SEMI-SUPERVISED LEARNING FOR VISUAL RELATION ANNOTATION

MITRA TAJROBEHKAR

School of Computer Science and Engineering

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

2022

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.

11/01/2022

.....

Date

.....

MITRA TAJROBEHKAR

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.

11/01/2022

.....

Date

.....

Adjunct Prof. Joo-Hwee Lim

Authorship Attribution Statement

This thesis contains material from 1 paper published in the following journal

Chapter 3 is published as Mitra Tajrobehkar, Kaihua Tang, Hanwang Zhang, Joo-Hwee Lim. "Align R-CNN: A Pairwise Head Network for Visual Relationship Detection," *IEEE Transactions on Multimedia, vol. 24, pp. 1266-1276, 2022.*

11/01/2022

.....

Date

.....

MITRA TAJROBEHKAR

Acknowledgements

Through my four-year PhD, I have come to realize that each person I talked with, worked with, met with has changed the trajectory of my life. First of all, I would like to express my sincere gratitude to my supervisors Professor Joo-Hwee Lim and Professor Hanwang Zhang that I was very fortunate to pursue my PhD under their supervisions. Beyond the detailed suggestions they both have given to my research, they have also given me much valuable advice, for my professional career and my outlook on life.

I would also like to thank Professor Shijian Lu, for his advice during my first year of my PhD. I would like to thank to my committee members Professor Cuntai Guan, and Dr. Liyuan Li for theirs comments on my research progress on TAC meetings.

I am also grateful to all my colleagues, especially Dr. Kaihua Tang, Dr. Yulei Niu, and Dr. Long Chen from MReaL group and Dr. Hongyuan Zhu and Dr. Hui Li from I2r.

This doctoral dissertation would not have been possible without funding from the Agency for Science, Technology and Research (A*STAR, Singapore). I would like to acknowledge School of Computer Science and Engineering, NTU for providing a wide range of valuable resources.

Finally and most importantly, I would like to thank my parents and my brother, who always provide me with their strong support from Iran. I really appreciate your love and effort to bring me up to be a better individual. This thesis is dedicated to you.

Mitra Tajrobehkar, January 2022

Abstract

Due to the powerful ability to learn low-level and high-level general visual features, deep neural networks (DNNs) have been used as the basic structure in many CV applications such as object detection, semantic segmentation, relation detection and annotation, *etc.* While most of the research focuses are on maximizing *overall* performance during training a machine learning model, not much attention is given to evaluate the robustness *i.e.* against visual content manipulation, until very recent years. Lack of model robustness especially with respect to consistency and discrimination can be due to various reasons, *e.g.* data distributions, inadequacy in learning process, model sensitivity to different regions of feature space, *etc.* *Discrimination* refers to the model capability of predictions to distinguish between individual class samples, while *consistency* refers to the model capability of predictions to remain stable despite input variations.

To address these challenges, the central focus of this thesis is on representation learning – for two most challenging Computer Vision (CV) tasks: Scene Graph Generation (SGG), and Visual Question Answering (VQA). For the first research direction, we propose a novel head network to tackle the problem of non-discrimination by learning semantic pairwise feature representation. The second research direction addresses the model instability by generating better representation. Through a consensus model, common feature representations that are reasoned from various samples are learned to increase the robustness of consensus. In summary, the major contributions of this thesis are as follows:

- **We propose a meta-architecture — learning-to-align —**, called ALIGN R-CNN, for dynamic object feature concatenation to deal with visual reasoning task. Taking scene graph generation as an example, humans learn to describe visual relationships between objects, semantically (*e.g.*, `riding` `behind`, `sitting on`). We propose a semantic transformation that parses an input image into

<subject-relation-object> triplet and then extracts visually grounded semantic features by re-aligning features of subject from its relative object and relation. We argues that the previous works are highly limited by naive concatenation and as a result, they fail to discriminate between `riding` and `feeding` for object pair of `person` and `horse`. Moreover, naive concatenated pairwise features may collapse less frequent but meaningful predicate *e.g.*, `sitting on` into more frequent but meaningless one *e.g.*, `on`. Compared with existing model relation representations that utilize scene graphs to connect the objects, the proposed ALIGN R-CNN has two key advantages: 1) maintains a good representations during training while removing the irrelevant features from the objects; 2) dynamic learning, that enables model deals with different pairs. These advantages prevent the proposed ALIGN R-CNN from over-fitting with the biased dataset. Note that the proposed framework can be utilized in a community which seeks zero-shot predictions.

- **We propose a framework to enhance model consistency** by generating desirable feature consistency. This line of research addresses lack of VQA models that measure the robustness of consensus against linguistic variations in questions. For instance, while reference question "`How many cars in picture?`" and its syntactic one "`How many automobiles are there?`" are semantically identical in meaning, model may predict an incorrect answer for syntactic one. Besides, the model should be powerful enough to predict the right answer for "`How many red cars seen in picture?`" Inspired from unsupervised feature representation learning, we use contrastive learning, which sufficiently learns better representation from both vision and language inputs. However, we argue that training the model with naive contrastive learning framework that using random intra-class and random non-target sequences as positive and negative examples is sub-optimal. Thus, it may not boost the model performance on robust VQA benchmark. The proposed method dedicates a principal head network to generate positive and negative samples for contrastive learning by adding adversarial perturbations. Specifically, it generates *hard* positive samples by adding large perturbations to both input images and questions to maximize the conditional likelihood. The proposed framework has two key advantages: 1) the generative model from embedding representation offers rich information to increase model stability; 2) by exploring the effects of single modalities and

multi-modal attacks, the model mitigates correlation between the bias and the learned features.

Contents

List of Figures

List of Tables

Abbreviations

AdaM	**A**daptive **M**oment Estimation
ADV	**A**dversial
AE	**A**uto-**E**encoder
AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etwork
AT	**A**dversarial **T**raining
BP	**B**ack**P**ropagation
BT	**B**ack **T**ranslation
CE	**C**ross-**E**ntropy
CL	**C**ontrastive **L**earning
CNN	**C**onvolutional **N**eural **N**etwork
CS	**C**onsensus **S**core
DL	**D**eep **L**earning
DNNs	**D**eep **N**eural **N**etworks
FC	**F**ully **C**onnected
FGSM	**F**ast **G**radient **S**ign **M**ethod
GT	**G**round-**T**ruth
IFGSM	**I**terative **F**ast **G**radient **S**ign **M**ethod
KL	**K**ullback **L**eibler
LSTM	**L**ong **S**hort-**T**erm **M**emory
PGD	**P**rojected **G**radient **D**escent
PhrDet	**P**hrase **D**etection
PredCls	**P**redicate **C**lassification

PredDet	**P**redicate **D**etection
ReLU	**R**ectified **L**inear **U**nit
RNNs	**R**obust **N**eural **N**etworks
ROI	**R**erion **O**f **I**nterest
mR	**m**ean **R**ecall
MMT	**M**ulti-**M**odal **T**ransformer
MV	**M**achine **V**ision
NCE	**N**oise **C**ontrastive **E**stimation
NLP	**N**atural **L**anguage **P**rocessing
NMT	**N**eural **M**achine **T**ranslation
SCL	**S**upervised **C**ontrastive **L**oss
SGCls	**S**cene **G**raph **C**lassification
SGD	**S**tochastic **G**radient **D**escent
SGDet	**S**cene **G**raph **D**etection
SGG	**S**cene **G**raph **G**eneration
SL	**S**upervised **L**earning
VG	**V**isual **G**enome
VidVRD	**V**ideo **V**isual **R**elation **D**etection
VQG	**V**isual **Q**uestion **G**eneration
VRD	**V**isual **R**elation **D**etection
ZSL	**Z**ero **S**hot **L**earning

Chapter 1

Introduction

1.1 Motivation

Several decades of Artificial Intelligence (AI) and Machine Learning (ML) research have led to significant advances in computer vision (CV), natural language processing (NLP), automated reasoning, and *etc.*
In the last few years, AI has shown significant progress by emulating human-like reasoning in some tasks [12, 13] by learning from data labels (supervised Learning). In particular, by inventing modern neural networks, computer vision based applications, *e.g.*, CNN-based image classification models [14–16], have outperformed human performances. Some other tasks like object detection [17–19] and image segmentation [20], have also shown significant success.

Although rapid improvements have been achieved in visual tasks at all levels, there is still a long way to go. Former studies point out a semantic gap between learning image feature representation by machines and human visual understanding mechanisms. Hence, a core goal in vision science is to understand what features the human visual system uses to process complex visual scenes. According to cognitive neuroscience research [21–23], the brain cognitive system is able to discriminate between entities by finding dissimilarities between their *e.g.* appearances via a hierarchical process that makes high-level representations from the observations. Deep learning, also known as representation-based learning [24], is a particular approach to machine learning that is gaining popularity due to its ability to capture hierarchy of representations. For this reason, former studies like [12, 13, 25–27]

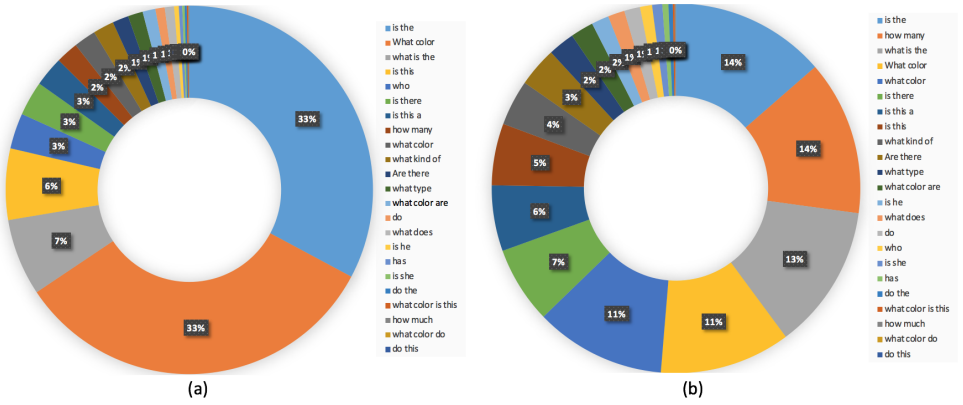

FIGURE 1.1: **Question Type Distribution** in VQA v2.0 dataset in (a) train (b) test set

have focused on reducing the semantic gap based on reasoning and learning rich and informative feature representation.

The extraction and synthesis of such representations (rich information) from a multi-dimensional data space requires the use of a mechanism that is to facilitate decision making. Especially, in more challenging vision-based tasks that need semantic feature understanding from complex or multi-modal inputs, effective representation of information are still bottlenecks. Most modern vision applications involve more than one modality (*e.g.* visual and textual modalities), such as embodied question answering, vision and language navigation, *etc.* Therefore, it is crucial to learn more complicated and cross-modal information from different modalities and data distributions, where deep multi-modal learning will be needed. A series of previous works [12, 13, 25–27], have built a structured representation that captures comprehensive semantic knowledge is a crucial step towards a deeper understanding of input scenes. Such representation cannot only offer contextual cues for fundamental recognition challenges, but also provide a promising alternative to high-level intelligence vision tasks. From another aspect, the key challenge is to extract visual attributes from one or more data modalities by learning how to fuse the extracted features into a common representation space, which is referred to as representation learning.

Moreover, the significant progress under deep neural network (DNN) models shows that the high performance in algorithms is limited by the lack of similar distribution between the training dataset and the testing set [28]. For instant, Figure. 1.1 displays non-uniform and long-tailed data distributions for a challenging task such

as visual question answering, over question modality. Some question types like "`who is`" are more common in training set than in test set (imbalance dataset distribution). Moreover, some questions like "`what is`", "`what color of`" are extremely frequent, while some like "`is she`", "`how much`" are very few (long-tail). Also, in visual relation annotation task, recalling frequent annotation "`riding`" for "`human`" and "`horse`" (rather than "`human- sitting on- horse`") for the long-tailed data distribution [10] is easily predicted. Since parameters are tuned by data during the learning process, even very deep neural networks are aligned to the most frequent samples in the training set, and it is dissimilar to a human-like reasoning system.

To tackle the aforementioned problem that is caused by dissimilarity in data distributions (DDD), a series of efficient works have been attempted to deal with the output instability of DNNs by utilizing either *robust learning techniques* or *unbiased feature representations*. Semi-supervised learning (SSL) aims to learn an unbiased model with a limited number of labeled samples. Due to the dissimilarity in data distributions, interest in training neural networks using semi-supervised methods has increased [29–31]. Note that the terms "*Robustness*" or "*Instability*" of DNNs refer to the tackling of *bias* prediction problem. However, for different tasks Robustness may deal with different limitations. For example, robustness can also refer to the model's ability to deal with attack [32] or external noise.

This dissertation addresses the overfitting issue from a representation learning perspective for two challenging tasks: Scene Graph Generation and Visual Question Answering that need deeper-level of understandings. we denote the "*high-level*" representation as unbiased feature representation and categorize it into discriminative and consistent representations. Discrimination helps the model to distinguish between different classes, while consistency allows the model to be stable within the same classes. Besides, this thesis focuses on high-level feature representations learning – for single and multi-modal CV tasks– therefore Deep Neural Networks will be our main use case in this thesis (Section. 1.2).

In the next section we review deep networks models we have used in this thesis. We then proceed in Section to provide an overview of the representation learning and outline a number of challenges in deep representation learning.

1.2 Deep Neural Networks

The intuitive concept of neural networks (NNs) refers to simple connected nodes (neurons), each producing a sequence of activations. Nodes are activated by their inputs based on the weights that belong to their connections. Hence, these weights are then associated with each connection together with input, and give the node activation.

In modern neural networks, introduced by Schmidhuber [33], nodes are organized into multiple (deeper) layers. For each layer, a non-linear transformation applies to that layer's input, which is associated with weights from previous connections. For example, each node in layer i is connected to all nodes in layer $i-1$. Notably, the term of "*depth*" refers to the multi-layer dedicated to non-linear transformations' layers in each DNN. The node activation provides feature representation of the input data, where each transformation leads towards better representation learning and attains the desired output [34]. In the literature, deep neural network models often divided into two types: multi-layer perceptrons (MLPs) and convolutional neural networks (ConvNets). MLP takes as input an observation x and passes it through a series of layers. The input to each layer i, gets pre-multiplied by a matrix of learnable weights W_i. Finally, a nonlinear activation function a_i is applied.

$$\begin{aligned} f_i(x) &= a_i(W_i.f_{i-1}(x) + b_i) \\ &\vdots \\ f_I(x) &= (W_I.f_{I-1}(x) + b_I) \end{aligned} \tag{1.1}$$

f_I denotes the feature representation for input x.

ConvNets are extensions of MLPs whose inputs are generally multi-dimensional. Each layer of ConvNets contains three components as convolutions, nonlinear, and pooling. Hence, the extension from MLP in Equation.1.1 to ConvNets is demonstrated in Equation.1.2.

$$\begin{aligned} f_i(x) &= c_i(a_i(W_i E_i(f_{i-1}(x)) + b_i\mathbb{I})) \\ &\vdots \\ f_I(x) &= W_I.f_{I-1}(x) + b_I\mathbb{I} \end{aligned} \tag{1.2}$$

where $\mathbb{I}$ is a vector of 1's of input dimension of each layer. A common nonlinear function is the ReLU, mentioned as a. c denotes a pooling function (*i.e.*max pooling

and average pooling). Furthermore, each row of W is often called a filter. Each row of f is called a feature map or a channel.

In contrast with traditional NNs, DNN-based computer vision systems, by overcoming the limitations of manual feature reasoning (*e.g.*, [19, 25, 35, 36]), have achieved remarkable progress. More importantly, they could learn representations that are useful for further tasks (*e.g.*, [24, 37]).

1.3 Representation Learning

Representation learning refers to the process of learning a parametric mapping from the raw input data domain to a feature vector or tensor. The process tries to capture related and useful concepts that can improve performance on a range of downstream tasks. One of the key ingredients for the success of deep learning is the ability to learn and extract through deep layers, some useful features from raw data. To date, varies deep learning techniques have been applied to learn semantic features.

