quality image, the entire thoracic cage should be included, inspiration should be complete so that the lung bases are not obscured, adequate penetration of x-rays so that critical anatomy is visible, posteroanterior (PA) frontal view should be taken when possible. Fig. 7.9 shows three common projections in chest radiography. The left is a PA projection and the middle is an anteroposterior (AP) projection. It is obvious the AP images are less than optimal for diagnoses. The right is a lateral projection that may be taken as a complement to a frontal projection if required to answer clinical questions. The left and right images are from Demner-Fushman et al. [169] and the middle image is from Wang et al. [146].

Fig 7.9 Common Projection and View of Chest Radiography

Most of the deep learning applications for chest radiographs use frontal view images to train a model for abnormality detection or disease diagnoses (See sections 7.1.3.1 and 7.1.3.2). When in use, if the model is fed a lateral view chest radiograph, the chance that the model produces a correct prediction will be low. A projection view classifier can help

to flag out such a scenario. Preparation of a large-scale imaging database with labels is laborious intensive. It is inevitable that label error exists, for example, when lateral view chest radiographs are labeled as frontal view. A highly accurate projection view classifier could be used as a tool to quickly find possible mislabeling.

Using a 50-layer ResNet model trained on ImageNet, we replace its application layer with a binary classifier. 150 frontal view images are randomly selected from [146] and 137 lateral view images are randomly selected from [169] to form a training dataset. The model is trained 50 epochs using stochastic gradient descent, and the initial momentum is set to 0.9. 59 frontal and 59 lateral images is randomly selected from the same databases to form a validation set, without overlapping with the train set. The classification accuracy on frontal view projection is 98.31%, i.e., one lateral view image is classified as frontal.

The validation set of [169] includes 1759 images labeled as frontal view and 973 images labeled as lateral view. Testing this validation set with our projection view classifier, 82 images that are labeled as frontal are classified as lateral. We manually check these 82 images and find that seven images are indeed lateral view, one useless image (no anatomy presents), and one frontal view image rotated 180 degrees. And out of 973 images labeled as lateral, eight of them are classified as frontal. A manual checking of these 8 images reveals 6 of them are indeed frontal view. With the help of our classifier, 15 mislabeled or useless images are quickly found. Excluding these 15 images, the frontal view projection accuracy is 97.24%.

After manually check the images that are wrongly classified as lateral view, we find the quality of the majority of these frontal view images is suboptimal, as shown in Fig. 7.10, for example, poor body position, lung partially excluded, poor x-ray penetration, etc.

Fig 7.10 Examples of frontal Views Miss-classified as Lateral View

7.2.2 Discussion

Beside poor image quality, the adversarial examples are another factor that could trigger deep learning-based pathology detection and diagnosis applications to produce the wrong results. Recently, a team from MIT and Harvard medical school shows that adversarial examples can manipulate deep learning systems across clinical domains [171]. In medical imaging domain, they use patch and gradient descent methods to generate chest x-ray adversarial examples. Using chest x-ray and nine pixel perturbation methods, another work [172] also shows deep learning systems are vulnerable as well. Besides gradient based adversarial method as we used in chapter 6, Kanzow et al. [172] also use distortion injection methods from the IQA field, such as Gaussian blur and contrast reduction, to generate attacking examples.

Finlayson et al. [171] suggest algorithmic and IT infrastructural defense as possible areas for future research. Combining our finding in Chapter 6 and explorative work in section 7.2.1, we hypothesize that the fit-for-task approach might be another possible area for future research.

7.3 Summary

In this chapter, recent chest image analyzing applications are studied. The trend that deep learning techniques are replacing traditional machine learning is unignorable. Performance of triaging adult chest radiography reaches a clinically acceptable level. Lung nodule detection from computed tomography achieves within inter-observer variability among experienced human observers. Automatically generating text report for chest radiograph is demonstrated as feasible. Applications built with deep learning can effectively detect multiple pathologies in one study case. Looking forward, more accurately locating individual pathology in the image is a research problem, and image quality and adversarial are practical challenge.

Chapter 8

Conclusion

8.1 Contributions

In this thesis, new algorithms of image quality assessment are proposed from the aspects of feature extractors, feature descriptors and learning approaches. Further insight is discovered through systematic investigation of the impact of distortion types and magnitudes and adversarial examples to image classification performance. The key contributions of this work are outlined as in the following:

- The feasibility and advantage of modeling no-reference image assessment with deep learning methodologies. Using convolutional neural network plus a transformation of image signal from RGB into YIQ color space, we propose a no-reference model and experimentally show the performance advantage of the new approach over traditional learning framework.
- New no-reference and full-reference image quality assessment algorithms.

The major advantages of the proposed algorithms are: a) Introduce a unique feature extraction schema that combines features from 32-by-32 pixels local patches and entire image to design NRGD and NRSTATS; explore different feature descriptors and learning methods; b) Propose a global feature notion to extract image level quality characteristics, and experimentally show its contribution to a better model performance. c) Comparison studies show that the proposed NRGD and NRSTATS outperform

leading no-reference image quality assessment algorithms but underperforms the convolutional neural network based method proposed previously.

- A transfer learning-based no-reference image quality assessment algorithm.

The major contributions of this work are: a) Through transfer learning, we use deep convolutional model trained on ImageNet as a feature extractor and build an end-to-end trainable model; b) Replacing the conventional fully connected application layer with a recurrent neural network, we show that this leads to a more accurate prediction of image quality where distortions are not homogenous globally. c) A thorough evaluation of model performance on single quality databases and generalization ability across other quality databases are conducted. Experiments show that the knowledge learned from a high-level image-understanding task (object recognition) can be effectively transferred to tackle a low-level image-understanding task, namely quality assessment. The proposed algorithms DTNIQ-f and DTNIQ-r achieve state-of-the-art quality prediction power and generalization capability in terms of correlation with the human visual system.

- A framework toward resilient classification through image quality assessment.

The major contributions of this work are: a) from an image quality assessment perspective, we thoroughly study the impacts of various input distortions to object classification outcome by semantic distance, specifically Wu-Palmar similarity. b) We quantitatively show the characteristic shared between image quality and adversarial perturbations and propose quality as a useful feature component for adversarial detection; c) Using the state-of-the-art CNN model trained on ImageNet and experimental data covering 1000 object classes from the ImageNet, we propose a fit-for-task framework and show that the proposed method can effectively detect poor quality images and protect the CNN from the adversarial methods that are tested.

- The applications of deep learning in the chest radiology and image quality related problem in this domain.

Radiography and computed tomography are the most common radiological imaging modalities for pulmonary examinations. Firstly, recent deep learning based applications in chest radiology are reviewed. Secondly, a deep learning classifier is designed and trained to detect poor quality chest radiographs, regards projection view and body position. Using this classifier, we demonstrate its potential for managing image quality of large-scale radiology datasets.

8.2 Limitation and Future Work

This thesis work has proposed a preventive framework in the direction of safe artificial intelligence (AI). The framework detects and stops wrong object recognition from being made when the input image quality is poor or is maliciously manipulated. The research field of adversarial example generation and defense is becoming active due to recent year massive trend of rolling out AI into various commercial and industrial sectors. Along the direction of treating malicious manipulation as a kind of image quality degradation, we believe there are a couple of interesting threads for further research as a possible extension of the thesis work.

8.2.1 Improving Fit-For-Task Preventive Framework

The fit-for-task classifier can fairly well protect the CNN from making ridiculously wrong predictions when input images are of poor quality or maliciously altered with the experimented methods. Our methods are developed around the perceptual quality of a

whole image. Examining the unrecognizable adversarial images, we observe that these images do not have a meaningful composition from a photography standard or distinctly different composition compared to natural images.

The composition of an image impacts the visual distinctness of objects in an image. It includes space and size of the foreground object, the complexity of background texture, contrast and many more that are beyond the scope of existing image quality assessment problems. We think that conjugating visual saliency and image composition information with image quality assessment could be a potential direction to look into in the future.

Another limitation of the proposed preventive framework is that it would not be effective when the adversarial examples don't introduce discernable quality distortion. We believe that, like today's cybersecurity sector there are many solutions and that the future of AI safety will not rely only on one mechanism. A comprehensive review of the attacking methods would be beneficial to research work from here.

8.2.2 Radiological Imaging Quality Management

The fit-for-task principle can be extended to protect AIs from producing an error-prone result when the image quality is inadequate for diagnosis tasks. It is crucial that when training such AIs with a very large-scale database from hospitals, the image quality and label quality are sufficient. Otherwise, it could end up in a garbage-in and garbage-out situation. The image quality in this domain has unique aspects, including the projection, the necessary anatomy inclusion, the x-ray penetration, etc. In this thesis work, using chest radiographs available to the public, we propose a projection view classifier for chest radiographs and demonstrate one use case in an AI training scenario.

Time limitation is an unavoidable obstacle when doing part-time Ph.D. research, especially while in the meantime attempting to take technology development onto an entrepreneurship route. We believe this part of work has a lot of room to be expanded in the future.

References

- [1] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." *Neural computation* 18.7 (2006): 1527-1554.
- [2] Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *nature* 529.7587 (2016): 484.
- [3] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [4] Bengio, Yoshua. "Learning deep architectures for AI." *Foundations and trends® in Machine Learning* 2.1 (2009): 1-127.
- [5] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [6] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *Cognitive modeling* 5.3 (1988): 1.
- [7] Hochreiter, Sepp, et al. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies." (2001).
- [8] Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks." *Advances in neural information processing systems*. 2007.
- [9] Deng, Li, and Dong Yu. "Deep learning: methods and applications." *Foundations and Trends® in Signal Processing* 7.3–4 (2014): 197-387.
- [10] Fukushima, Kunihiro. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." *Biological cybernetics* 36.4 (1980): 193-202.
- [11] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.
- [12] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International Journal of Computer Vision* 115.3 (2015): 211-252.
- [13] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

- [14] Cadieu, Charles F., et al. "Deep neural networks rival the representation of primate IT cortex for core visual object recognition." *PLoS computational biology* 10.12 (2014): e1003963.
- [15] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *Computer vision–ECCV 2014*. Springer International Publishing, 2014. 818-833.
- [16] Yosinski, Jason, et al. "Understanding neural networks through deep visualization." *arXiv preprint arXiv:1506.06579* (2015).
- [17] Mahendran, Aravindh, and Andrea Vedaldi. "Understanding deep image representations by inverting them." *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.
- [18] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- [19] Arora, Sanjeev, et al. "Provable bounds for learning some deep representations." *International Conference on Machine Learning*. 2014.
- [20] Krizhevsky, Alex, and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Vol. 1. No. 4. Technical report, University of Toronto, 2009.
- [21] Jégou, Hervé, et al. "Aggregating local image descriptors into compact codes." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34.9 (2012): 1704-1716.
- [22] Briggs, William L., and Steve F. McCormick. *A multigrid tutorial*. Siam, 2000.
- [23] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
- [24] Oquab, Maxime, et al. "Learning and transferring mid-level image representations using convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
- [25] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." *International journal of computer vision* 88.2 (2010): 303-338.
- [26] Razavian, Ali, et al. "CNN features off-the-shelf: an astounding baseline for recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014.

- [27] Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." *arXiv preprint arXiv:1312.6229* (2013).
- [28] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
- [29] Karayev, Sergey, et al. "Recognizing image style." *arXiv preprint arXiv:1311.3715* (2013).
- [30] Donahue, Jeff, et al. "Decaf: A deep convolutional activation feature for generic visual recognition." *International conference on machine learning*. 2014.
- [31] Kruthiventi, Srinivas SS, Kumar Ayush, and R. Venkatesh Babu. "Deepfix: A fully convolutional neural network for predicting human eye fixations." *IEEE Transactions on Image Processing* 26.9 (2017): 4446-4456.
- [32] Bylinskii, Zoya, et al. "Mit saliency benchmark." (2015): 402-409.
- [33] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." *arXiv preprint arXiv:1508.06576* (2015).
- [34] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
- [35] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." *arXiv preprint arXiv:1211.5063* (2012).
- [36] Graves, Alex. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850* (2013).
- [37] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).
- [38] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
- [39] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- [40] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *European conference on computer vision*. Springer, Cham, 2014.

- [41] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- [42] Taylor, Graham W., Geoffrey E. Hinton, and Sam T. Roweis. "Modeling human motion using binary latent variables." *Advances in neural information processing systems*. 2006.
- [43] Salakhutdinov, Ruslan, and Geoffrey Hinton. "Semantic hashing." *RBM* 500.3 (2007): 500.
- [44] Alain, Guillaume, and Yoshua Bengio. "What regularized auto-encoders learn from the data-generating distribution." *The Journal of Machine Learning Research* 15.1 (2014): 3563-3593.
- [45] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507. (b)
- [46] Ranzato, Marc Aurelio, et al. "Unsupervised learning of invariant feature hierarchies with applications to object recognition." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007.
- [47] Vincent, Pascal, et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *The Journal of Machine Learning Research* 11 (2010): 3371-3408.
- [48] Vincent, Pascal, et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *The Journal of Machine Learning Research* 11 (2010): 3371-3408.
- [49] Ponomarenko, Nikolay, et al. "TID2008-a database for evaluation of full-reference visual quality assessment metrics." *Advances of Modern Radioelectronics* 10.4 (2009): 30-45.
- [50] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." *Image Processing, IEEE Transactions on* 13.4 (2004): 600-612.
- [51] Kang, Le, et al. "Convolutional neural networks for no-reference image quality assessment." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