In the literature, approaches to learning representations of data are often divided into two main categories: generative or discriminative modelling. Discriminative approaches learn representations by directly modelling the conditional distribution $p(y|x)$ with a parameterised model that takes the data sample x as an input while providing the label variable y as output. In contrast, generative approaches do not include labels y and learn representations by modelling the data distribution $p(x)$. It is based on the assumption that a good model $p(x)$ that can generate realistic data samples, must also in turn capture the underlying structure related to the explanatory variables y. The evaluation of conditional distribution $p(y|x)$ for some discriminative tasks can then be obtained by using Bayes' rule.

In general, either of the presented models can work properly based on the assumption that an useful feature representation will be captured by the encoder. Hence, the task of learning a good decoder/generator can be linked with the task of learning a good feature encoder. Hence, many ML efforts have concentrated on engineering input features, the primary goal of which is to enrich the input data based on prior knowledge of relevant aspects to solve a given task [38].

Due to feature representation, existing approaches have focused on content-based features, context-based features, or both [39]. For example, content-based representations are mostly used for visual tasks, while context-based features are generated from external context (*e.g.* annotations). Most past efforts in feature representation have been focused on the content of inputs *e.g.* in traditional vision-based content features [38, 40, 41], in deep learning-based methods [42–46]. Compared to traditional content-based features, content features are extracted from the pre-trained deep learning methods [15, 47–51]. As mentioned earlier, all of the proposed mainstream approaches to learning effective visual representations can be categorized into one of two general classes: generative or discriminative. Finally, we introduce one additional terminology commonly used when referring to semantic and robust feature representation. Robust feature representation problem is about encoding and then combining either discriminative, or consistent or both information to provide useful representation for final predictor [52]. All approaches are described in Sections 1.3.1, 1.3.2 and 1.3.3.

1.3.1 Discriminative Representation

Achieving effective prediction requires learning the subset of relevant representations. A representation that focuses only on relevant information content is less sensitive to noise, more discriminative across imbalanced/biased data, and more optimized to learn. Prior efforts proposed effective representation through discriminative approaches [3, 4, 53–56]. However, extracting the discriminative features from the input samples can make learning easier by separating representations of different samples, and help design effective predictors to distinguish different class labels. In general, the discriminative reasoning task stands for both visual understanding and semantic representation as a critical stage in many AI systems' ability to perform well. Discriminative visual features are extracted based on "context-based" information such as pixels, colors, objects, "context-based" information such as image tags, or "both". In the following, we introduce the mentioned types of discriminative features.

Content-based Feature Extraction- To extract semantic content-based features of images, deep features from the intermediate layers of pre-trained deep learning models have been used. Intermediate layers in well-known models such as

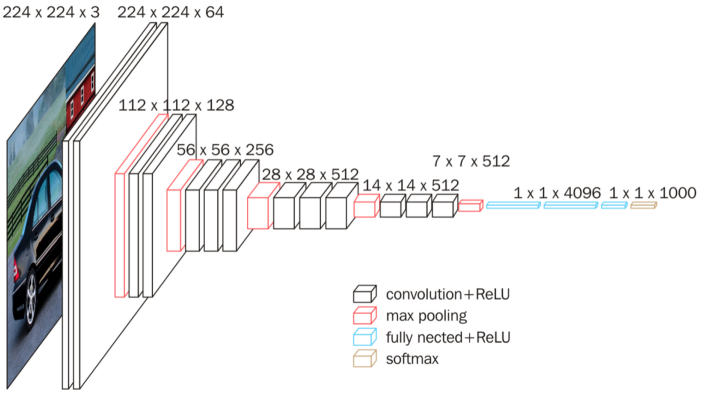

FIGURE 1.2: **VGG structure** [1]

.

VGG [1], ResNet152 [15], GoogleNet [57], CNN-LSTM [58] extract semantic features of images. Although these content features are shown to be highly efficient in many visual tasks (like [42–46, 58, 59]), the information they learn may be insufficient to discriminate complex images with inter-class similarities and intra-class dissimilarities, such as in higher-level tasks that need deeper understanding, like visual relation annotation and complex scene recognition [39, 60].

Some works, such as [49, 51], investigated separable representations derived from the fusion of different or higher pooling layers of GoogleNet [57] VGG16 [1] models. Figure. 1.2 shows the pooling layers of the VGG structure. However, these methods demand massive datasets and a rigorous hyper-parameter tuning process to learn the separable features for the better discrimination of each input image. Their methods ignore object-level information, which could be critical for better differentiation of *i.e.*complex scene images.

Aside from them, other works extracted mid-level representations by augmenting data with pre-trained deep learning models [50, 61]. For instance, Zhang *et al.* [50] performed random cropping of images into multiple crops and extracted the visual features from the AlexNet [25] model. Then, visual features that are extracted from multiple crops are concatenated to be learnt. However, such methods suffer from high feature dimensionality [39]. More importantly, while data augmentation increases the training sample diversity, the performance of these methods is highly limited by the chance of random choices. However, most of the proposed content-based solutions [42–45] are trained well only with frequent samples.

Context-based Feature Extraction- In comparison with content-based feature extractions, there are very few works [39, 50, 62] that have used context information by adding annotations or descriptions corresponding to images. Then, by fusing the external context information with the content information, more discriminating information can be offered. However, finding such external knowledge is impossible for some kinds of data, and generating them is expensive. Although context features can provide more unbiased representation for unbiased prediction, content features have a better representation ability for the non-ambiguous information present, *e.g.* in scene image datasets.

1.3.2 Generative Representation

The concept of "generative representation" mainly refers to the model's being robust against adversarial attacks [63]. Generative models can be thought of as self-supervised but with different objectives. In this thesis, we extend from discriminative to generative features towards robust visual question answering task.

Deep feature by Adversarial Training For DNNs, generalization is a significant characteristic. The majority of efforts in this community are focused on improving adversarial training in the presence of noise or attacks. The studies on the generalization of adversarial training are mainly categorized into standard generalization, adversarially robust generalization, and generalization on unseen attacks.

In some cases, they produce probabilistic predictions; in others, they only yield the most likely class to assign. In all cases, they approach the classification problem without explicitly modelling any of the data-generating distributions. In contrast, the primary goal of methods based on generative models is to model the process that generated the data. When such a generative model is conditioned on a given label y, it can also be used for classification.

As the first research direction, we propose a novel head network to actively explore discriminative content-based representation using feature refinement (Chapter 3). The second research direction addresses the model instability problem by generating consensus representation for the same class. Through a consensus model, common feature representations that are reasoned by various samples are learnt to increase the robustness of consensus (Chapter 4).

1.3.3 Learning Consistent Feature Representation

Besides two mentioned representations, from model robustness perspective, the consistency refers to that the discriminator' outputs do not change too much with respect to small changes in the input.

After learning a new task, the representation of the old relation in the space may change. In order to prevent the encoder from not altering the knowledge of the old task while learning the new task, we propose two replay strategies to learn consistent representation for alleviating this problem: contrastive replay and knowledge distillation.

Another perspective of model robustness is that it is consistent across similar classes of samples. From an ML model perspective, consistency means that given semantically similar inputs, the user would expect similar outputs, or in other words, consistently correct outputs. In the pioneering work by [29], consistency regularization has been introduced as an essential assumption in which a robust model should output predictions on the perturbed versions of the same input image [64]. Moreover, the importance of "*visual consistency*" has emerged in unsupervised contrastive learning. The similarity of the positive sample(s) and reference sample has been taken into account to make the learning more efficient [64, 65].

In the second part of this thesis, we argue that consistency in visual representation encourages consistent similarities between generated variations of samples that are semantically similar. In particular, we have improved the consistency not only for visual variations but also for language paraphrases (in Chapter 4.1).

1.4 Objectives and Major Contributions

The contributions of this dissertation are two-fold. First, in Chapter 3 we propose a meta-architecture approach to learning deep feature representations for visual relation annotation and scene graph generation. Previous works in this area learned shallow features or our approach which is frequentist and inspired by the algorithm in Chapter 3. This work with any level of supervision: the algorithm can take advantage of sequences that change points are fully or partially labeled. Alternatively, it can still learn feature representations in an entirely unsupervised manner.

Chapter 4, we seek robust visual reasoning to present a consistent learning paradigm for multi-modal reasoning tasks such as visual question answering, to overcome some limitations toward robustness. To construct the proposed robust model against input variations, ADVCL, we utilise contrastive learning approach to contrast intra-class samples with negative samples.

The major contributions are listed as follows:

For learning discriminative representation, we propose a model to explore semantic content-based representation of the ground truth visual input that enhances the model efficiency within fine-grained prediction.
For example, to construct a sensible scene graph, models should be able to distinguish more informative relationships rather than the 'more probable but less informative' ones (*e.g.* `person-on-skateboard` replaced by `person-sitting on-skateboard`) by capturing more contextual semantics. Our proposed model deals with fine-grained prediction and achieves improvement in the comprehensive metric [5, 66], mean Recall@K (mR@K) on VG [67] benchmark. For example, our model improves the mR@50 and mR@100 from 8.0% and 8.5% in MOTIFS [6] to 9.7% and 10.2%, and from 7.5% and 7.9% in VCTree [5] to 11.1% and 11.7% on the scene graph classification task, respectively.

We progress from discriminative to generative features in order to solve robust visual question answering problems.
For training a model to learn the representations of the ground truth multi-modal input, we apply a contrastive learning approach using *hard*-samples generated with adversarial perturbations.
We show that generating hard representation based on adversarial attacks leads the model to be consistent against semantic data variations.

1.5 Structure of the Dissertation

- In Chapter 1, we introduce the general idea of robust feature representation and learning for DNN models and the reason of why they are important in the computer vision community. Afterwards, we discuss the importance of two perspective of robustness to the motivation and basic ideas of our proposed methods in the following chapters.

- In Chapter 2, we review all the related works and preliminary concepts that are critical in this thesis.

- In Chapter 3, we present the design of a meta-architecture that explores the discriminative pairwise features by novel content-based feature representation, helping visual reasoning tasks like visual relation annotation and scene graph generation to overcome naive object feature concatenation limitation and dataset bias. To construct the proposed pairwise head network, ALIGN R-CNN, we design a novel attention-based multiple region alignment module that can be jointly optimized with SGG. We develop a learning scheme which actively refines pairwise features due to ground-truth relationship labels. Experimental results on large-scale and biased benchmark: Visual Genome for scene graph generation show conclude the importance of Feature alignment and how to use it for the guidance of future designs of discriminative models. In comparison with other methods, ALIGN R-CNN outperforms state-of-the-art results, while exploring interpretable visual semantic representation. Note that our pairwise head network ALIGN R-CNN can be widely applied in the community which particularly seeks unbiased and zero-shot predictions.

- In Chapter 4, we seek robust visual reasoning to present a consistent learning paradigm for multi-modal reasoning tasks such as visual question answering, to overcome some limitations toward robustness. To construct the proposed robust model against input variations, ADVCL, we utilise contrastive learning approach to contrast intra-class samples with negative samples. Moreover, we argue the naive contrastive positive and negative samples that are randomly picked are sub-optimal. Hence, to overcome this issue, we generate hard positive samples from adversarial attack samples. Experimental results on two benchmarks, VQA2.0 and Praphrased visual Q&A, show that ADVCL is robust to question variations and less-frequent data.

- In Chapter 5, we draw the conclusion of the overall thesis by summarizing the contributions of this thesis on robust feature representation and model robustness. Finally, we discus the importance of causality in DNN models and how it can improve the robustness for our future works.

Chapter 2

Literature Review

2.1 Introduction

In this section, we provide the review and analysis of the related tasks to the proposed methods in this thesis according to the following strategies.
First, we review two practical problems, including Scene Graph Generation, and Visual Question Answering which are described in Section 2.2 and Section 2.3, respectively. Second, the learning methods for feature representations that can be applied to many different practical problems have been analyzed. We categorise these methods into "Zero-Shot Learning" and "Contrastive Learning", which have been reviewed in Section 2.5 and Section 2.6, respectively.

2.2 Scene Graph Generation

Scene graph [68, 69] is a comprehensive and coherent, visually-grounded structure representation which explicitly models objects (*e.g.* boy, umbrella, bike), attributes of objects (`umbrella-is- green`), and relationships between paired objects (`boy-hold-umbrella`). Specifically, the fundamental elements of a scene graph are objects, attributes and relationships where the substructure is annotated in the form of relationship triplets, <`subject-relation-object`> or <`subject- relation-attribute`>, abridged as <o_1`, r,` o_2>. "`subject1`" and "`subject2`" are assigned with object classes. Objects are the fundamental building blocks of an input image

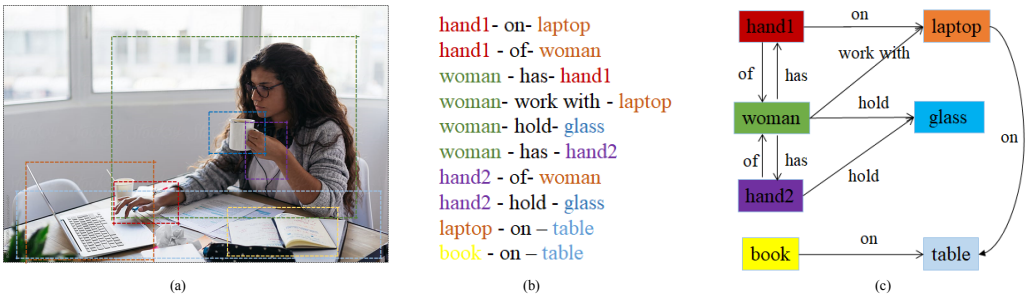

FIGURE 2.1: **Image with annotations of different semantic levels** Visual Relationship Detection: Given an image as input, we detect multiple relationships in the form of <Subj-Rel-Obj>. (a) For each triplet, both objects are localized in the image as bounding boxes. In this example, we detect the following relationships (b) and finally generate the graph (c).

that can be located using a set of bounding-boxes (BB). An object may contain zero or more attributes, which can be color (*e.g.* green, red), states (*e.g.* standing), *etc.* Relation is the predicate label of the node pair such as `person-ride-bike` and `dog-on-bed`. The relation can be categorized as spatial (*e.g.* above, behind), actions (*e.g.* looking, walking), descriptive verbs (*e.g.* wearing), prepositions (*e.g.* with), and comparative (*e.g.* smaller than), *etc.* Figure. 3.1 shows the straight-forward model to generate a graph from input image by extracting object bounding-boxes and pairwise relation features from each pairs of object.