- [52] Rensink, Ronald A. "The dynamic representation of scenes." *Visual cognition* 7.1-3 (2000): 17-42.
- [53] *IQA Research Webpage*, sse.tongji.edu.cn/linzhang/IQA/IQA.htm.
- [54] Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." *Signals, Systems and Computers, 2004*.
- [55] Sheikh, Hamid Rahim, and Alan C. Bovik. "Image information and visual quality." *Image Processing, IEEE Transactions on* 15.2 (2006): 430-444.
- [56] Chandler, Damon M., and Sheila S. Hemami. "VSNR: A wavelet-based visual signal-to-noise ratio for natural images." *Image Processing, IEEE Transactions on* 16.9 (2007): 2284-2298.
- [57] Zhang, Lin, D. Zhang, and Xuanqin Mou. "FSIM: a feature similarity index for image quality assessment." *Image Processing, IEEE Transactions on* 20.8 (2011): 2378-2386.
- [58] Zhang, L., Shen, Y., & Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *Image Processing, IEEE Transactions on*, 23(10), 4270-4281.
- [59] Narwaria, Manish, and Weisi Lin. "Objective image quality assessment based on support vector regression." *Neural Networks, IEEE Transactions on* 21.3 (2010): 515-519.
- [60] Liu, Tsung-Jung, Weisi Lin, and C-CJ Kuo. "A multi-metric fusion approach to visual quality assessment." *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*. IEEE, 2011.
- [61] Charrier, C., Lézoray, O., & Lebrun, G. (2012). Machine learning to design full-reference image quality assessment algorithm. *Signal Processing: Image Communication*, 27(3), 209-219.
- [62] Wang, Zhou, Alan C. Bovik, and B. L. Evan. "Blind measurement of blocking artifacts in images." *Image Processing, 2000. Proceedings. 2000 International Conference on*. Vol. 3. Ieee, 2000.
- [63] Li, Xin. "Blind image quality assessment." *Image Processing. 2002. Proceedings. 2002 International Conference on*. Vol. 1. IEEE, 2002.

- [64] Marziliano, Pina, et al. "A no-reference perceptual blur metric." *Image Processing. 2002. Proceedings. 2002 International Conference on*. Vol. 3. IEEE, 2002.
- [65] Caviedes, Jorge, and Sabri Gurbuz. "No-reference sharpness metric based on local edge kurtosis." *Image Processing. 2002. Proceedings. 2002 International Conference on*. Vol. 3. IEEE, 2002.
- [66] Zhu, Xiang, and Peyman Milanfar. "A no-reference sharpness metric sensitive to blur and noise." *Quality of Multimedia Experience, 2009. QoMEX 2009. International Workshop on*. IEEE, 2009.
- [67] Moorthy, Anush K., and Alan C. Bovik. "A two-step framework for constructing blind image quality indices." *Signal Processing Letters, IEEE* 17.5 (2010): 513-516.
- [68] Moorthy, Anush Krishna, and Alan Conrad Bovik. "Blind image quality assessment: From natural scene statistics to perceptual quality." *Image Processing, IEEE Transactions on* 20.12 (2011): 3350-3364.
- [69] Mittal, Anish, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain." *Image Processing, IEEE Transactions on* 21.12 (2012): 4695-4708.
- [70] Ruderman, Daniel L. "The statistics of natural images." *Network: computation in neural systems* 5.4 (1994): 517-548.
- [71] Ye, Peng, and David Doermann. "No-reference image quality assessment using visual codebooks." *Image Processing, IEEE Transactions on* 21.7 (2012): 3129-3138.
- [72] Xue, W., Mou, X., Zhang, L., Bovik, A. C., & Feng, X. (2014). Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *Image Processing, IEEE Transactions on*, 23(11), 4850-4862.
- [73] Geusebroek, Jan-Mark, and Arnold WM Smeulders. "A six-stimulus theory for stochastic texture." *International Journal of Computer Vision* 62.1-2 (2005): 7-16.
- [74] Wainwright, Martin J., and Odelia Schwartz. "10 Natural Image Statistics and Divisive." *Probabilistic models of the brain: Perception and neural function* (2002): 203.

- [75] Lyu, Siwei, and Eero P. Simoncelli. "Nonlinear image representation using divisive normalization." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.
- [76] Mittal, Anish, Rajiv Soundararajan, and Alan C. Bovik. "Making a “completely blind” image quality analyzer." *Signal Processing Letters, IEEE* 20.3 (2013): 209-212.
- [77] Zhang, Lin, Lei Zhang, and Alan C. Bovik. "A feature-enriched completely blind image quality evaluator." *Image Processing, IEEE Transactions on* 24.8 (2015): 2579-2591.
- [78] Xue, Wufeng, Lei Zhang, and Xuanqin Mou. "Learning without human scores for blind image quality assessment." *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013.
- [79] Chandler, Damon M. "Seven challenges in image quality assessment: past, present, and future research." *International Scholarly Research Notices* 2013 (2013).
- [80] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011): 27.
- [81] Sheikh, H. R., et al. "LIVE image quality assessment database release 2 (2005)." 2007 07-17]. <http://live.ece.utexas.edu/research/quality> (2005).
- [82] Ponomarenko, Nikolay, et al. "Color image database TID2013: Peculiarities and preliminary results." *Visual Information Processing (EUVIP), 2013 4th European Workshop on*. IEEE, 2013.
- [83] Larson, Eric C., and Damon M. Chandler. "Most apparent distortion: full-reference image quality assessment and the role of strategy." *Journal of Electronic Imaging* 19.1 (2010): 011006-011006.
- [84] Saad, M. A., Bovik, A. C., & Charrier, C. (2012). Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8), 3339-3352
- [85] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of Machine Learning Research* 9.2579-2605 (2008): 85

References

- [86] Zhang, Lin, et al. "A comprehensive evaluation of full reference image quality assessment algorithms." *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012.
- [87] Wang, Zhou, and Alan C. Bovik. "Mean squared error: love it or leave it? A new look at signal fidelity measures." *Signal Processing Magazine, IEEE* 26.1 (2009): 98-117.
- [88] Lin, Weisi, and C-C. Jay Kuo. "Perceptual visual quality metrics: A survey." *Journal of Visual Communication and Image Representation* 22.4 (2011): 297-312.
- [89] Zhang, Yi, and Damon M. Chandler. "No-reference image quality assessment based on log-derivative statistics of natural scenes." *Journal of Electronic Imaging* 22.4 (2013): 043025-043025.
- [90] Gu, Zhongyi, et al. "Learning quality-aware filters for no-reference image quality assessment." *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. IEEE, 2014.
- [91] Treisman, Anne M., and Garry Gelade. "A feature-integration theory of attention." *Cognitive psychology* 12.1 (1980): 97-136.
- [92] Kovesi, Peter. "Image features from phase congruency." *Videre: Journal of computer vision research* 1.3 (1999): 1-26.
- [93] Lasmar, N-E., Youssef Stitou, and Yannick Berthoumieu. "Multiscale skewed heavy tailed model for texture analysis." *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009.
- [94] Moorthy, Anush K., and Alan C. Bovik. "Statistics of natural image distortions." *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010.
- [95] Egmont-Petersen, Michael, Dick de Ridder, and Heinz Handels. "Image processing with neural networks—a review." *Pattern recognition* 35.10 (2002): 2279-2301.
- [96] "Video Quality Experts Group (VQEG)." *Video Quality Experts Group (VQEG)*. N.p., n.d. Web. 15 Mar. 2016.
- [97] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

- [98] Kümmerer, Matthias, Thomas SA Wallis, and Matthias Bethge. "DeepGaze II: Reading fixations from deep features trained on object recognition." *arXiv preprint arXiv:1610.01563* (2016).
- [99] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 1-40.
- [100] Ghadiyaram, Deepti, and Alan C. Bovik. "Massive online crowdsourced study of subjective and objective picture quality." *IEEE Transactions on Image Processing* 25.1 (2016): 372-387.
- [101] Ghadiyaram, D., & Bovik, A. C. (2014, December). Blind image quality assessment on real distorted images using deep belief nets. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on* (pp. 946-950). IEEE.
- [102] Bianco, Simone, et al. "On the use of deep learning for blind image quality assessment." *Signal, Image and Video Processing* 12.2 (2018): 355-362.
- [103] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014.
- [104] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958.
- [105] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (pp. 487-495).
- [106] Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.
- [107] Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal Adversarial Perturbations." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.
- [108] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010.

- [109]Hinton, Geoffrey E., et al. "Improving neural networks by preventing co-adaptation of feature detectors." *arXiv preprint arXiv:1207.0580* (2012).
- [110]Litjens, Geert, et al. "A survey on deep learning in medical image analysis." *Medical image analysis* 42 (2017): 60-88.
- [111]Xu, Huazhe, et al. "End-to-end learning of driving models from large-scale video datasets." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [112]Vasiljevic, Igor, Ayan Chakrabarti, and Gregory Shakhnarovich. "Examining the Impact of Blur on Recognition by Convolutional Networks." *arXiv preprint arXiv:1611.05760* (2016).
- [113]Dodge, Samuel, and Lina Karam. "A study and comparison of human and deep learning recognition performance under visual distortions." *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*. IEEE, 2017.
- [114]Karahan, Samil, et al. "How image degradations affect deep cnn-based face recognition?." *Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the*. IEEE, 2016.
- [115]Hosseini, Hossein, Baicen Xiao, and Radha Poovendran. "Google's Cloud Vision API Is Not Robust To Noise." *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE, 2017.
- [116]Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).
- [117]Lu, Jiajun, Theerasit Issaranon, and David A. Forsyth. "SafetyNet: Detecting and Rejecting Adversarial Examples Robustly." *ICCV*. 2017.
- [118]Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [119]Sitawarin, Chawin, et al. "DARTS: Deceiving Autonomous Cars with Toxic Signs." *arXiv preprint arXiv:1802.06430* (2018).

- [120]Feng, Yeli, and Cai Yiyu. "No-reference image quality assessment through transfer learning." *Signal and Image Processing (ICSIP), 2017 IEEE 2nd International Conference on*. IEEE, 2017.
- [121]Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." *arXiv preprint arXiv:1801.00553* (2018).
- [122]Li, Xin, and Fuxin Li. "Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics." *ICCV*. 2017.
- [123]Metzen, Jan Hendrik, et al. "On detecting adversarial perturbations." *arXiv preprint arXiv:1702.04267* (2017).
- [124]Lu, Jiajun, Theerasit Issaranon, and David A. Forsyth. "SafetyNet: Detecting and Rejecting Adversarial Examples Robustly." *ICCV*. 2017.
- [125]Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- [126]Meng, Dongyu, and Hao Chen. "Magnet: a two-pronged defense against adversarial examples." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017.
- [127]Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." *arXiv preprint arXiv:1704.01155* (2017).
- [128]Liang, Bin, et al. "Detecting adversarial examples in deep networks with adaptive noise reduction." *arXiv preprint arXiv:1705.08378* (2017).
- [129]Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection." *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994.
- [130]Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).
- [131]Gu, K., Zhai, G., Yang, X., & Zhang, W. (2014). Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1), 50-63.

- [132]Liu, Yutao, et al. "Blind quality assessment of camera images based on low-level and high-level statistical features." *IEEE Transactions on Multimedia* 21.1 (2019): 135-146.
- [133]Liu, Lixiong, Tianshu Wang, and Hua Huang. "Pre-attention and Spatial Dependency Driven No-Reference Image Quality Assessment." *IEEE Transactions on Multimedia* (2019).
- [134]Zhang, Lin, et al. "Training quality-aware filters for no-reference image quality assessment." *IEEE MultiMedia* 21.4 (2014): 67-75.
- [135]Hou, Weilong, et al. "Blind image quality assessment via deep learning." *IEEE transactions on neural networks and learning systems* 26.6 (2014): 1275-1286
- [136]Bosse, Sebastian, et al. "Deep neural networks for no-reference and full-reference image quality assessment." *IEEE Transactions on Image Processing* 27.1 (2017): 206-219.
- [137]Kim, Jongyoo, and Sanghoon Lee. "Deep learning of human visual sensitivity in image quality assessment framework." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [138]Ma, Kede, et al. "End-to-end blind image quality assessment using deep neural networks." *IEEE Transactions on Image Processing* 27.3 (2017): 1202-1213.
- [139]Wang, Zhou, and Alan C. Bovik. "Modern image quality assessment." *Synthesis Lectures on Image, Video, and Multimedia Processing* 2.1 (2006): 1-156.
- [140]Ruder, Sebastian. "An overview of gradient descent optimization algorithms." *arXiv preprint arXiv:1609.04747* (2016).
- [141]Marketforintelligence.com. (2016). *What's Next?* | [online] Available at: <http://www.marketforintelligence.com/talks/whats-next/>.
- [142]Pesapane, Filippo, Marina Codari, and Francesco Sardanelli. "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine." *European radiology experimental* 2.1 (2018): 35.
- [143]van Ginneken, Bram. "Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning." *Radiological physics and technology* 10.1 (2017): 23-32.