Scene Graph Generation (in short SGG) has attracted much attentions in multimedia community [5, 6, 9, 70–79], especially due to its potential to underpin many multi-modal downstream tasks, such as image captioning [80–83], VQA [84–87], and image retrieval [68, 88]. In general, two approaches have been proposed to generate scene graph. The straight-forward approach, known as compositional approach, [6, 9, 68, 69, 74, 89–96] first detects objects and then classifies the relation of each object pair. The second group of approaches [6, 7, 9, 66, 75, 97–103] jointly infer the objects and their relations based on the object region proposals. The main difference between the two groups of approaches is whether the relation features (features of each pair's union area/ pairwise object feature) are used to refine the object features. To generate a complete scene graph, both approaches above should detect all existing objects or object proposals in the image as far as possible, and group them into pairs and use the features of their union area (denoted as relation features), as the basic representation for predicate inference. In this thesis, we consider compositional approach. Given an input image, scene graph generation method first generally generates triplet proposals with Region Proposal Network

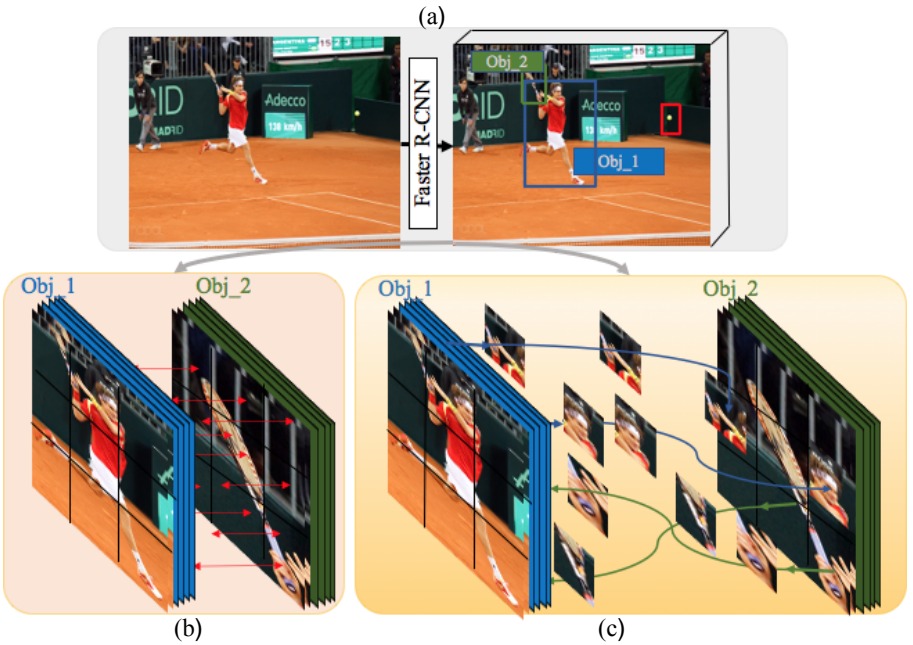

FIGURE 2.2: **Pairwise feature embeddings** [2].(a) Given paired objects (Person and Racket); (b) pairwise relation method in which feature maps are directly concatenated and (c)the proposed Align R-CNN relation method in which discriminative/relationship-specific object parts are aligned before concatenation

(RPN). Each triplet proposal is made up of subject, object and predicate ROIs, respectively. The predicate ROI is the box that tightly covers both the subject and the object. Then, in feature representation, for each object proposal, we can get appearance, spatial information, label, depth, mask, and for each predicate proposal, we can get the appearance, spatial, depth and mask. These multi-modal features are vectorized and can be combined and refined in the third step of Feature Refinement using message passing mechanisms, attention mechanisms and visual translation embedding approaches. Finally, the classifiers are used to predict the categories of objects and predicates, and the scene graph is generated.

In this section, SGG models will be reviewed and analyzed according to "feature representation" and "feature refinment". First, we review the feature representation methods for objects, subjects and predicates in SGG. We classify pairwise object feature modeling for relationship prediction into "Triplet-based" and "Context-based" categories. Second, the feature refinement methods for objects, subject and predicate are presented. We categorise these methods into "Message Passing", "Attention Mechanism", "Visual Embedding" and Others.

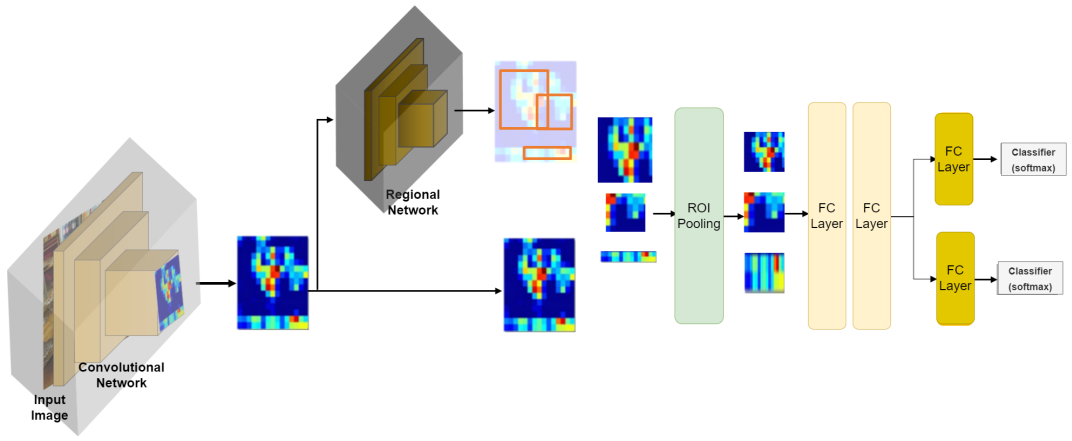

FIGURE 2.3: **Faster-RCNN network for object detection.**

2.2.1 Object Detection

Object detection task plays an important role in the applications where one needs to localize the position of objects in input images or video frames and recognizing the category of the objects before further processing. Recent advances in deep learning techniques for object detection is closing the gap with respect to human performance [17–19]. It is also very important for many CV tasks such as scene detection, robotics, autonomous car driving, and *etc.* These applications depend on several task-specific predictors and estimators to provide accurate predictions. To ensure that the extracted features are accurate and is not affected by the unrelated task-specific information, the designing of a robust model which can handle task-specific noise or adversarial manipulation in the input data, has became critical in advanced applications.

The most popular object detection pipelines [17, 104] involves generating a number of object proposals classifying each of them. In this thesis, for both applications we use Faster-RCNN [17] as a detector. As seen in Figure. 2.3, the framework of Faster-RCNN is is a two-stage network for object detection. In Figure. 2.3, object proposals are produced based on feature maps that are produced in the network. Then, the generated proposals are passed to several fully connected (FC) layers to output the objects' bounding-boxes and the class labels corresponded to objects.

2.2.2 Feature Representations

The focus of the SGG is to use object visual features to generate a powerful pairwise object feature representation model to connect objects for high-level reasoning [75, 80, 85, 105]. The simplest way to predict relationships in the SGG is to take the original features or their concatenation (extracted in the feature extraction stage of the general process depicted in Figure 2.2.b) as inputs to produce a confidence score for each relationship category of the classifier. Hence, the core technique for SGG is pairwise object feature modeling for relationship prediction, which can be summarized into two main categories as follows:

Triplet-based- Triplet-based modelings [69, 93, 106–110] explore the information from each given pair of objects. These approaches first detect objects using pretrained detector, and then use the outputs of the object detector as fixed inputs of a relationship detection module. These models are merely local and thus overlook the beneficial global context.

Context-based- Unlike triplet-based methods, context-based methods [5, 9, 66, 71–74, 76–78, 111, 112] enrich the triplet-based approach with contextual features by using graph message passing. Visual context-based methods that inferring object-level context reasoning provide a powerful bias that connect objects for high-level reasoning [75, 80, 85, 105]. For example, detecting a pair of objects "`person`" and "`bike`" can be informative to detect the predicate "`ride`", which preserves information to answer the question "`who is riding on the bike?`" by "`person`" [5]. However, these two prior structures are sub-optimal and easily fooled by biased dataset.

The common core of the two categories mentioned above is the pairwise feature modelling, that is to say, any improvement in pairwise modeling can directly improve the performances of any relationship prediction model. Unfortunately, although SGG techniques have advanced significantly in recent years, the pairwise feature is still the naive feature concatenation. As shown in Figure. 2.2 (b), the naive pairwise feature is built based on the fusion of individual object features per each pair of objects. However, we argue that the naive concatenation suffers from the following limitations:

Lack of Object Part Interactions. The object features abandon spatial cues and hence their concatenation cannot encode any object part interactions. For example, to distinguish between the two predicates `riding` and `carrying` for the relationship `person-?-surfboard`, concatenation can only provide the class information about `person` and `surfboard`, thus, the resultant prediction is inevitably biased to the most frequent combination, *e.g.* `person-carry- surfboard`, without even looking at the visual interactions [5].

Mis-alignment of Object Parts. One may want to directly concatenate the two feature maps that retain the spatial cues [6, 113]. However, we argue that such concatenation are fully object specific and may *mis-align* the object parts. Hence, they may fail in complex visual features due to the mis-aligned object parts. Opposed to those works, our proposed model exhibits different pairwise features by mining *region alignment* for different object interactions. For example, to distinguish between `holding` in `person-racket` (see Figure. 2.2), if we directly concatenate the feature maps, the parts of `hand` object cannot align to the corresponding `racket` parts, leading to non-discriminative features for prediction.

2.2.3 Feature Refinement

The idea of learning better features by feature refining is extensively studied in scene graph generation [9, 66, 75]. The intuition behind it is to exploit the fruitful contextual information for both object and relation detection in a Visual Relation Detection (VRD) task. Regarding relation detection, the previous works pre-estimate dependencies among triplet <object1-relationship-object2> components and make the task context-dependent. For instance, the refined features lead the predictor to predict more biased relations, *e.g.* in `person-?-bike`, `riding` is indeed more prominent than `feeding` [2]. However, graph initialization is essential here before going through node and/or edge refinement.

2.2.3.1 Message Passing Mechanism

The message passing explore the effect of context occurring between elements of the scene graph, including objects and relationships. We can understand the effect

of context at three levels. First, for a triplet, the predictions of different phrase components are dependent on each other. For example, the visual connection of the subject (man), which appears to be sitting on something and an object (horse) with the appearance of human on it help to enhance the evidence of the predicate "ride". In return, the specific visual features for "ride" also helps to infer the subject (man) and object (horse) as well. Second, since the triplets are not isolated, messages can be passed between them. The message passing at the subgraph level is based on the assumption that the objects which have relationships are semantically dependent, and the relationships which have overlapped object(s) are also semantically related to each other and even share appearance features. Third, visual relationships are image-specific, thus learning feature representations from a global view is meaningful for relationship prediction. The global information is scattered over each object, specifically, for each object proposal generated by RPN, other proposals all contain its contextual information.

2.2.3.2 Attention Mechanism

Machine learning-based Attention was first proposed by [114] by applying Neural Machine Translation that simultaneously learns to align and translate. Attention mostly attempts to optimize the task of prediction by exploring key visual features and passing information only between embedded nodes.

In the task of scene graph generation, attention mechanisms are always used to refine object features and extract relationship features. In most multi-modal alignment-based works, attention mechanism plays an important role in identifying the word alignments with respect to vision features and vice versa. The benefit of such approaches is the feature semantic representation reasoned by alignment, which provides the model with flexibility to attend to non-strict regions. Unlike multi-modal approaches, in this thesis we introduce the head network that applies a region alignment module to calculate the alignments with respect to semantic-rich regions belonging to the correspondence object in a pair for pairwise feature context selection (see Chapter 3).

In multi-modal task, for example, by given a set of image features $V \in \mathbb{R}^{n \times d_v}$ and bounding box features $B \in \mathbb{R}^{n \times d_b}$ and textual features $E \in \mathbb{R}^{m \times d_e}$, the root attention has two roles. The first role is to generate an attention map for object

level visual features based on language representations:

$$\alpha^{object} = softmax(ATT(V, B, E)) \tag{2.1}$$

$$Head : H_i = A_i.V.W_i^v \tag{2.2}$$

$$Ouput : V^{ATT} = Embed([H_1, ..., H_h]W^H), \tag{2.3}$$

where W is all trainable parameter; A denotes the soft- attention multi-lead $H = 1, ..., h$, and [.] is the concatenation operation. The network function is shown as $Embed()$ and the residual operation as in:

$$O = \alpha^{object}.V^T \tag{2.4}$$

ATT denotes any network functions which can generate object region attentions. The second role of root attention is to generate a fused visual feature V^{ATT}.
The attention network function ATT can be categorised as *soft* or *hard* architecture, based on the alignment score. The soft-attention network re-weights and learns all the global context feature for each time [114], over all patches in the input. While, hard-attention approach learns the single patch at any time. In multi-modality task, hard-attention passing only takes "hard" representation's weight that are related to input queries (*e.g.* input visual questions). In Chapter 3 by inspiring [115, 116], we take the benefits of soft parsing attention to consider the context of two objects per pair. Then we apply hard-attention to negate the irrelated region information as pairwise feature for prediction.

Towards semantic scene understanding Semantic scene understanding requires identification of the type of objects and their visual relationships. Thanks to progress in deep learning (DL) approaches, semantic scene understanding has seen substantial progress through advances in image classification, semantic segmentation, relation detection and annotation, *etc.*

In the process of human reasoning, abstract reasoning of a visual scene is to try to ignore irrelevant details. Inspired by this, visual understanding using high-level; semantic representation is introduce. In visual reasoning, the achievement of deep learning significantly improved the accuracy of results. Image features are primarily used as input to get answers. However, the image features are too redundant to learn accurate characterizations within a limited complexity and time. The model

using semantic representation as input verifies that more accurate results can be obtained by introducing a high-level semantic representation.

In this thesis, we modify previous representation learning approach to facilitate model building in this way and demonstrate how it can reason semantic visual feature between pairs of objects to efficiently learn about their visual interactions.

Furthermore, [93, 107, 109, 111, 117, 118] propose relationship proposal networks by employing pairwise regions in images for fully or weakly supervised visual relation annotation. However, most of them are designed for whole scene graph description.

This thesis aims to propose a model uses an attention-based module to encode the semantic relation between object regions and classes. Subsequently, it transfers information to sparse regions to make a prediction that is closer to the features extracted from the informative regions than the features from different classes. Some attention-based works like [119, 120] have been known as the efficient relational reasoning models over the past few years. The attention-based methods can model dependency between the triplet components, aggregating information from the feature embeddings of all pairs from the input (*e.g.* pixels). The aggregation weights are learned and driven by the target task. [121] build a commonsense knowledge graph for this objective.
Close to [121, 122], however, our work considers object part instances from object feature maps instead of image pixels as the primitive elements. Besides, the proposed model by this thesis, learns discriminative pairwise features by capturing the local semantic dependency among object instances.

2.3 Visual Question Answering

The task of Visual question answering (VQA) was first introduced by [123] to gain a deep understanding of visual content by bringing together advancements in natural language processing (NLP) and computer vision (CV). Image-based QA [12] requires wide understanding of input multi-modal image-question pair to predict a correct answer. In the terms of performance, there has been significant progress over recent years [124–135] by either applying an attention module [129, 136] or by enhancing the reasoning of the joint representation [124, 125, 130–135]. To extract visual feature representations, early works used pre-trained VGG

or ResNet. Later, bottom-up and top-down network [129] enhanced the visual features by extracting from an object detector [17].

Further efforts [124, 126, 128, 137–139] extended traditional attention methods for better visual representations. For instance, co-attention [128, 140, 141] implements soft-attention based on a fused representation of multi-modal feature representation. Most modern and advanced attention methods proposed models like BAN [126], DCN [141]. Inter-intra network [127] further improved computation efficiency of co-attention through bilinear attention. More recent works by [8, 130–134] have applied multi-modal transformers (MMT) to two modalities' representations from detected object features and BERT language features [142]. Similar to [130], in this paper we use the multi-modal transformer architecture for our experiments. Moreover, many re-generated VQA datasets including human-crafted datasets like CLEVR [13], and large-scale long-tailed datasets like VQA v1.0 [12] and VQA v2.0 [143] have been proposed.

Visual representation reasoning, which is the main core of this thesis, is critical to perform VQA model robustness, for variations of language questions as well as biased and complicated questions which are needed to high-level reasoning to answer. Despite these significant progress, balancing between accuracy as well as robustness of today's VQA models against input variations leave much to be desired. In this thesis, we show that the desirable feature consistency leads VQA model to be robust and efficient. These methods significantly improve overall VQA performance. It is worth noticing that conventional VQA methods lack the ability of produce robust prediction against multi-modal input variations.

2.3.1 Robustness in Visual Question Answering

Robustness of VQA models concerning multi-modal vision and language input has been studied from various aspects.

Some efforts like [144–147] proposed robust VQA models with respect to the language bias caused by the available imbalanced training dataset. For instance, "what is color of apple" in most cases predict "red" due to imbalance training set. To tackle the lack of robustness, some past works proposed new benchmarks [11, 144]. VQA-CP [144], the first robust VQA benchmark, has proposed a balanced dataset

to evaluate question-oriented language bias in VQA models. It has been built by reshuffling data in VQA v2.0 [143], and qualifies language bias affected by input questions. GQA-OOD [147] improves from VQA-CP, and proposes to study robustness against the question-answer distributions.

Besides language bias, VQA-Rephrasings [11] exposes the brittleness of VQA models to linguistic variations in input questions. VQA-Rephrasings made by human-written rephrasings of questions in VQA v2.0 dataset. [11, 148] studied the VQA robustness in relation to question paraphrases and subsequently proposed two augmented datasets by generating various rephrasings of questions. VQA-CC [11] studied the VQA robustness in relation to question paraphrases and trained a Visual Question Generation (VQG) model to generate paraphrases of questions to augmented the training dataset. VQA-Aug [148] augmented the training dataset by generating paraphrases of questions via Back-Translation (BT). While, VQA-P2 ([149]) augmented data with low-level changes such as simple lexical substitutions. In the Chapter 4, we study the consistency of VQA prediction on reference questions and their corresponding paraphrases. In contrast to previous approaches, our approach is agnostic to model architecture.