- [144]LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436.
- [145]Chartrand, Gabriel, et al. "Deep learning: a primer for radiologists." *Radiographics* 37.7 (2017): 2113-2131.
- [146]Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [147]Krause, Jonathan, et al. "A hierarchical approach for generating descriptive image paragraphs." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [148]Kaggle.com. (2017). *Data Science Bowl 2017 | Kaggle*. [online] Available at: <https://www.kaggle.com/c/data-science-bowl-2017#description>.
- [149]Shin, Hoo-Chang, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." *IEEE transactions on medical imaging* 35.5 (2016): 1285-1298.
- [150]Yao, Li, et al. "Learning to diagnose from scratch by exploiting dependencies among labels." *arXiv preprint arXiv:1710.10501* (2017).
- [151]Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225* (2017).
- [152]Li, Zhe, et al. "Thoracic disease identification and localization with limited supervision." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [153]Dunnmon, Jared A., et al. "Assessment of convolutional neural networks for automated classification of chest radiographs." *Radiology* (2018): 181422.
- [154]Yates, E. J., L. C. Yates, and H. Harvey. "Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification." *Clinical radiology* 73.9 (2018): 827-831.
- [155]Annarumma, Mauro, et al. "Automated triaging of adult chest radiographs with deep artificial neural networks." *Radiology* (2019): 180921.

- [156] Shin, Hoo-Chang, et al. "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [157] Jing, Baoyu, Pengtao Xie, and Eric Xing. "On the automatic generation of medical imaging reports." *arXiv preprint arXiv:1711.08195* (2017).
- [158] Wang, Xiaosong, et al. "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [159] Liao, Fangzhou, et al. "Evaluate the malignancy of pulmonary nodules using the 3d deep leaky noisy-or network." *arXiv preprint arXiv:1711.08324* (2017).
- [160] Ciompi, Francesco, et al. "Towards automatic pulmonary nodule management in lung cancer screening with deep learning." *Scientific reports* 7 (2017): 46479.
- [161] Anthimopoulos, Marios, et al. "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network." *IEEE transactions on medical imaging* 35.5 (2016): 1207-1216.
- [162] Gao, Mingchen, et al. "Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.1 (2018): 1-6.
- [163] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.
- [164] Juarez, A. Garcia-Uceda, H. A. W. M. Tiddens, and M. de Bruijne. "Automatic airway segmentation in chest CT using convolutional neural networks." *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, Cham, 2018. 238-250.
- [165] González, Germán, George R. Washko, and Raúl San José Estépar. "Deep learning for biomarker regression: application to osteoporosis and emphysema on chest CT scans." *Medical Imaging 2018: Image Processing*. Vol. 10574. International Society for Optics and Photonics, 2018.

- [166] Lessmann, Nikolas, et al. "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions." *IEEE transactions on medical imaging* 37.2 (2018): 615-625.
- [167] González, Germán, George R. Washko, and Raúl San José Estépar. "Automated agatston score computation in a large dataset of non ecg-gated chest computed tomography." *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016.
- [168] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." *arXiv preprint arXiv:1511.07122* (2015).
- [169] Demner-Fushman, Dina, et al. "Preparing a collection of radiology examinations for distribution and retrieval." *Journal of the American Medical Informatics Association* 23.2 (2015): 304-310.
- [170] Johnson, Alistair EW, et al. "MIMIC-CXR: A large publicly available database of labeled chest radiographs." *arXiv preprint arXiv:1901.07042* (2019).
- [171] Finlayson, Samuel G., et al. "Adversarial attacks against medical deep learning systems." *arXiv preprint arXiv:1804.05296* (2018).
- [172] Taghanaki, Saeid Asgari, Arkadeep Das, and Ghassan Hamarneh. "Vulnerability analysis of chest x-ray image classification against adversarial attacks." *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, Cham, 2018. 87-94.
- [173] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
- [174] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
- [175] Kanzow, Christian, Nobuo Yamashita, and Masao Fukushima. "Levenberg–Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints." *Journal of Computational and Applied Mathematics* 172.2 (2004): 375-397.
- [176] Rosenblat, F. "The perceptron, A perceiving and recognizing automation." *Cornell Aeronautical Laboratory* (1957).

- [177] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995): 273-297.
- [178] Martin, David, et al. "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." Vancouver:: Iccv, 2001.
- [179] Dobrowski, Julian. "Good vs Bad Quality Chest x-Rays and How to Identify Them." Chest X-ray Essentials. 2019.
- [180] Hua, Yi, Lixiong Liu, and Qingjie Zhao. "Medical image quality assessment via contrast masking." *2015 8th International Congress on Image and Signal Processing (CISP)*. IEEE, 2015.
- [181] Jayageetha, J., and C. Vasanthanayaki. "Medical Image Quality Assessment Using CSO Based Deep Neural Network." *Journal of medical systems* 42.11 (2018): 224.
- [182] Zhang, Zhicheng, et al. "Can signal-to-noise ratio perform as a baseline indicator for medical image quality assessment." *IEEE Access* 6 (2018): 11534-11543.
- [183] Chow, Li Sze, and Raveendran Paramesran. "Review of medical image quality assessment." *Biomedical signal processing and control* 27 (2016): 145-154.
- [184] Fintelmann, Florian, et al. "Repeat rates in digital chest radiography and strategies for improvement." *Journal of thoracic imaging* 27.3 (2012): 148-151.
- [185] Wikipedia Contributors. "Spearman's Rank Correlation Coefficient." *Wikipedia*, Wikimedia Foundation, 7 Sept. 2019, en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient. Accessed 9 Sept. 2019.

Appendices

A. Connection Between Linear Autoencoder and PCA

Principal component analysis (PCA) is often used as dimensional reduction technique. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix, each of the n rows is a data point and each of the p columns has zero mean. A linear transformation $\mathbf{Y} = \mathbf{XV}$ maps data matrix into a new space spanned by $\mathbf{V} \in \mathbb{R}^{p \times p}$. Applying SVD the data matrix can be decomposed to $\mathbf{X} = \mathbf{USV}^T$, where $\mathbf{S} \in \mathbb{R}^{n \times p}$ is a diagonal matrix, $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. Substitute $\mathbf{X}$ in its covariance matrix with its SVD we obtain:

$$\mathbf{C} = \mathbf{VS}^T \mathbf{U}^T \mathbf{USV}^T / (n - 1) = \mathbf{V} \frac{\mathbf{S}^2}{n-1} \mathbf{V}^T \quad (\text{a-1})$$

$$\mathbf{S}^2 = \begin{cases} \sigma_{i,j}^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad \forall i, j \in \{1, 2, \dots, p\} \quad (\text{a-2})$$

$\mathbf{V}$ is eigenvector of $\mathbf{C}$ and the diagonal entries of $\frac{\mathbf{S}^2}{n-1}$ are eigenvalues. For dimensional reduction only first k largest components of the eigenvalues and corresponding columns (principal axis) of $\mathbf{V}$ need to be kept.

Now let the same data matrix $\mathbf{X}$ be the input to a linear autoencoder with 1 hidden layer of k units, $\mathbf{Y}$ be output of the encoder and $\mathbf{X}'$ be output of the decoder. We have:

$$\mathbf{Y} = \mathbf{XW}_1^T + \mathbf{ub}_1^T, \quad \mathbf{X}' = \mathbf{W}_2\mathbf{Y} + \mathbf{b}_2\mathbf{u}^T \quad (\text{a-3})$$

$$J = \|\mathbf{X} - \mathbf{X}'\|^2 = \|\mathbf{X} - \mathbf{W}_2\mathbf{Y} - \mathbf{b}_2\mathbf{u}^T\|^2 \quad (\text{a-4})$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights of the encoder and decoder.

For a weight tied autoencoder $\mathbf{W}_2 = \mathbf{W}_1^T$. $\mathbf{b}_1$ and $\mathbf{b}_2$ are bias and $\mathbf{u} \in \mathbb{R}^n$ is a column vector with each entry equals 1. The training goal is to minimize J with respect to the weights and bias.

Minimizing J with respect to $\mathbf{b}_2$ we have $\mathbf{b}_2 = \frac{1}{n}(\mathbf{X} - \mathbf{W}_2\mathbf{Y})\mathbf{u}$, as $\mathbf{u}^T\mathbf{u} = n$. Substitute it into eqn. b-4 we have $J = \left\| (\mathbf{X} - \mathbf{W}_2\mathbf{Y})(\mathbf{I} - \frac{1}{n}\mathbf{u}\mathbf{u}^T) \right\|^2$. Minimizing J with respect to $\mathbf{W}_2$ we have $\mathbf{X} = \mathbf{W}_2\mathbf{Y}$. Substitute $\mathbf{X}$ with best rank k approximation of its SVD decomposition we have $\mathbf{W}_2\mathbf{Y} = \mathbf{U}\mathbf{S}_k\mathbf{V}^T$, which can be rewritten into:

$$\mathbf{W}_2 = \mathbf{U}\mathbf{I}^{-1}, \quad \mathbf{Y} = \mathbf{I}\mathbf{S}_k\mathbf{V}^T$$

where $\mathbf{I}$ is an identify matrix. $\mathbf{V}$ is eigenvector spans the space where the encoder's output resides and the diagonal entries of $\mathbf{S}_k$ are eigenvalues, with a constraint that $\mathbf{W}_2$ be the left singular vector of $\mathbf{S}$. Comparing with eqn. b-1 and b-2 we see the connection between the code (representation) a linear autoencoder learns and the PCA transformation.

B. Distortions Types and Semantic Similarity

Fig B.1 Semantic Similarity Between Objects Recognized by ResNet-152

Examples of objects recognized from reference and distorted images by the ResNet-152 model trained on the ImageNet and Wu-Palmar similarity between objects, which is range from 0 to 1. Each object is represented by a synset in the WordNet. The Wu-Palmar similarity between warplane and airliner is 0.88, and 0.50 between catamaran and safety pin. The images are from the LIVE and TID2013 datasets.

Table B.1 Full Term of Distortion Types in LIVE and CSIQ

Abbreviation	Full Term
JPEG	Jpeg compression
JP2K	JPEG2000 compression
GB	Gaussian blur
AWN	Additive Gaussian white noise
FF	Fast fading Rayleigh
PGN	Additive Gaussian pink noise
GCD	Global contrast decrements

Table B.2 Full Term of Distortion Types in TID2013

No.	Abbreviation	Full Term
1	AWN	Additive Gaussian white noise
2	ANMC	Additive noise in color components (more intensive than additive noise in the luminance component)
3	SCN	Spatially correlated noise
4	MH	Masked noise
5	HFN	High frequency noise
6	IMN	Impulse noise
7	QN	Quantization noise
8	GB	Gaussian blur
9	DEN	Image denoising
10	JPEG	JPEG compression
11	JP2K	JPEG2000 compression
12	JGTE	JPEG transmission errors
13	J2TE	JPEG2000 transmission errors
14	NEPN	Non eccentricity pattern noise
15	Block	Local block-wise distortions of different intensity
16	MS	Mean shift (intensity shift)
17	CTC	Contrast change
18	CCS	Change of color saturation
19	MGN	Multiplicative Gaussian noise
20	CN	Comfort noise
21	LCNI	Lossy compression of noisy images
22	ICQD	Image color quantization with dither
23	CHA	Chromatic aberrations
24	SSR	Sparse sampling and reconstruction

C. A Gaze Intervention Framework For Children With Autism

Individuals with autism exhibit deficits in social communication and interaction. Researchers [1] have been using interactive virtual reality technologies helping children with autism to improve their communication and learning. For example, children with autism interact with virtual dolphins with hand gesture in Parsons et al. [12], learn social skills in Bellani et al. [3] and safety skills in Strickland et al. [4]. In this chapter, we propose a low cost intervention system that aims to enhance the visual responsiveness of children with autism. In the proposed prototype, children with autism use gaze to interact with tasks that are designed with game mechanics. The eye fixation data of players are recorded, which can be utilized to design more intervention mechanism such as through manipulating regional image quality.

C.1 Background

Since the early adoption of virtual reality (VR) technology in the research of autism spectrum disorder in 1996 by Strickland et al. [4], many research activities have been channeled to using VR technology to understand the unique behavior of children with autism; and to develop effective intervention programs to treat autism. In 2002 Parsons et al. [1] reviewed the potential usefulness of VR in providing a safe and non-threatening environment and a realistic setting for role-play of desired behaviors, so that can be repeatedly practiced in a consistent way. A decade later Parsons et al. [2] and et al. Bellani [3] reviewed the state-of-the-art in this field. They both concluded that children with autism can learn rule-based social skills from VR in varying degree and a few can transfer the acquired skills to real word. Other latest researches [5-7] suggest VR is

promising in improving understanding of empathy and social interaction in children with autism. Our work focuses on the non-verbal social interaction deficits in children with autism that using eye contact as a mean of communication.

This chapter introduces a low cost intervention system aiming to enhance the visual responsiveness of children with autism. Section C.2 reviews the deficits in social skills of children with autism spectrum and the recent changes in autism diagnosis. Section C.3 discusses the evolution of autism treatment. Section C.4 presents the intervention system we design and develop for autistic children in Singapore special education schools. Section C.5 describes data analysis methodology and evaluation of the effectiveness of the application. Discussion and summary are provided in sections C.6 and C.7.

C.2 Evolution in Autism Research

During the 1940s Leo Kanner and Hans Asperger independently reported children who presented “a powerful desire for loneliness” and “an obsessive insistence on persistent sameness”. Since then psychiatrists and psychologists looked into the life experience impacts for explanation and intervention for this behavior disorder in the domains of social deficits, communication and language deficits. From the 1970s to 90s many behavioral treatments based on the operant conditioning principles were developed targeting behaviors that include social skills, language, daily living skills, academic skills and aberrant behaviors Matson et al. [8]. Behavioral treatments systematically apply reinforcement, punishment and extinction techniques or a combination to develop effective treatment strategies. The trend continued to increase; today it is generally referred to as applied behavior analysis (ABA). ABA and many variations are widely

practiced in the field of autism and developmental disabilities in special education in the US, UK, and many other countries.