Apart from previous perspectives of VQA robustness, [146, 150] proposed the extension of visual grounding in VQA models with respect to semantic changes in input images. Causal VQA [146] analysed the predictor robustness for prediction consistency to questions from reference images and their corresponding noisy images. Further studies investigate robustness against reasoning. Robustness across reasoning analyses the model ability across variety of situations *i.e.* compositional and positional reasoning. For instance, [151] evaluated the robustness through special compositions of different types of questions. GQA [152] studied and qualified VQA models' ability on positional reasoning as well as relational reasoning by generating questions from class-label scene graphs. A work by [153] introduced a novel split for VQA dataset via providing perception-related questions for reference questions in dataset.

2.4 Deep Multi-modal Learning

In this section, we summarize two perspectives from the current literature on deep multimodal learning, namely: multimodal data representation, multimodal fusion (*i.e.*both traditional and deep learning-based schemes), multitask learning, multimodal alignment, and zero-shot learning.

2.5 Zero-Shot Learning

Zero-shot (zS) learning is a task which aims to recognize a *i.e.* object, class whose instances may not have been seen in the training set [94, 154–158], but in test set. Due to the importance of zero-shot learning due to the overall performance improvement, in the case where there is not enough labeled data for some samples in test set in training data, the number of proposed zero-shot learning models has increased during the past few years.

2.5.1 Zero-Shot Learning in Scene Graph Generation

In the task of scene graph generation, zero-shot recall aims to recognize unseen triplet components with no labelled instances in the training set by leveraging supplementary information, such as semantic representations. For instance, the Visual Genome (VG) dataset [67] contains 2,098 components that never occur in the training set, hence they can be used for zero-shot evaluations [159]. For instance, `cat-looking at-animal` component has never seen in training set. Tang *et al.* [10] reported zero-shot problem on VG for the first time. However, they didn't propose any particular solutions to improve it.

To tackle the problem of zero-shot learning by inspiring human reasoning systems, we propose a novel pairwise head encoder inferring robust and discriminative relation features. The proposed network is able to detect unseen components (*e.g.* `cat-looking at-animal`) by composing and grounding existing visual concepts from other patterns (3).

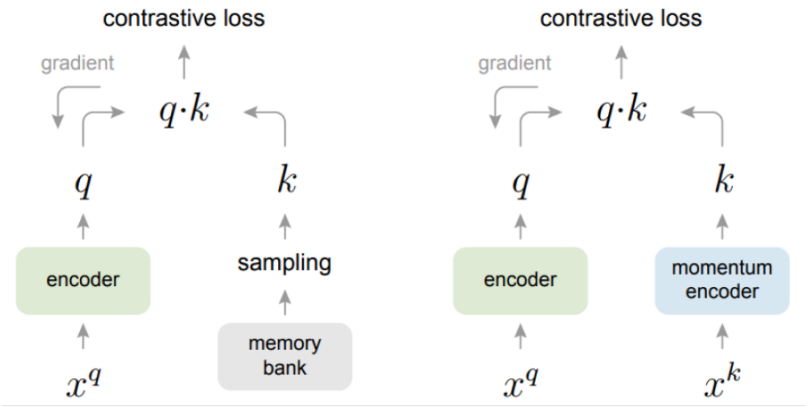

FIGURE 2.4: **Conceptual comparison of key representations from memory bank [3] and on-the fly [4].**

2.6 Contrastive Representation Learning

Contrastive learning is a technique that aims to learn consistent and high-level representations by distinguishing the positives from negative representations of data. Learning high-level visual representations by using contrastive learning [160] has achieved great success in both supervised and unsupervised manners [4, 160, 161]. Specifically, some contrastive learning methods like [4] adopt contrastive learning to train the models by decreasing the distance between the feature representations of different augmented views of the same input images while increasing the distance between different images. In this thesis, we use contrastive learning to push multi-modal features closer to debiased features while keeping them far away from negatively biased features (Chapter 4).

Semi-supervised Contrastive Loss

Contrastive learning has been known as main model for unsupervised visual feature reasoning and learning. An effective contrastive loss function is in Eq 2.5 [162]:

$$\mathcal{L}_{q,k^+,k^-} = -log\frac{exp(q.k^+)/\tau}{exp(q.k^+)/\tau + \sum_{k^-} exp(q.k^-)/\tau} \tag{2.5}$$

where q is a query of representation, k^+ denotes a positive representation sample, and k^- are representations of the negative key samples. τ is temperature hyper-parameter.

Supervised Contrastive Loss

One variation of a supervise contrastive loss (SCL) [163] enables joint multi-modal representations from samples whose questions are semantically paraphrased to be closer than the rest. In this thesis, we are inspired by [8] to optimize training paradigm. However, we discuss that the key challenge of contrastive learning for multi-modal representations benefits more from true samples (*i.e.* those that are semantically similar to an anchor point) and may be damaged from those negative samples with other anchors.

For a set of N, sample pairs are randomly selected, $\{\mathbf{x}_k,\ \mathbf{y}_k\}_{k=1...N}$, the corresponding batch that use for training contains $2N$ pairs, $\{\bar{\mathbf{x}}_l,\ \bar{\mathbf{y}}_l\}_{l=1...2N}$, where $\bar{\mathbf{x}}_{2k}$ and $\bar{\mathbf{x}}_{2k-1}$ are two random sample augmentations, *e.g.* pairs of $\mathbf{x}_k(k = 1...N)$ and $\bar{\mathbf{y}}_{2k-1} = \bar{\mathbf{y}}_{2k} = \mathbf{y}_k$.
As in Eq 2.6 [163] that shows the straight forward to generalize Eq. 2.5 to incorporate supervision.

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} -\frac{1}{|P_i|} \sum_{p \in P_i} log \frac{exp(z_i.z_p/\tau)}{\sum_{a \in A_i} exp(z_i.z_a/\tau)} \tag{2.6}$$

2.6.1 Contrastive Representation Learning for Visual Question Answering

Contrastive learning techniques have achieved great success in unsupervised learning [4, 160]. The main task under contrastive learning is to learn in a way that encoded features to be similar to positive samples while being far from dissimilar ones by employing the contrastive loss. In image-based applications [4, 164–167] , many efforts have been adopted for choosing/generating positive and negative samples that is critical for the quality representation.

According to the robustness of VQA models, (described in Section 2.3.1), [168] replaced discriminative loss function (*e.g.* cross-entropy) with robust loss to boost the generalization performance of the model in semi-supervised learning approach. The common approach that can increase the robustness is to simply replace the discriminative loss with a loss that is robust to distribution. However, due to the nature of contrastive loss, CE loss can be replaced by SCL that makes anchor and its paraphrasings closer.

Using the advantages of augmented training set, *SCL* has been applied to learn better representations from anchor and its paraphrased question. In supervised and metric learning settings, "hard" (true negative) key samples enable a learning method to correct its mistakes more quickly [169, 170]. In this thesis, we furthermore show the importance of positive samples by generating "hard" positive samples using adversarial attack samples instead of random intra-class samples (Chapter 4).

2.7 Summary

In this chapter, we provided some theoretical and background knowledge that are related to this the works in this these. In this thesis, we argue that while most research focuses on maximizing overall performance during training a machine learning model, not much attention has been given to evaluating its robustness, *i.e.*, against biased datasets and visual content manipulation, until very recent years. Lack of unbiased representation learning, especially with respect to consistency and discrimination, can be due to various reasons, *e.g.* data distributions, inadequacy in the learning process, model sensitivity to different regions of feature space, *etc.*

To mitigate the mentioned challenges, we have proposed different feature representation learning approaches to enhance learning performance for relation annotation and for question answering.
As a first contribution, we incorporate an unbiased learning approach in Chapter 3 by exploring the importance of learning feature alignment for dynamic object feature concatenation in relation annotation, where objects may not only have large overlaps in their spatial locations but also have a lot of commonalities in their semantic concepts.
In Chapter 4, we further explore unbiased learning by exploring the present learning paradigm of consistency in multi-modal reasoning tasks such as visual question answering, to overcome some limitations in robustness.

Chapter 3

Align R-CNN: A Pairwise Head Network for Visual Relationship Detection [1]

3.1 Introduction

3.1.1 Scene Graph Generation

Scene Graph Generation (SGG), or visual relationship detection (VRD) was first formalized by [95]. Scene graph generation has attracted much attention in the multimedia community [5, 6, 9, 70–79]. Due to these important needs, previous efforts have tried to provide a powerful *pairwise object feature* model to connect objects for better feature reasoning [75, 80, 85, 105].

However, in this thesis we argue that most existing methods of studying scene graph generation performance which are based on pairwise triplet-based prediction, may fail to disambiguate between single pair interacting in multiple ways (*e.g.* `person-horse` may interact by multi categories: `riding`, `feeding`, *etc.*) and/or multiple pairs interacting in the same way (*e.g.* predicate `holding` may be used for several pairs: `person-book`, `person-glass`, *etc.*). This problem stems

[1] The work in this chapter has been published in the paper : Mitra Tajrobehkar, Kaihua Tang, Hanwang Zhang, Joo-Hwee Lim. "Align R-CNN: A Pairwise Head Network for Visual Relationship Detection." Proceedings of the IEEE Transactions on Multimedia, 2021.

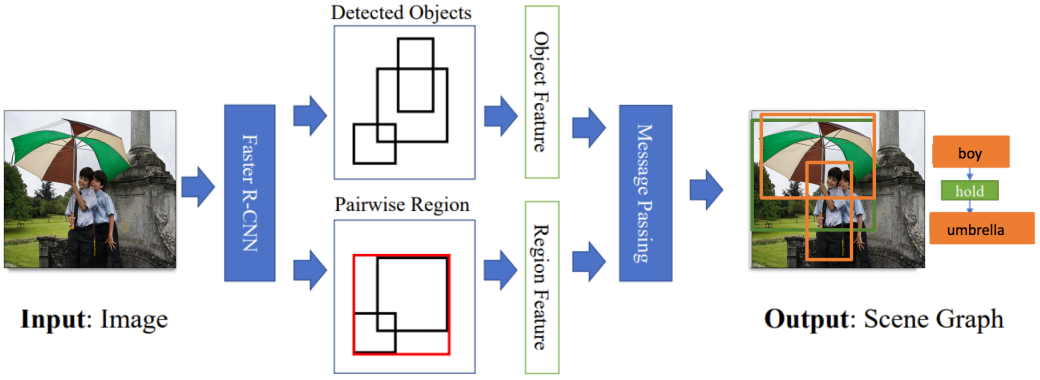

FIGURE 3.1: **A Common Framework of Scene Graph Generation**.

from non-discriminative relationship feature embeddings. In this chapter, we propose a network (ALIGN R-CNN) that aims to demonstrate that feature extraction in pairwise relation models can be improved by augmenting a pairwise head with the attention component for dynamic object feature concatenation. ALIGN R-CNN) innovates a feature representation approach to address the data distribution mismatch for semi-supervised learning.

We implement ALIGN R-CNN on the new standard of the SGG diagnosis toolkit that has been recently proposed by [10] and evaluate the Align R-CNN pairwise relation model by applying it to the state-of-the-art models [5–7, 9]. We evaluate ALIGN R-CNN on a large and well-known benchmark: Visual Genome [67]. We gain comparable improvement on standard tasks in SGG. In particular, ALIGN R-CNN helps high-level vision models fight against the dataset bias. To construct a sensible scene graph, models should be able to distinguish more informative relationships than more probable but less informative ones (*e.g.* `person-on-skateboard` replaced by `person-sitting on-skateboard`) by capturing more contextual semantics. The proposed ALIGN R-CNN deals with fine-grained prediction and achieves improvement in the comprehensive metric [5, 66], mean Recall@K (mR@K) on VG [67] benchmark. For example, our model improves the mR@50 and mR@100 from 8.0% and 8.5% in MOTIFS [6] to 9.7% and 10.2%, and from 7.5% and 7.9% in VCTree [5] to 11.1% and 11.7% on the scene graph classification task, respectively.

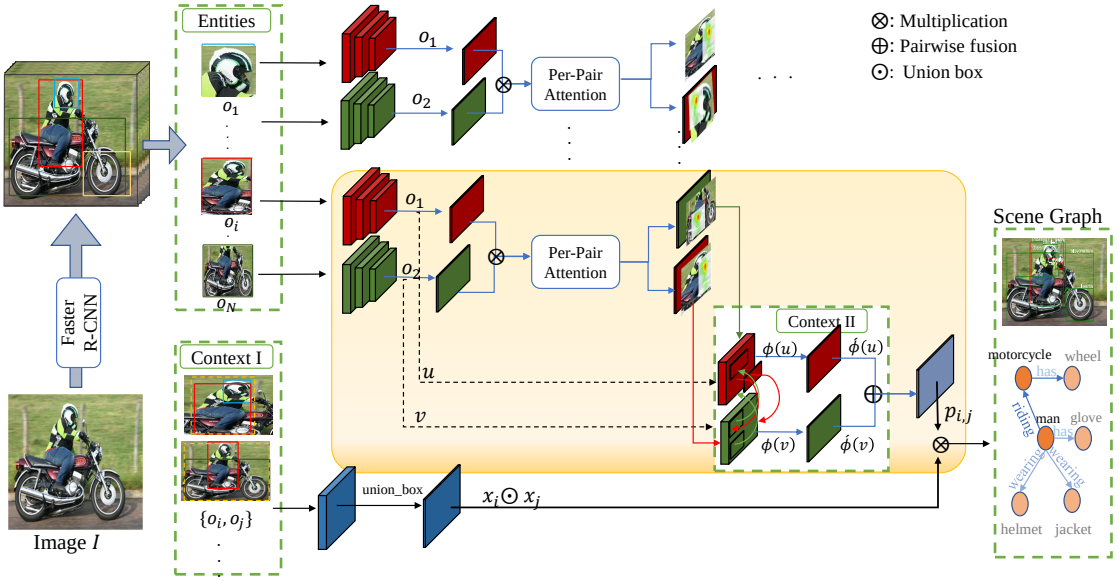

FIGURE 3.2: **Overview of proposed Align R-CNN framework**. The model contains attention-based multiple region alignment module to generate new pairwise visual features for relation prediction as showcased in the yellow box

3.2 Approach

Our proposed ALIGN R-CNN network can be summarized into the following stages. (1) Object detection (Section 3.2.1) to detect object proposals. The visual feature of detected object i is shown as $\boldsymbol{o}_i$, concatenating a RoIAlign feature [19] $\boldsymbol{v}_i \in \mathbb{R}^{2048}$. (2) Given a set of object proposals in an image, in Section 3.2.3, pairwise object feature and learning-by-Alignment conducts region-wise alignment for dynamic object feature concatenation. Then, we employ *Decoder* to decode the discriminative pairwise information using the constructed ALIGN R-CNN (Context II in Figure. 3.2). For *Decoder*, we adopt bidirectional LSTMs (BiLSTMs) (shown in Figure 3.3.b) , Bidirectional Tree LSTM (BiTreeLSTM) [171], and fully connected layers according to MOTIFS [6], VCTree [5] (shown in Figure 3.3.c), and VTransE [7], respectively to encode the content, through different experiments.

In MOTIFS, the Bi-LSTMs model is used to capture the global context and structural regularities in scene graphs. MOTIFS constructs the global context from all detected objects in a given image. Then, it leverages the global context to refine feature-level representations for individual objects and possible relationships between them.

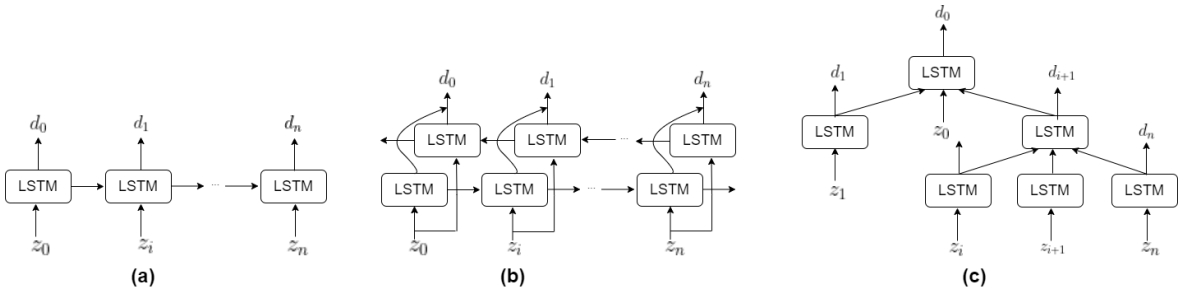

FIGURE 3.3: **a) Unidirectional LSTM (LSTM); b) Bidirectional LSTM (BiLSTM); and c) Bidirectional (BiTreeLSTM).**

(3) The encoded contexts will be decoded for graph generation in Section 3.2.4, that jointly utilizes the object feature embeddings (Context I) and aligned features (Context II), to generate the final graph. The full architecture diagram is shown in Figure 3.2

3.2.1 Object Detection and Feature Extraction

Consider the input image as $\boldsymbol{I}$, which is represented by a set of N objects and bounding-boxes in form of $O = \{o_1, ..., o_i..., o_N\}$ and $B = \{b_1, ..., b_N\}$ and feature maps μ, extracted by object detector (here we used Faster R-CNN [17]). As an object feature extractor, Faster R-CNN [17] object classifier uses ROIAlign [19] to output visual features $M = \{m_i | i = 1, ..., N\}$ and their corresponding initial object labels $C = \{c_i | i = 1, ..., N\}$. The concatenation of three contextual features: m_i, b_i and c_i is passed into *Encoder* $\mathbf{f}$ to obtain object embeddings $\boldsymbol{v}_i$.

$$\mathbf{f}(o_i, c_i, b_i) = \boldsymbol{v}_i \tag{3.1}$$

$\boldsymbol{v}_i$ then fed into a *Decoder* $\mathbf{g}$ to predict fine-tuned object label $\boldsymbol{l}_i$.

$$\mathbf{g}(v_i) = \boldsymbol{l}_i \tag{3.2}$$

According to Equation. 3.1, we simply extract the pairwise object feature from each pair of objects (o_i, o_j) as $(v_i, v_j) | i \neq j; i, j = 1...n$.

3.2.2 Object Pairs Proposals

To handle the intractable number of possible pairwise object combinations, some works like [74] use a simple filter to remove many of the unnecessary object pairs. [172] cluster the phrase regions into some important ones and pass messages between them. Our proposed method is mostly related to the work by [102] in which both works propose a relation proposal network to estimate the relatedness of each object pair based on the predicted class probabilities. However, [102] does it without semantic embedding. Taking a different approach compared to their work, our first proposal *Align-RCNN* uses semantic embedding to choose the most semantically inter-dependent object pairs.

3.2.3 Align R-CNN Construction

Align R-CNN construction aims to demonstrate that feature extraction in pairwise relation models can be improved by augmenting a pairwise head to the attention component for dynamic object feature concatenation. To this end, head network aims to learn *object-part-alignment* node embedding $\boldsymbol{S}$, which approximates the task-dependent possibility between each object pair.
Predicting the relationship $\boldsymbol{r}_{i,j}$ between objects $\boldsymbol{o}_i$ and $\boldsymbol{o}_j$ typically takes three inputs: (1) object embeddings: $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, (2) word embeddings of the object predictions: $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, and (3) the visual features of the union region $\boldsymbol{x}_{ij}$. Given a pair of object features as $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, the naive pairwise feature is considered as a combination of representations from $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Unlike previous works that encode the pure visual feature representations of subject and object, Align R-CNN supports the assignment of varying importance values to different pair features, rather than allocating an equal weight for all pairwise feature embeddings. For example, in `person-horse` pair, different visual features from `person` and `horse` in different samples will help to distinguish whether the relationship is `person-riding-horse` or `person-feeding-horse`. Therefore, we propose an $Encoder_{\boldsymbol{r}}$ to adaptively capture contextual semantics by learning *object-part alignment* node embedding $\boldsymbol{S}$, which refines the pairwise visual feature for more discriminative relationship prediction. For simplicity, we denote $\boldsymbol{u}$ and $\boldsymbol{v}$

instead of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the rest of this chapter.

$$\boldsymbol{S} = \boldsymbol{f}(\boldsymbol{u}, \boldsymbol{v}) \tag{3.3}$$

Therefore, in Eq.3.3 we define $\boldsymbol{f}$ as an attention-based part-alignment module which discovers and re-aligns local pairwise features that contribute the most to final prediction. $\boldsymbol{f}$ takes node features $\boldsymbol{u}$ and $\boldsymbol{v}$ as input i. $\boldsymbol{S}$ reflects the importance of the edge ($\boldsymbol{u}$, $\boldsymbol{v}$), which can be used to measure the importance of any pairs task-dependency for any predicate category.

The proposed pairwise alignment head is branched into two steps: **Message** and **Pairwise Alignment**. During prediction, the Message module emulates searching through whole region unit in every $\boldsymbol{u}$, $\boldsymbol{v}$- pair as object positions to extract semantic-rich features [115]. Subsequently, it re-aligns the low-level appearance features via transferring semantic-rich information from `object1` $\Leftrightarrow$ `object2` as well as `object2` $\Leftrightarrow$ `object1` (see Fig. 3.4). Then, the Pairwise Alignment module generates pairwise representations for all object pairs due to the obtained contextual information. The output of Pairwise Alignment module is fed into *Decoder* to predict $\boldsymbol{r}_{i.j}$.

3.2.3.1 Message

The straightforward method to apply message passing for extracting alignments between paired objects is to use pairwise attention activation A [173]. Hence, $\boldsymbol{u}$ and $\boldsymbol{v}$ are then used as key and query to attend over graph node features to get attended features $\varphi(\boldsymbol{u})$ and $\varphi(\boldsymbol{v})$ (Eq. 3.4).

$$A \quad = softmax(ATT(\frac{W_h^{\boldsymbol{u}}.\boldsymbol{u} \cdot W_h^{\boldsymbol{v}}.\boldsymbol{v}^T}{\sqrt{\boldsymbol{d}}})) \tag{3.4}$$

where, A represents relative attention, which is computed by a softmax function over all the regions in the pairwise feature maps ($\boldsymbol{u}$, $\boldsymbol{v}$). $W_h^{\boldsymbol{u}}$ and $W_h^{\boldsymbol{v}}$ are transformation matrices which embed the corresponding features into a common subspace. $\boldsymbol{d}$ denotes the channel dimension of embedded features ($\boldsymbol{d} = \boldsymbol{d}_{model} = 256$).

To extract more discriminative features, we apply the K-Hard attention scheme ATT. This scheme represents the K most relevant visual information which needed to be considered as object part interaction and simultaneously processed to re-align the features, *i.e.*decode visual context. For example, to predict more precisely, the

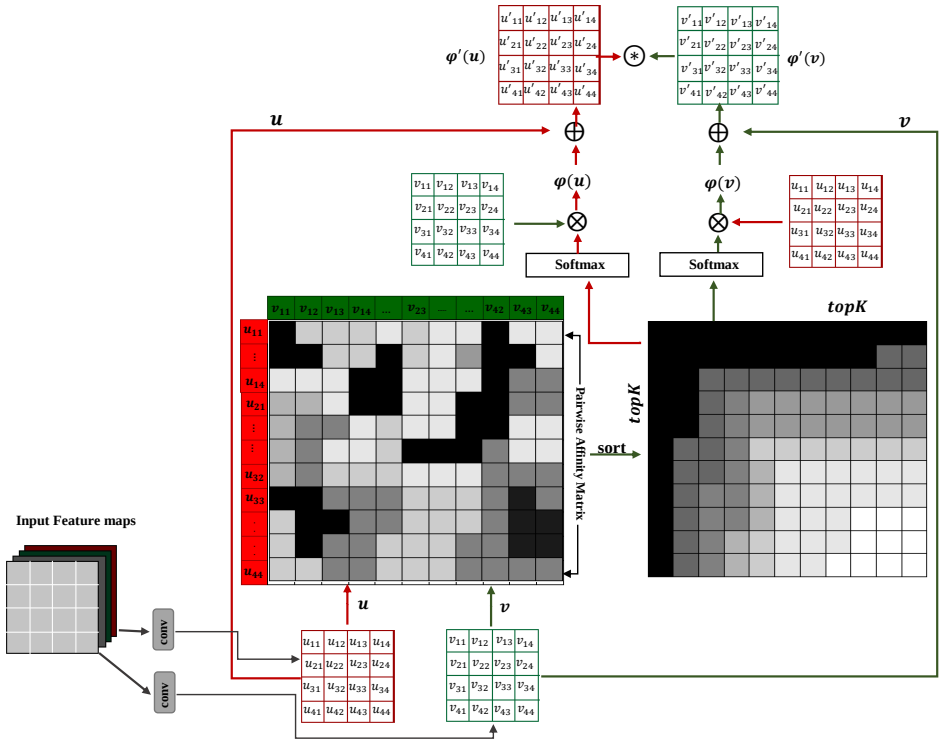

FIGURE 3.4: **Proposed attention-based multiple region feature alignment module.** $\boldsymbol{u} = \{u_{11}...u_{ij}...u_{IJ}\}$, $\boldsymbol{v} = \{v_{11}...v_{ji}...v_{IJ}\}$: input feature embedding representations are belong to object1 and object2 in i and j region positions $(i, j-1, ..., I \times J)$. The alignment module generates pairwise features as $\varphi^{'}(\boldsymbol{u}), \varphi^{'}(\boldsymbol{v})$ for final prediction.

`riding` category in pair (`person, motorcycle`), `person hand` and `motorcycle handle` as well as `person leg` and `motorcycle body` that needs to be selected as discriminative regions and transferred into corresponding regions:

$$ATT_i = topK(\boldsymbol{A}_i^P) \tag{3.5}$$

top_K returns the indices of the K largest values of an input vector, and $\boldsymbol{A}_i^P$ denotes the pairwise affinity matrix belongs to input i (K is set to 6 by default in our experiments). As shown in Figure 3.4, we select object1 feature map state representations $\boldsymbol{u}$ with the K maximal attention value for corresponding object2 ($\boldsymbol{v}$) and vice versa. We average attention activation matrices $\boldsymbol{A}$ into pairwise affinity matrix $\boldsymbol{A} \in \mathbb{R}^{I \times J}$ to extract a hard-attention between object unit regions. Thus, message module learns a weight of attention for each pair and then gather information from

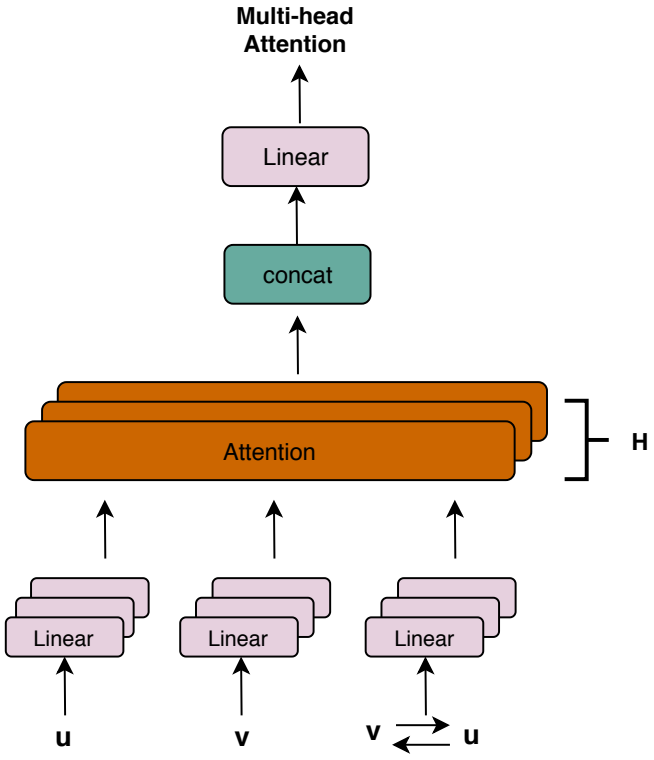

FIGURE 3.5: **Multi-Head Attention consists of several attention layers running in parallel.**

most relevant regions based on those weights as:

$$\begin{aligned} \varphi(\boldsymbol{u}) &= \sigma(\sum_{j \epsilon \mathbb{N}_i} \boldsymbol{A}.\boldsymbol{v}) \\ \varphi(\boldsymbol{v}) &= \sigma(\sum_{j \epsilon \mathbb{N}_i} \boldsymbol{u}.\boldsymbol{A}) \end{aligned} \tag{3.6}$$

To prevent information dilution and negative knowledge transferring, we implement multiple independent attentions or the attention calculations [119], [158]. According to the multi-head structure shown in Figure 3.5, inputs are entered in a linear transformation and then fed into H-scaled dot product attention heads (Eq.3.7). For each head h, $(h = \ 1, ..., H)$, the attention-based alignment function obtains different and independent representations of inputs. Finally, the H-th order expansion results are concatenated, and the outputs by the linear transformation are used as the multi-head attention weights.

$$Multi - head(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{v}) = concat(head_1, ..., head_h) \tag{3.7}$$

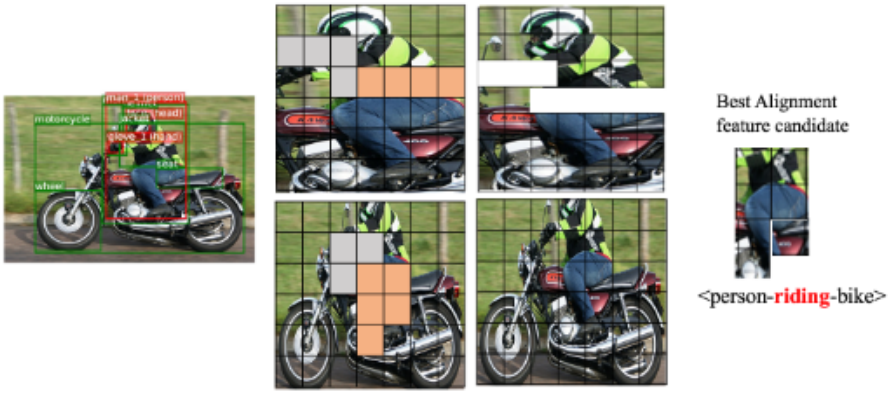

FIGURE 3.6: **Visualization of the proposed pairwise object features alignment procedure.**

In Eq. 3.7, we employ $H = 4$ as attention layers, or heads. For each layer, $\boldsymbol{d}_{model}/4 = 64$.

3.2.3.2 Pairwise Object Alignment

Finally, the pairwise visual context representation $(\varphi(\boldsymbol{u}),\ \varphi(\boldsymbol{v}))$ for each object pair $(o_i,\ o_j)$ are generated according to Eq. 3.9, separately:

$$\varphi(\boldsymbol{u}) = concat_{h=1}^{H}(\sigma(\sum_{j \epsilon \mathbb{N}_i} \boldsymbol{A}^h.\boldsymbol{v})) \tag{3.8}$$

$$\varphi(\boldsymbol{v}) = concat_{h=1}^{H}(\sigma(\sum_{j \epsilon \mathbb{N}_i} \boldsymbol{u}.\boldsymbol{A}^h)) \tag{3.9}$$

where "concat" represents a concatenate operation, σ is a nonlinear activation function, and $\boldsymbol{A}^h$ denotes the weight obtained by the h-th head of attention mechanism. In general, hard-attention can help to reduce the set of feature components that are needed to be processed.