In the 1985 article [10] “Does the autistic child have a ‘theory of mind’?” Baron et al. suggest that the lack of ‘theory of mind’ (ToM) is the underlying cognitive condition attributed to the impaired social communication in autistic children. The ToM is the ability to attribute beliefs, intentions and other mental states to self and others to understand others’ belief and intention. Later studies by Parsons et al. [2] indicate that teaching ToM to people with children autism is helpful for them to pass ToM tasks that have been taught but not new tasks. With the advance in non-intrusive corneal reflection eye tracking and the functional magnetic resonance imaging (fMRI) technologies, since early 2000s studies [11] have been carried out to search for the characterization of eye movement of individuals with autism during social processing and the connection between the neurodevelopmental condition and behavioral dysfunctions. Evidence of atypical gazing pattern and hypoactivation in different nodes of the brain has been found although details are varying in findings.

Over time autism is diagnosed more frequently. In 2010 the CDC monitoring network reports ‘The global prevalence of autism has increased twentyfold to thirtyfold since the earliest epidemiologic studies were conducted in the late 1960s and early 1970s’. Although the underlying reasons for the apparent high increase rate are unclear, several factors likely contributed such as improved awareness and services to the general public, and changes in diagnostic criteria. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) groups all sub-conditions such as Asperger syndrome into two categories: impaired social communication and or interaction, and restricted and or repetitive behaviors.

C.3 Atypical Gaze Pattern

Gaze plays a vital role in social interaction. Typically developed infants connect with caregivers in moments of gazing at caregivers' face, following caregivers' gaze direction. Around one year old they are capable of establishing joint attention. In everyday social processing typically developed adults frequently utilize eye contact to signal messages such as friendship, agreement or anger. One of the core symptoms of autism is the abnormality in eye contact with people and objects during social interaction.

In the autism research literature, many studies investigated such abnormality. Using human faces as stimuli Pelphrey et al. [12] report the scan paths of adult males with autism seemed erratic and disorganized. They had fewer fixations on the core features of the faces, and less accurate in identifying the six basic emotions, fear in particular. Klin et al. [13] use naturalistic social situations to study the difference of fixation on social and non-social regions of a dynamic scene. They found adolescents and young adults with autism fixated significantly less on the eye but more on the mouth, body and object regions when watching videotape clips. Within a group of autism participants with higher social competence had more fixations on the mouth region. Von Hofsten et al. [14] report children with autism had fewer fixations and shorter fixation duration on the faces of the people involved in a conversation but no evidence of impairment on tracking moving objects. Chawarska et al. [15] report that the distribution of attention from toddlers with autism on inner and outer facial features are different than typically developed peers, hence are less effective in facial recognition. Attention to eyes, nose, and mouth features decreased in older toddlers with autism that suggested a possible sign of detrimental over time.

C.4 A Gaze Intervention Framework

Autism is a neurodevelopmental disorder. People with autism have a different hence less effective mechanism to gather information and communicate in the world favors of typically developed individuals. They experience a varying degree of difficulties in maintaining a friendship, relationship, and employment that can lead to social exclusion. However, the development in the neuroplasticity field by Kolb et al. [16] give hope to ABA intervention. The connectivity of the brain changes throughout the lifetime and is sensitive to the patterns produced through experience. A change in behavior signifies some change in the brain. Through ABA experience designed on repetition and reinforcement, the emerging brain of children with autism is rewired and retrained to engage in behaviors that are better accepted socially.

ABA intervention requires expertise and long-term commitment. The accessibility and affordability could be the factors preventing it from reaching children with autism in need. In the next section we propose a system that allows children with autism to receive regular intervention without the need for ABA expertise always. The exercise can be carried out at school, home or community facilities such as the library. The system aims to help them to enhance visual responsiveness in the new media environment. First, the enabling technology eye tracking is introduced. Then we describe the system framework and application development.

C.4.1 Eye Tracking Technology

The eye tracking technology was first used in autism research by Boraston et al. more than ten years ago. Duckowski et al. [19] show tremendous improvements have been

introduced since then. Most eye tracking devices used in the field of autism study fall into one of the four categories. See Fig. C.1 table-mounted with chin rest, head-mounted, glass-wear, and remote systems. These systems are video-based, rely on the corneal-pupil reflection of infra-red or ambient lights directed to the eye to estimate the point of gaze.

Fig C.1 Examples of Eye Tacking Devices.

(a) Cambridge Research Systems High-Speed VET. (b) SR Research EYELINK II. (c) SMI eye tracking glass. (d) Tobii EyeX controller

The corneal reflections are known as Purkinje images. Four Purkinje images are used commonly. The first and second Purkinje images are the reflections from the outer and inner surface of the cornea respectively. The third and fourth Purkinje images are the reflections from the outer and inner surfaces of the lens respectively. See Fig. C.2. Gaze location is computed by extracting the pupil center and one or more of the corneal reflections with an image-processing algorithm. Calibration is required to map the estimated gaze location onto the viewing scene coordinates, for example, x and y coordinates of the monitor.

Fig C.2 Illustration of Corneal Reflections.

Left: Purkinje images. Right: Illustration of pupil and reflection extracted by image-processing algorithm.

We use a Tobii EyeX Controller, a remote system can be mounted on both desktop and laptop setups. It doesn't need regular re-calibration and the user can sit, stand and move freely. The operating distance is between 45cm and 80cm. The freedom of head movement is 48cm in width and 39cm in height.

C.4.2 System Technology

The gaze intervention system is built on top of the EyeX engine [20] and the Unity game engine [21]. The EyeX engine filters the raw gaze data from the EyeX controller and reports to gaze-enabled applications or interprets into meaningful interactions that are exposed to applications as integrators. See Fig. C.3. The Unity game engine is a cross-platform 3D rendering engine that comes with intuitive tools such as scene editor, scripting and animation workflow.

As shown in Fig. C.4 the gaze intervention system creates training tasks with the Unity game engine. The training tasks are simple games played via eye interaction. In each training task, areas of interest (AOI) are defined as inter-actors. A fixation is a gaze on a single location for duration of time that is long enough to trigger an AOI event. The AOI

event will give a reward to a player on the computer screen. In the meantime, the player's gaze data during an entire session are captured for analysis.

Fig C.3 Framework of Gaze-Enabled Applications

Fig C.4 A Schematic Diagram of the Proposed Gaze Intervention System

C.4.3 Application Development

The training tasks we design and develop restrict children to advance their gameplay with eye interaction only. In training, task regions are defined to give reward, cue or punishment. For example in Fig. C.5, during the play, if the child fixes his or her eye on the dragons' eyes, he or she will be rewarded a coin. If the child sets sight on a distracter, a flyer in this design, a visual cue that one of the dragons plays a cheering animation is given.