In addition, the strong Visual Relationship Detection (VRD) task needs to consider the individual object feature. Hence, we use the benefits of information and visual context from both object categories to boost the predicate feature mutually. For this, we fuse each object feature in a pair with it's aligned features in channel dimensions using convolutional layers with a ReLU activation function to form the final feature maps. Hence, the selected pair $(\boldsymbol{u},\ \boldsymbol{v})$ will be assigned into a new pair

of $(\varphi^{'}(\boldsymbol{u}), \varphi^{'}(\boldsymbol{v}))$ from Eq. 3.11:

$$\varphi^{'}(\boldsymbol{u}) = LeakyReLU(\mathbf{MLP}[\boldsymbol{u},\ \varphi(\boldsymbol{u})]) \tag{3.10}$$

$$\varphi^{'}(\boldsymbol{v}) = LeakyReLU(\mathbf{MLP}[\boldsymbol{v},\ \varphi(\boldsymbol{v})]) \tag{3.11}$$

where [.] denotes the concatenate function. The visual context $\boldsymbol{S}$ (Eq. 3.3) is, then, computed as a weighted sum of aligned objects. We used $LeakyRelu$ as an activation function from Eq. 3.12 which is a variant of $ReLU$. where a is constant gradient ($a = 0.01$).

$$\begin{cases} z & z > 0 \\ a & z <= 0 \end{cases} \tag{3.12}$$

3.2.4 Scene Graph Generation

After completion of state re-alignment of all object pairs, the final predicate logits $\boldsymbol{p}_{i,j}$ is generated by fusing two inputs, one is the output of alignment module (generated pairwise feature embeddings as shown in Figure 3.6), and another is contextual union region features $\boldsymbol{x}_{ij}$ from ROIAlign [19] layer before we project the visual context feature into a feature space of $\mathbb{R}^{4096}$ using convolutional layer ($Convs$) (Eq. 3.14).

$$\boldsymbol{p}_{i,j} = (\varphi^{'}(\boldsymbol{u}) \oplus \varphi^{'}(\boldsymbol{v}) \oplus \boldsymbol{x}_{ij}) \tag{3.13}$$

$$\boldsymbol{x}_{ij} = Convs(RoIAlign(\mu,\ \boldsymbol{b}_1 \cup \boldsymbol{b}_2)) \tag{3.14}$$

$\boldsymbol{b}_1 \cup \boldsymbol{b}_2$ indicates the union box of two RoIs. In Eq. 3.13, $\oplus$ denotes fusion function proposed by [174] that we use here instead of naïve pairwise feature concatenation:

$$x \oplus y = ReLU(x + y) - (x - y)^2 \tag{3.15}$$

Then, $\boldsymbol{p}_{i,j}$ is fed into $Decoder_{\boldsymbol{r}}$ a fully connected layer followed by a softmax classifier layer to predict predicate label $\boldsymbol{r}_{i,j}$.

As mentioned before in Section 3.2.1, following MOTIFS [6] and VCTree [5], we utilize LSTM and TreeLSTM as *Decoders* to capture the co-occurrence among object labels. The input of each LSTM/ TreeLSTM cell is the concatenation of feature and the previous label: $[\boldsymbol{v}_i;\ \boldsymbol{l}_{i-1}]$. Besides, following [10], the language

prior is calculated as: $L = W_z[\boldsymbol{l}_i \otimes \boldsymbol{l}_j]$, where $\otimes$ generates the one-hot unique vector $\mathbb{R}^{N \times N}$ for the pair of N-way object labels.

After predicting relationship for all object pairs, we finally obtain the generated scene graph in the form of $G = V = (\boldsymbol{c}_i, \boldsymbol{b}_i)$, $E = \boldsymbol{r}_{i,j} | i, j = 1...n$, where V and E denote the set of nodes and edges, respectively. $\mathbf{c}_i \in C$ is the object class of ith node $(i = 1, ..., n)$, $\boldsymbol{b}_i \in R^4$ is the location of node $\boldsymbol{v}_i$. $\boldsymbol{r}_{i,j} \in R$ is the predicate label between ith and jth node $(i \neq j)$.

3.3 Experiments on Scene Graph Generation

We conduct several experiments to study whether the introduced method is able to effectively generate the discriminative pairwise features in the task of SGG. We perform a fine-grained as well as zero-shot analysis by replacing our proposed aligned pairwise features in predicate prediction with the final context features proposed by various state-of-the-arts models [5–7] and a classical model [9] (Section 3.3.5).

3.3.1 Benchmark

Visual Genome (VG) [67] is a widely used benchmark for SGG. It consists more than 10^5 images, each contains multi objects predicate relation categories, also it is annotated inconsistently [9, 70, 74, 75, 175], *i.e.*resulting in a wide range of classes with little variation and few examples. Therefore, some attempts have been done to clean these annotations [9, 74, 75, 91, 101, 175–178]. They explored to make clean data by applying predicate class merging and filtering, thus improve the overall performance by using their own VG benchmark. Moreover, there is no unique and standard split for train, validate, and test sets of VG dataset. Recently, two works of [179] and [180] proposed to dedicate a new split on a large-scale version of VG. However, [180] work has not been benchmarked yet. Like many other works, we have adopted a split following [9], which selects top-150 object categories and top-50 predicate categories by frequency [5]. VG is divided into the training set and test set by 70%, and 30%, respectively. We further used 5,000 images from training set as the validation set for hyper-parameter tuning. We train the detector on the VG dataset [67] using the SGD optimizer with a batch size of 6, momentum of 0.9,

	Model	Backbone	SGDet		SGCls		PredCls		mean
			R@50	R@100	R@50	R@100	R@50	R@100	
Constraint	IMP+ [9, 66]	VGG16	20.7	24.4	34.6	35.4	59.3	61.3	39.3
	FREQ [5, 6]	VGG16	26.2	30.1	32.3	32.9	60.6	62.2	40.7
	MOTIFS [6]	VGG16	27.2	30.3	35.8	36.5	65.2	67.1	43.6
	Kern [66]	VGG16	27.1	29.8	36.7	37.4	65.8	67.6	44.1
	VCTree [5]	VGG16	27.9	31.3	38.1	38.8	66.4	68.1	45.1
	LinkNet [122]	VGG16	27.4	30.1	41.0	41.7	67.0	68.5	45.9
	ARN [71]	VGG16	-	-	38.2	40.4	56.6	61.3	-
	ATR-Net [70]	ResNeXt-101	21.4	26.4	36.0	37.0	65.8	67.8	42.4
	IMP+ [9, 10]	ResNeXt-101	25.9	31.2	37.5	38.5	61.1	63.1	42.9
	Align-IMP+	ResNeXt-101	**26.1**	**32.0**	**38.0**	**39.0**	**61.2**	**64.1**	43.4
	VTransE [7, 10]	ResNeXt-101	26.9	30.9	38.2	39.1	57.3	61.7	42.4
	Align-VTransE	ResNeXt-101	**32.0**	**36.1**	**38.4**	**39.3**	**62.2**	**64.5**	45.4
	MOTIFS [6, 10]	ResNeXt-101	32.7	37.2	38.9	39.8	65.1	67.0	46.7
	Align-MOTIFS	ResNeXt-101	**33.1**	**37.5**	**39.6**	**40.3**	**66.0**	**68.1**	47.4
	VCTree [5]	ResNeXt-101	31.5	36.2	**46.6**	**47.6**	65.4	67.2	49.0
	Align-VCTree	ResNeXt-101	**32.1**	**36.4**	45.5	46.4	**66.3**	**68.7**	49.2
No Constraint	IMP+ [9, 66]	VGG16	22.0	27.4	43.4	47.2	75.2	83.6	49.8
	FREQ [6]	VGG16	25.3	30.9	40.5	43.7	71.3	81.2	48.8
	KERN [66]	VGG16	30.9	35.8	45.9	49.0	81.9	88.9	55.4
	ARN [71]	VGG16	-	-	41.4	46.0	61.6	68.9	-
	IMP+ [9, 10]	ResNeXt-101	27.0	33.9	46.8	51.2	76.8	85.0	53.5
	Align-IMP+	ResNeXt-101	26.8	34.0	**47.6**	**51.6**	**77.8**	**85.8**	53.9
	VTransE [7]	ResNeXt-101	29.6	35.7	47.8	51.5	68.4	78.7	52.0
	Align-VTransE	ResNeXt-101	**36.0**	**42.5**	47.8	51.5	**78.1**	**86.2**	57.0
	MOTIFS [6, 10]	ResNeXt-101	36.6	43.4	48.5	51.9	81.0	88.2	58.3
	Align-MOTIFS	ResNeXt-101	**34.3**	**40.7**	**49.4**	**53.0**	**81.9**	**89.0**	63.8
	VCTree [5, 10]	ResNeXt-101	35.7	42.3	**58.3**	**62.7**	67.2	81.6	57.9
	Align-VCTree	ResNeXt-101	**36.1**	**42.6**	56.7	61.0	**80.4**	**87.8**	60.8

TABLE 3.1: The SGG performances (%) of various models on Recall@K. Four SGG models [5–7, 9] with ResNetx-101 backbone were re-implemented under codebase proposed by [10] for fair comparison. First seven rows show the baseline models that originally were implemented using VGG-16 backbone.

Model	Backbone	SGDet		SGCls		PredCls		mean
		mR@50	mR@100	mR@50	mR@100	mR@50	mR@100	
IMP+ [9, 66]	VGG16	3.8	4.8	5.8	6.0	9.8	10.5	6.8
FREQ [5, 6]	VGG16	6.1	7.1	7.2	8.5	13.0	16.0	9.6
Kern [66]	VGG16	6.4	7.3	9.4	10.0	17.7	19.2	11.7
MOTIFS [6]	VGG16	5.7	6.6	7.7	8.2	14.0	15.3	9.6
VCTree [5]	VGG16	6.9	8.0	10.1	10.8	17.9	19.4	12.2
IMP+ [9, 10]	ResNeXt-101	4.2	5.3	6.2	6.5	10.9	11.8	7.5
Align-IMP+	ResNeXt-101	**4.6**	**5.9**	**8.7**	**9.3**	**13.4**	**14.6**	9.5
VTransE [7, 10]	ResNeXt-101	5.0	6.0	8.5	8.7	12.2	13.9	9.2
Align-VTransE	ResNeXt-101	**6.7**	**7.8**	**9.1**	**9.5**	**13.9**	**15.0**	10.4
MOTIFS [6, 10]	ResNeXt-101	5.5	6.8	8.0	8.5	14.6	15.8	9.9
Align-MOTIFS	ResNeXt-101	**7.5**	**8.8**	**9.7**	**10.2**	**16.6**	**18.0**	11.8
VCTree [5]	ResNeXt-101	5.7	6.9	7.5	7.9	14.9	16.1	9.8
Align-VCTree	ResNeXt-101	**6.7**	**7.8**	**11.1**	**11.7**	**16.9**	**18.0**	12.0

TABLE 3.2: **The SGG performances (%) of various models on mean-Recall@K**. We use the same notations as in Tab. 3.1

and weight decay of 0.0001. The learning rate is initialized as 0.001 and is divided by 10 when the mAP of the validation set plateaus. We use the training/test split in [6] for evaluation.

3.3.2 Protocols

Generally, five evaluation tasks have been introduced in Scene Graph Generation so far, but prior works mostly evaluate on three tasks. We preserve the tasks' names as defined in [95] and [9], despite inconsistencies on whether they are more related to classification or detection tasks:

- **Predicate Detection (PredDet)** [95]: Given an image associated with the ground truth bounding boxes and labels of objects, as well as which object pairs do interact, it outputs each pair's predicate class.
- **Predicate Classification (PredCls)** [9]: Given an image associated with the ground truth bounding boxes and labels of objects, it decides which object pairs interact and outputs each selected pair's predicate class.
- **Scene Graph Classification (SGCls)** [9]: Given an image associated with the ground truth bounding boxes, it classifies objects, decides which pairs interact and further classifies each pair's predicate.

- **Scene Graph Generation (SGDet)** [9], also known as Relationship Detection (RelDet) [95]: Given a raw image with no prior information, it detects objects, decides which pairs interact, and further classifies each pair's predicate.

- **Phrase Detection (PhrDet)** [95]: Similar to SGDet however it evaluates the IoU of the predicate bounding box per pair. Like most prior works (*e.g.* [6, 9]), we evaluated *AttAlign* on all the three above standard tasks per dataset: PredCls, SGCls, and SGDet.

3.3.3 Metrics

From [95], a standard metric for evaluating SGG is conventional Recall@K (R@K) where K gets three values: 20,50,100. R@K measures the number of times that correct predicate comes from the top-k confident predictions. Early works like [95, 175, 178, 181], considered Visual Relationship Annotation as a multi-class classification problem and they give k as 1 to reward the correct top-1 prediction for each pair of objects. However, more recent works [74, 182], tackled VRD by knowing that each pair can include more than one relationship category (multi-label problem). Thus, they asserted k equal to the number of relationship classes to allow for predicate co-occurrences [70]. Due to inconsistency in annotation, past works like [6, 70, 179] have made a modification into R@K, by considering k as the maximum number of predicates validated between a pair of objects. Thus, in this thesis similar to the some validate works, *graph constraint* denotes the case in which $k = 1$, while *no graph constraint* is for larger k ($k > 1$).

Besides, there are other perspective in Recall metric: Micro-Recall or Macro-Recall that mostly used by Pioneer works on VRD [95] and on VG200 [9], respectively.

$$R^{Micro} = \frac{\sum_{i=1}^{N} tp_i}{\sum_{i=1}^{N} gt_i} \tag{3.16}$$

Given the number of testing images and the number of ground-truth predicates as N and gt per image ith. Micro-Recall counts true positives tp per image i, as Eq.3.16. Similarly, Macro-Recall rewards the detected annotation in terms of

images:

$$R^{Macro} = \frac{1}{N}\left(\frac{\sum_{i=1}^{N} tp_i}{\sum_{i=1}^{N} gt_i}\right) \tag{3.17}$$

In this work, we picked Recall@50 and Recall@100 following [95] to evaluate our model performance.

$$R@X_k = \frac{1}{Y_k}\sum_{k=2}^{Y}\frac{T_{yk}^{X}}{g_{yk}} \tag{3.18}$$

Furthermore, due to the fact that the SGG models trained on imbalanced datasets such as VG, it generate lower performances for less frequent predicate categories. To this end, we evaluate architectural robustness of the proposed Align R-CNN using unbiased/comprehensive metric called **Mean Recall (mR@K)** which was introduced by [5, 66]. In Eq.3.19, Mean Recall calculates the recall from Eq.3.18 on each category k, and subsequently averages the values. Thus, all categories are measured equally [5]. mR@K reduces the negative effect of some meaningless yet frequent relationships, *e.g.* "`on`", and more importantly leads attention to infrequent but reasonable categories, *e.g.* "`standing`" and "`carrying`", which are semantics for high-level reasoning.

$$mR@X = \frac{1}{K-1}\sum_{k=2}^{K} R@X_k \tag{3.19}$$

3.3.4 Implementation Setting

Scene Graph Generation- Each triplet composition is annotated in the form of t = <`subject-relation-object`> inside a scene graph, which is generated correctly if the classes of entities and relations between entities match those of ground truth. For **object detection**, following the previous works [6, 9, 10], Faster R-CNN [17] is pre-trained and frozen as a detector and following [10] it used a ResNeXt-101-FPN [16, 183] backbone. For more detail, the sizes of detected objects that resized to the network size (called as anchor box size) and aspect ratio are adjusted similar to [184], and the ROI pooling layer is replaced with the RoIAlign layer [19]. The detector was trained on the VG dataset [67] using the SGD optimizer with a batch size of 8 and the learning rate is initialized as 0.008 and is divided by 10 when the mAP of the validation set plateaus [10].

For **relationship predication**, three branches: union-boxes, object class labels, and pairwise features are passed into the final predictor. Unlike [5, 6], we pooled RoIAlign [19] object pair features to coarse 4×4 units with 256 channels, and passed object pairs into the alignment module (details in Sections 3.2 and 3.2.4). Our models are trained by SGD optimizer, where the batch size and learning rate are initialized to be 16 and 0.16, 12 and 0.12, 8 and 0.08 for **SGCls**, **PredCls** and **SGDet**, respectively. For **SGDet**, top-64 object proposals were selected after non-maximal suppression (NMS) with 0.3 IoU. We set background/foreground ratio for predicate classification to 3, and capped the number of training samples at 64 (retained all foreground pairs if possible). Our model is optimized by SGD with momentum, using learning rate $lr = 6 \cdot 10^{-3}$ and batch size $b = 5$ for supervised learning, and $lr = 6 \cdot 10^{-4}, b = 1$ for reinforcement learning. Besides, all SGG models are trained using the conventional cross-entropy losses for predicates and objects predicted by the object context layer [6].

3.3.5 Comparison with state-of-the-arts

3.3.5.1 Comparing Methods

The novelty of our proposed method in active alignment for relationship predication is applicable to a variety of SGG methods. As previously mentioned, we follow a very recent codebase proposed by Tang *et al.* [10], to implement our network on four baselines: two classic IMP+ [9] and VTransE [7] and two state-of-the-art models: Stacked Motif Networks (MOTIFS) [6] and Visual Context Tree model (VC_{Tree}) [5], which were called Align-IMP+, Align-VTransE, Align-MOTIFS, and Align-VC_{Tree}, respectively. The goal is to evaluate Align R-CNN with respect to its predicate feature generation across four different model formulations with their original representations. Furthermore, we compare our model with some other well-known works such as VRD [95] Message-Passing [9], FREQuency baseline (FREQ) [6], Knowledge-Embedded Routing Network (KERN) [66], and two attention-based models: Attention Translation Relation Network (ATR-Net) [70] and Attentive Relational Network (ARN) [71].

In summary, most state-of-the-arts to SGG [5, 6, 66, 185] use three types of features to represent relationships:

1) Visual Features: The CNN features of the two objects or their combination.
2) Spatial Features: Coordinates of the two objects which encodes their spatial layouts.
3) Semantic Features: Class labels of the two objects which provide a strong prior of the predicate.
Most of those methods, if not all, combine the three features in an early stage to learn a compositional feature for relationship prediction. The contribution of each feature is thus implicit and probably not optimized. Furthermore, according to message passing, [5, 6, 9, 66] use joint context coding. In contrast, [95] and FREQUENCY [6] use independence context coding. We re-implement them by replacing the proposed aligned pairwise feature embeddings with the context features proposed by each mentioned models and using the same hyper-parameters and the pre-trained detector backbone, for fair comparison. The quantitative results are provided in Tables 3.1- 3.3, in which our results in two constraint (limitation with number of relation per object pair) and no-constraint (no limitation with number of predicted relationships) modes are compared with the mentioned state-of-the-arts.

3.3.5.2 Quantitative Studies

The quantitative results are provided in Tables. 3.1, 3.2, and 3.3. Our method outperforms the state-of-the-art models significantly on mean-Recall and zero-shot setting. According to Table. 3.1, in which performances are evaluated with constraint (just consider one relationship per pair) and without constraint (no limitations) [186] conditions, it is observed that despite the simple architecture of the proposed method, it can achieve comparable performance under the constraint mode. For two tasks of **SGDet** and **SGCls**, proposed ALIGN-MOTIFS and ALIGN-VCTREE serve as state-of-the-arts.

Fine-grained Relationship Prediction Recall metric is dominated by the performance of most frequent predicates. As the discriminative model cannot be effectively evaluated by Recall, mean Recall is a more important metric than overall Recall especially when the class distribution of the database is severely skewed. In Tab. 3.2, the model performance of fine-grained relationship predictions evaluated using unbiased metric (mean-Recall [5, 66]) and compared with four baselines. Surprisingly, for all tasks: **SGDet**, **SGCls**, and **PredCls**, ALIGN R-CNN models achieve significant improvements compared with model baselines.

To give a comprehensive analysis of this circumstance, Figure 3.7 focuses on the comparison of the proposed models versus. model baselines, visually for case-by-case analyses. According to Figure 3.7, our model can see fair improvements on some fine-grained relation classification. For example, ALIGN R-CNN is able to detect less frequent but more informative predicates like `covered in`, `walking on`, `against`, *etc.*in comparison with the top 35 most frequent samples in the training set. However, VC_{Tree} and MOTIFS baselines are obviously inclined to most frequent predicates, like `on` and `behind` that show why their recall on **SGCls** and **PredClas** are comparable to ALIGN R-CNN. Specifically, in the model by MOTIFS, 54% of predicates such as `parked on`, `walking on`, and `sitting on` are misclassified to "frequent but the less informative" predicate such as `on`, while our model only fails on 37% of that.

Zero-Shot Recall. In addition, we prove the strength of reasoning through more discriminative visual features in zero-shot learning. Here we address the problem of zero-shot learning by generalizing discriminative pairwise visual feature that prevents penalizing components containing infrequent or zero-shot visual predicates. According to Tab. 3.3, our performance results outperforms over all baselines significantly as they demonstrate that generalizing discriminative pairwise visual features by object part alignment provides a stark improvement over naïve and even other context models.

In particular, comparing with two state-of-the-art baselines: MOTIFS and VCTree, there are over 57% and 22% performance improvements on **SGCls**zR@50 and **PredCls**zR@50 respectively. In real world images, some triplet compositions (*e.g.* `person-riding-motorcycle`) occur more frequently than others (*e.g.* `dog-riding- motorcycle` or `person- beside- motorcycle`), which creates a strong frequency bias. This co-occurrence effect was analysed statistically for the first time by [6] (see FREQ in Table. 3.1 and Tab. 3.2). We study the effect of Frequency bias (FREQ), proposed in [6], on zR performance by adding and removing it. As such the FREQ was considered beyond the default model, our results in Table. 3.3 highlight that FREQ drops predicate performances in zero-shot. For example, by removing FREQ from our ALIGN-MOTIFS (no FREQ) and ALIGN-VC_{Tree}, **PredCls**zR@50 is improved from 3.9 to 18.6 and from 4.1 to 17.5, respectively.

	SGCls		PredCls	
Model	zR@50	zR@100	zR@50	zR@100
IMP+	3.3	3.9	17.6	20.3
Align-IMP+	**4.9**	**5.8**	**18.4**	**21.2**
VTransE	4.9	5.7	11.5	14.8
Align-VTransE	**5.1**	**6.1**	**18.8**	**22.0**
MOTIFS	0.7	1.1	3.2	5.3
Align-MOTIFS	**1.1**	**1.6**	**3.9**	**6.2**
MOTIFS(no FREQ)	2.2	3.1	11.0	14.7
Align-MOTIFS(no FREQ)	**5.3**	**6.2**	**18.6**	**21.7**
VCTree	1.2	2.1	3.2	5.5
Align-VCTree	**2.3**	**3.0**	**4.1**	**5.7**
Align-VCTree(no FREQ)	6.3	7.5	17.5	20.8

TABLE 3.3: **Quantitative results of Zero-Shot Recall**.

3.4 Qualitative Analysis

To better understand the impact of the ALIGN R-CNN approach, we compare the effectiveness of our work (three samples shown in top) to predict more accurate and discriminative relationships with the MOTIFS baseline [6] (three samples shown in bottom) via Figure 3.7.

MOTIFS achieves remarkable improvement when equipped with the proposed ALIGN R-CNN network: (1) ALIGN-MOTIFS encourages the model to predict more accurate and informative relationships compared with baseline. In the first row, we show the ability of ALIGN-MOTIFS to generate discriminative pairwise features and as a result detect accurate (*e.g.* `wearing`), informative (*e.g.* `looking at`, `eating`, `belong to`) relationships, while MOTIFS is more likely to fail on some infrequent predicates (`to` instead of `looking at`). Those infrequent predicates, normally carry more semantic information. Thus, our method tends to predict informative predicates rather than trivial ones, which is more capable to understand complicated scenes. Interestingly, learning *object-part alignment* caused pairwise features to be class-specific (rather than object-specific), that predict specific predicate which matches with aligned pairwise features. However, ALIGN-MOTIFS fails to detect some non-trivial relationships like `on`. In Figure. 3.9 we visualize several **SGCls** samples that were generated by the proposed ALIGN-MOTIFS model. First section, which is entitled by fine-grained, demonstrates the model's ability to precisely predicts informative predicates such as `riding` and `sitting on` than frequent but less informative like `on` by generating discriminative pairwise

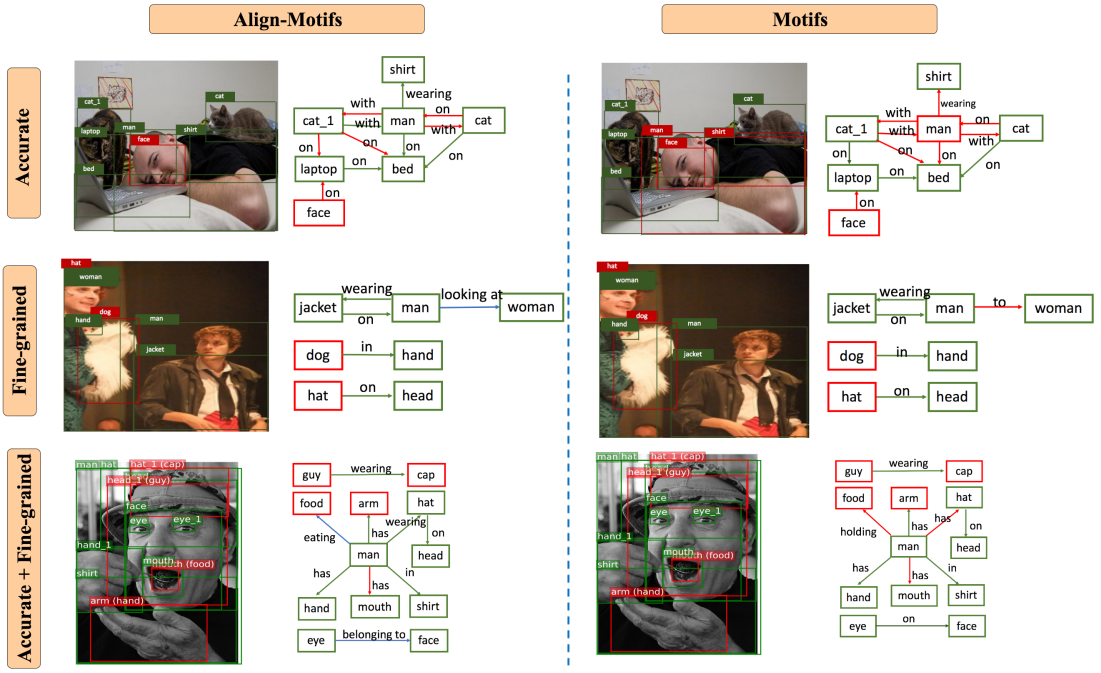

FIGURE 3.7: **Qualitative results showing comparisons between MOTIFS baseline and Align-MOTIFS in SGCls**. The denotations of the bounding box/node colors are as follows. Green: detected boxes with the ground-truth, red: ground-truth with no match. Green, red, and blue edges are labeled with true, false, and informative predicted by each model at the R@20 setting, respectively.

features. Interestingly, there is no much similar pairwise features for one single pair interacting (*e.g.* `woman-dog`) in multiple ways (*e.g.* `holding` and `looking at`) in our model.

Second section (entitled zero-shot) shows the power of ALIGN R-CNN that has been effected to detect unseen components (*e.g.* `cat-looking at-animal`) by composing and grounding existing visual concepts from other patterns. In summary, results point that there is some similar pairwise features for multiple pairs interacting in the same way. ALIGN R-CNN learns features by alignment to enable the predictor to:

1) distinguish the generated features for single pair interaction in multiple ways (*e.g.* different pairwise alignment in `man-surfboard` pair leads the model to predict multiple relationships (*e.g.* `riding`, `holding`, *etc.*)).

2) predict visually and semantically same relationship for multiple pairs (*e.g.* pairwise alignment for `looking at` predicate in `woman-phone` pairs can lead model to further predict `woman-dog` or `cat-looking at-animal` (zero-shot) in test set).

FIGURE 3.8: **Visualization of aligned pairwise feature embeddings in Align-MOTIFS for Fine-grained prediction**. For each section, the first column shows the grounding object regions (Green boxes point detected boxes with the ground-truth, red boxes point ground-truth with no match). Second and Third columns point the pairwise alignment estimation which explore and transfer informative regions that are inferred via Top-K pairwise regions' Message Passing. Last column highlight corresponding relationship that is predicted for sample selected pair.

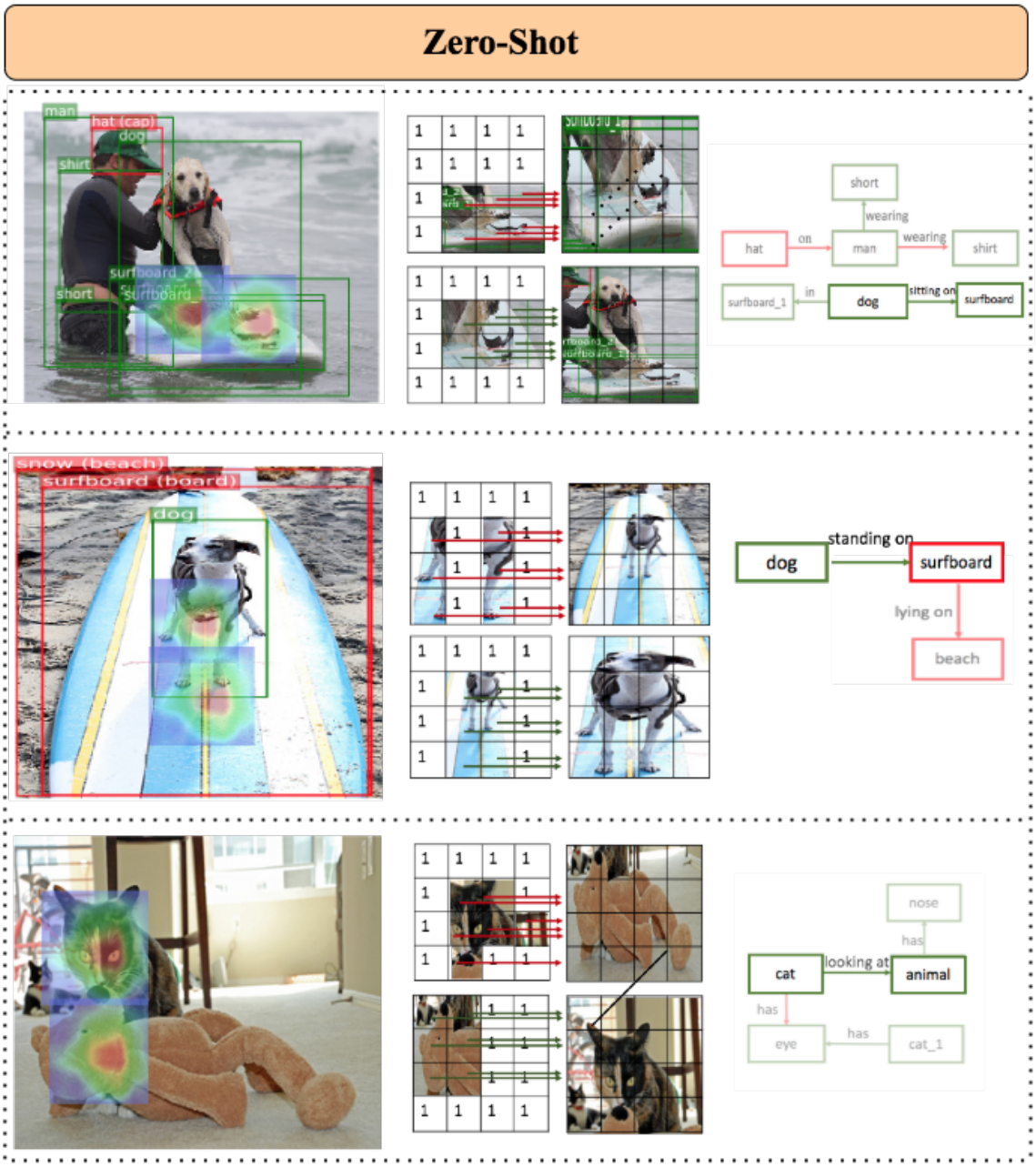

FIGURE 3.9: **Visualization of aligned pairwise feature embeddings in Align-MOTIFS for Zero-shot prediction**. For each section, the first column shows the grounding object regions (Green boxes point detected boxes with the ground-truth, red boxes point ground-truth with no match). Second and Third columns point the pairwise alignment estimation which explore and transfer informative regions that are inferred via Top-K pairwise regions' Message Passing. Last column highlight corresponding relationship that is predicted for sample selected pair.