Studies by Guillon et al. [17] reveal that the gazing difference between typically developed individuals and with autism does not generalize across static and dynamic social and non-social stimuli. Therefore two types of evaluation tasks are developed adopting widely used methods in the autism eye-tracking research field. In a paired comparison format, a sequence of image pairs that show a human face or person-absent are flashed to the viewer. Scan path, start, and end gaze points are captured for analysis, as shown in Fig. .6. Video clip of cartoon characters having a conversation is used as dynamic stimuli, attention and eye movement activities are captured for evaluation.

Fig C.5 Example of Training Task

During play mode the mask of rectangle regions and gaze and fixation tracker (white and grey dots) are turned off by default.

C.4.4 Data Analysis

Three types of playing data stream are captured; they are gaze points, fixation points and triggered events. The triggered events data forms a profile of each child's eye

movements. From the gaze points, we can infer whether the child with autism pays attention to the screen, how frequently he or she moves eye among different regions in the screen. See Fig. C.7 and C.8. The fixation points reveal what features attract the player's attention if they are not the defined reinforcement regions.

This type of information can be utilized in real-time to customize the gameplay that suits an individual child, for example, the difficulty level of controlling with the gaze.

Fig C.6 Example of Evaluation Task Using Static Stimuli.

Line depicts the scan path of gazing and black squares are the start and end points.

C.4.5 Discussion

In the prototype, moving objects (distractors) are used to facilitating the eye fixation to the desired region. With this low cost and easy to expand framework, more visual guiding methods can be explored and experimented. For example, how would children with autism react to noise in images? Visual attention study by Röhrbein et al. [22] find out that normal subjects give less number of fixations on distorted images, but longer fixation durations and the gaze distribution is mainly manipulated by the intensity of distortion rather than the types of distortion. Would children with autism exhibit similar gaze

response to the noise manipulation in images? Could image quality manipulation be turned into a new gaze intervention parameter? A potential direction could be investigated in the future through extending the prototype.

Fig C.7 Saccades of Eye During An Evaluation Task

Fig C.8 Distribution of Attention During Training in Seconds

C.5 Summary

How effective the behaviors and responses built through practice in virtual reality context can be transferred into the real world is debatable in the current research community Parsons et al. [1]. In our future work evaluation task of the real world will be designed to explore this matter. Evaluation tasks can be watching a clip of cartoon video available on the Internet, having a video chat with parents or peer children over Skype. Using object detection and locating technique, features such as human and character eyes and mouths can be extracted, compared with gaze and fixation points captured, and experiment with intervention through image noise manipulation.

C.6 References

- [1] Parsons, Sarah, and Sue Cobb. "State-of-the-art of virtual reality technologies for children on the autism spectrum." *European Journal of Special Needs Education* 26.3 (2011): 355-366.
- [2] Parsons, Sarah, and Peter Mitchell. "The potential of virtual reality in social skills training for people with autistic spectrum disorders." *Journal of intellectual disability research* 46.5 (2002): 430-443.
- [3] Bellani, Marcella, et al. "Virtual reality in autism: state of the art." *Epidemiology and psychiatric sciences* 20.3 (2011): 235-238.
- [4] Strickland, Dorothy. "A virtual reality application with autistic children." *Presence: Teleoperators & Virtual Environments* 5.3 (1996): 319-329.
- [5] Cai, Yiyu, et al. "Design and development of a virtual dolphinarium for children with autism." *IEEE transactions on neural systems and rehabilitation engineering* 21.2 (2013): 208-217.
- [6] Ehrlich, Justin A., and James R. Miller. "A virtual environment for teaching social skills: AViSSS." *IEEE computer graphics and applications* 29.4 (2009): 10-16.

- [7] Cheng, Yufang, et al. "Enhancing empathy instruction using a collaborative virtual learning environment for children with autistic spectrum conditions." *Computers & Education* 55.4 (2010): 1449-1458.
- [8] Matson, Johnny L., et al. "Behavioral treatment of autistic persons: A review of research from 1980 to the present." *Research in developmental disabilities* 17.6 (1996): 433-465.
- [9] Matson, Johnny L., Megan A. Hattier, and Brian Belva. "Treating adaptive living skills of persons with autism using applied behavior analysis: A review." *Research in Autism Spectrum Disorders* 6.1 (2012): 271-276.
- [10] Baron-Cohen, Simon, Alan M. Leslie, and Uta Frith. "Does the autistic child have a "theory of mind"?" *Cognition* 21.1 (1985): 37-46.
- [11] Dichter, Gabriel S. "Functional magnetic resonance imaging of autism spectrum disorders." *Dialogues in clinical neuroscience* 14.3 (2012): 319.
- [12] Pelphrey, Kevin A., et al. "Visual scanning of faces in autism." *Journal of autism and developmental disorders* 32.4 (2002): 249-261.
- [13] Klin, Ami, et al. "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism." *Archives of general psychiatry* 59.9 (2002): 809-816.
- [14] von Hofsten, Claes, et al. "How children with autism look at events." *Research in Autism Spectrum Disorders* 3.2 (2009): 556-569.
- [15] Chawarska, Katarzyna, and Frederick Shic. "Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder." *Journal of autism and developmental disorders* 39.12 (2009): 1663.
- [16] Kolb, Bryan, and Ian Q. Whishaw. "Brain plasticity and behavior." *Annual review of psychology* 49.1 (1998): 43-64.
- [17] Guillon, Quentin, et al. "Visual social attention in autism spectrum disorder: Insights from eye tracking studies." *Neuroscience & Biobehavioral Reviews* 42 (2014): 279-297.
- [18] Boraston, Zillah, and Sarah Jayne Blakemore. "The application of eye-tracking technology in the study of autism." *The Journal of physiology* 581.3 (2007): 893-898.

- [19] Duchowski, Andrew T. "Eye tracking methodology." *Theory and practice* 328 (2007): 614.
 - [20] Tobii EyeX – experience eye tracking in apps or games. tobii.com/en/eye-experience/eyex. Accessed 17 Oct 2014
 - [21] Unity - Game Engine. <http://unity3d.com/>. Accessed 17 Oct 2014
 - [22] Röhrbein, Florian, et al. "How does image noise affect actual and predicted human gaze allocation in assessing image quality?." *Vision research* 112 (2015): 11-25.
 - [23] Laganière, R. (2011). *OpenCV 2 Computer Vision Application Programming Cookbook: Over 50 recipes to master this library of programming functions for real-time computer vision*. Packt Publishing Ltd.
- Röhrbein, Florian, et al. "How does image noise affect actual and predicted human gaze allocation in assessing image quality?." *Vision research* 112 (2015): 11-25.