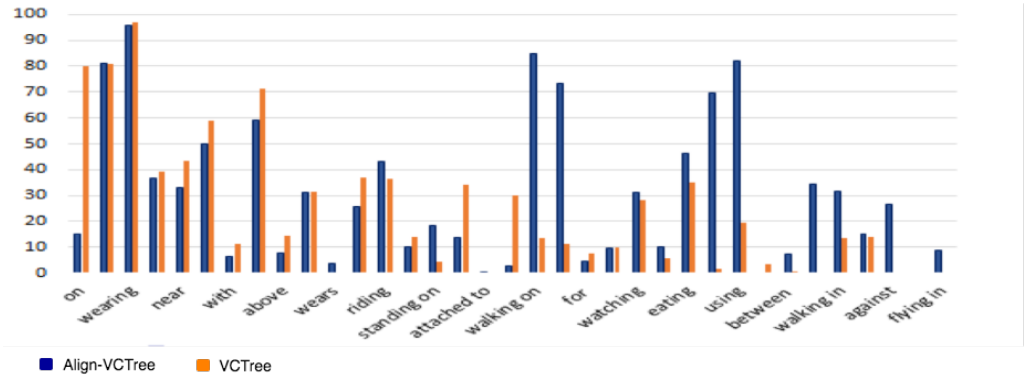

FIGURE 3.10: **Comparison of R@100 on PredCls task for the most frequent 35 predicates** between ALIGN-VCTREE and VCTREE [5]

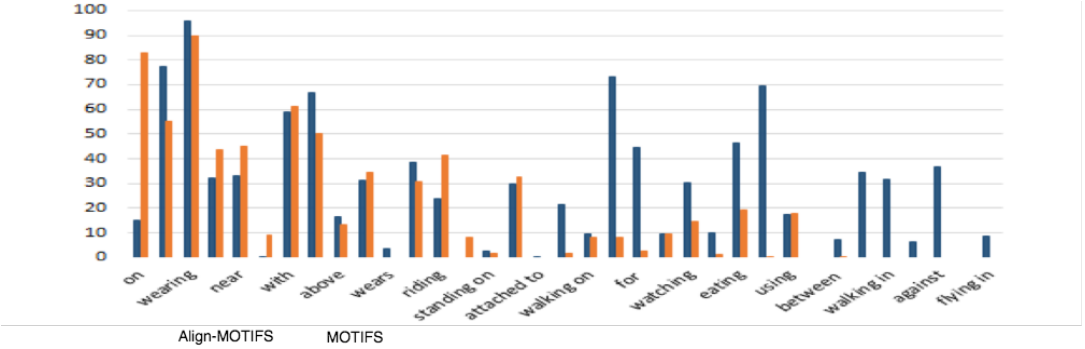

FIGURE 3.11: **Comparison of R@100 on PredCls task for the most frequent 35 predicates** between ALIGN-MOTIFS and MOTIFS [6]

3.5 Ablation Studies

To give a comprehensive analysis of this circumstance, we further focus on the comparison of the proposed models versus model baselines, visually for case-by-case analyses. According to Figures 3.10, 3.11, and 3.12, our pairwise head network can see fair improvements on some fine-grained relation classification. For example, ALIGN-VC$_{Tree}$ and ALIGN-MOTIFS are able to detect less frequent but more informative predicates like `covered in`, `walking on`, `against`, *etc.* in comparison with the top 35 most frequent samples in the training set. However, VC$_{Tree}$ and MOTIFS baselines are obviously inclined to most frequent predicates, *e.g.* `on` and `behind` that show why their recall on **SGCls** and **PredClas** are comparable to ALIGN R-CNN.

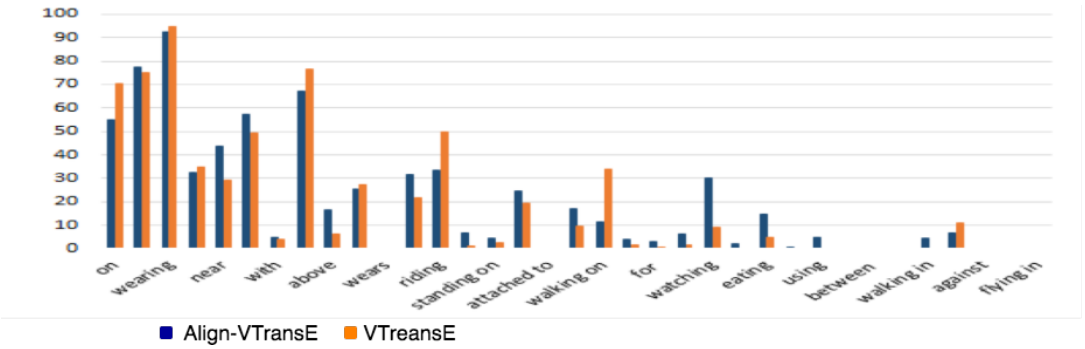

FIGURE 3.12: **Comparison of R@100 on PredCls task for the most frequent 35 predicates** between ALIGN-VTREANSE and VTRANSE [7]

3.6 Conclusion

In this chapter, we demonstrated the importance of learning feature alignment for dynamic object feature concatenation in relation prediction, where objects may not only have large overlaps in their spatial locations but also have a lot of commonalities in their semantic concepts. To this end, we introduced the ALIGN R-CNN framework with an object context encoding alignment module to explicitly learn complementary object alignment for predicting each object pair relation. After generating the complementary aligned features, the generated pairwise features and the region features are effectively combined as the final relation feature. ALIGN R-CNN transfers information between the pairwise nodes using hard attention in a differentiable manner to make the relationship prediction discriminative and robust against imbalanced dataset.

Chapter 4

Multimodal Contrastive Learning for Robust VQA

4.1 Introduction

Visual Question Answering (VQA) [12, 123, 143] is an active problem in multi-modal learning tasks that refers to predict and answer a to language question q about given image content v. Towards multi evaluation metrics for VQA, different research efforts have been made to build more accurate VQA model [124–127, 129, 187]. For instance, to improve to the VQA accuracy, model supposed to be discriminative (*e.g.* generate more correct answers as ground truth). However, as we mentioned in Chapter 2.3.1, despite significant progress in VQA, balancing between accuracy as well as robustness of VQA models much needed to explore. Thus, to improve the VQA overall performance and robustness *i.e.*against input variations, model supposed to be disciminative as well as consistence. Generally, the proposed solutions fall into the following categories: data augmentation [11, 188], robust learning process [32, 189], and classifier [190]. Among them, [11] tries to alleviate this issue by data augmentation (Chapter 4.2). Due to learning process, some robust training paradigms like *Adversarial Training* have been a proposed to better feature representation for similar samples in train set.

Inspired with recent progress in vision and language communities in robustness under adversarial training, and contrastive learning, we improve robustness-aware

supervised training by learning representations that are consistent under question input variations.

To summarize, our major contributions are threefold:

- We address the VQA robustness problem via dual/variant contrastive learning, which sufficiently learn better representation for both vision and language inputs.
- We propose variant true positive and negative samplings to optimize contrastive loss.
- We apply two separate generator modules for visual and texture modalities $(v, q, a) \rightarrow (v^+, q, a)$ and $v, q, a \rightarrow (v, q^+, a)$, for any given `Image-Question` and `Question- Answer` pair, respectively.

In the rest of this chapter, we dedicate a recap on the straight-forward model, mentioned solutions for model robustness and then proceed to our model in Section. 4.4.3.

4.2 Data Augmentation

Data augmentation is generating new training data from existing data samples. Both area of computer vision and natural language processing have made enormous progress on distribution problem by generating positive samples for data. In VQA, a straightforward strategy to improve model's robustness is to augment the training data with positives, which are examples that the prediction model does not classify correctly with high confidence, but that are visually similar to easy positives. To make VQA systems robust, few existing approaches [8, 11, 148] have trained variety of existing VQA models like [125, 129, 130, 187] by augmenting the training data with variations of the input question. Among them, there is no augmentation in visual modality except one recent work [148]. In following, we provide more details about language and visual augmentations.

4.2.1 Question Paraphrases

In the area of natural language processing (NLP), there are wide works for creating paraphrases of a sentence like using Long-short term memory (LSTM) [191], Deep Reinforcement Learning [192], Variational Autoencoders [193], Neural Machine Translation (NMT) [194–196], and textual DNN attackers [197, 198].

Question Paraphrases [11, 148] studied the VQA robustness in relation to question paraphrases and propose two augmented datasets by generating various rephrasings of questions. Shah *et al.* [11] use a visual question generation (VQG) model to generate paraphrased question given an image and ground-truth answer. Then they proposed cyclic-consistent training scheme to enable model predicting same answer for original and its rephrasing questions.

Similar to the method in [148], that used Neural Machine Translation [196] Back-Translation (BT) model to generate paraphrases for visual questions without any supervision. In this work, we use both augmented positive samples; **VQG** and **BT**, to evaluated our proposed approach.

4.2.2 Visual Augmentation using perturbed samples

Augmentation is widely used in the computer vision area especially in the task of image classification [25] by using mirror reflection, random crops, rotations, *etc.* However, as mentioned earlier, due to the semantic structure of VQA triplets, visual augmentation may not be helpful to make consistency of the original triplet [148]. It was proposed in [148] to use generated adversarial samples as the augmented data – for both the image and the question.

In an adversarial attack, given a clean input sample x, the adversary tries to add human unnoticeable perturbations δ in order to fool a DNN-based model:

$$x_{adv} = x + \delta \tag{4.1}$$

where δ is the adversarial perturbation. This kind of attack can leverage data augmentation techniques to process the input samples and make them resistant to adversarial attacks. In this thesis, we re-use question paraphrases as augmented

FIGURE 4.1: **Three examples of Paraphrase Generation in NLP**

data and further exploit the advantages of adversarial perturbation for generating hard samples for training structure.

4.3 Robust Training Strategy

A small number of methods have been investigated for adversarial training on the VQA task. However, they merely attack the image and do not discuss how the

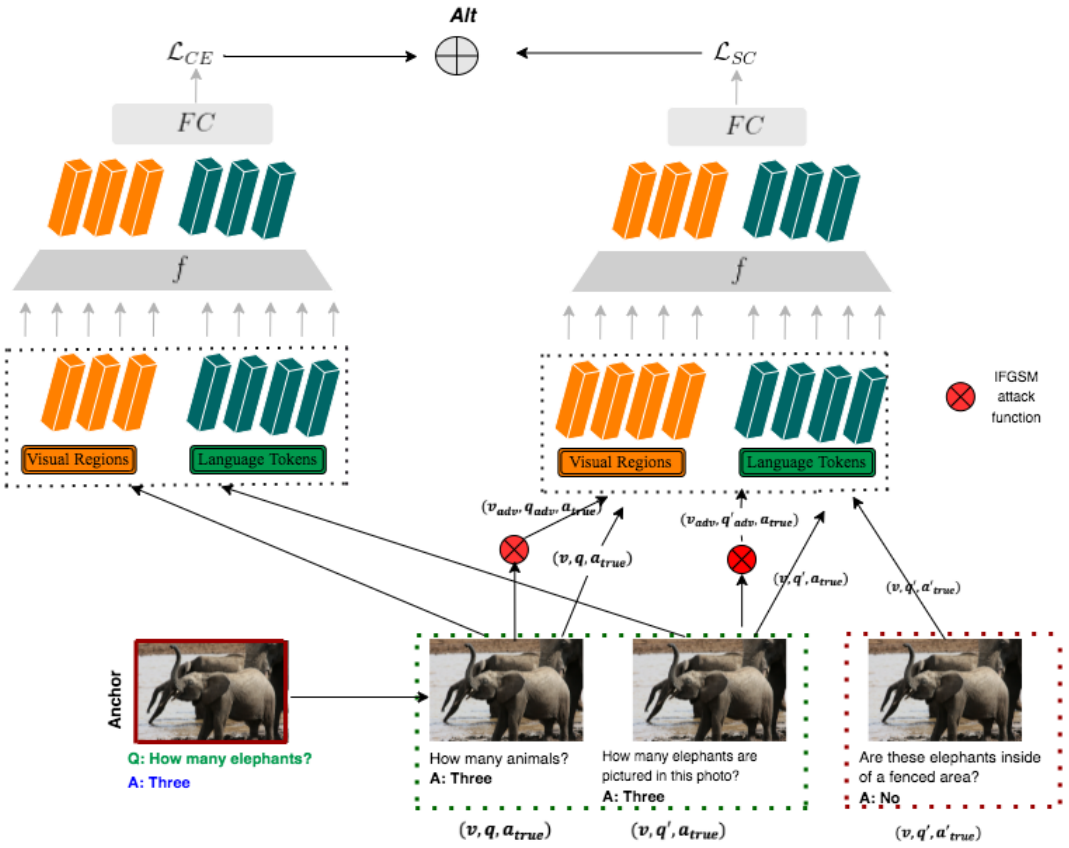

FIGURE 4.2: **Overview of proposed VQA framework**

adversarial examples can benefit the VQA model.

Adversarial Training (AT) that has been developed by [63, 199] is one of the most effective paradigms for defending DNN models against adversarial examples. AT involves generating constrained loss that maximizes adversarial examples from each batch of training data via Eq. 4.2, computing the expected adversarial loss $\mathcal{L}$ over all perturbed samples x_{adv} in the batch and subsequently training the model to minimize $\mathcal{L}$.

$$x_{adv} = x + \delta^* \quad where \delta^* = argmax \mathcal{L}(x + \delta) \tag{4.2}$$

TRADES proposed by [200] maximized the KL divergence loss ($\mathcal{L}_{KL}$) between clean and adversarial logits, as opposed to the adversarial cross-entropy loss, during PGD (inner maximization). Their training loss is the following joint formulation of clean and robust loss:

$$\begin{aligned} \mathcal{L}_{TRADES} = \mathcal{L}_{CE}((v_{clean};\ q);\ a;\ \theta) + \\ \beta \mathcal{L}_{KL}((v_{clean};\ q);\ (v_{adv};\ q);\ \theta) \end{aligned} \tag{4.3}$$

4.3.1 Attack methods

In this part, we review different types of adversarial attacks introduced by [63, 199, 201–206]. In general, attacks refer to finding adversarial examples for well-trained models [207] and are divided into two types of "black-box" and "white-box". The difference comes from having all the information about the targeted neural network or not. Among proposed white-box attack methods, CW [201] and PGD [63] attacks have achieved the robustness goal of machine learning models. Besides, some attack methods rely on the predicted scores (*e.g.* logits) of the model to deal with adversarial examples. While, some are more agnostic and only rely on the final decision of the model.

In this thesis we consider adversarial attack to generate unbiased samples' representations. Due to Eq. 4.5, adversary tries to generate a perturbation δ to maximize the loss function like Eq. 4.4, so that

$$\begin{aligned} \mathcal{L}_{TRADES} = \mathcal{L}_{CE}((v_{clean};\ q);\ a;\ \theta) + \\ \beta\mathcal{L}_{KL}((v_{clean};\ q);(v_{adv};\ q);\ \theta) \end{aligned} \tag{4.4}$$

Taking prediction as an example, $f(v_{clean};\ q;\theta)$ that predicts label a, which θ indicates the parameters of f. Given the perturbation budget δ, the adversary tries to find a perturbation δ^* to maximize the loss function $\mathcal{L}$, so that $f(x+\delta) \neq f(x)$. Therefore, δ^* is estimated as follows:

$$\delta^* := argmax\mathcal{L}(\theta;\ x+\delta;\ y) \tag{4.5}$$

where y is the label which in thesis is denoted by answer a of input sample x which refers to input image v and question q:

$$\begin{aligned} \delta_v^* &:= argmax\mathcal{L}(\theta;\ v+\delta;\ q;\ a) \\ \delta_q^* &:= argmax\mathcal{L}(\theta;\ v;\ q+\delta;\ a) \end{aligned} \tag{4.6}$$

In some cases, δ is small so that the perturbations are invisible.

The common but effective attacks following this formulation, such as Fast Gradient Sign Method (FGSM) [199], iterative FGSM (IFGSM) [202], and Projected

Gradient Descent (PGD) attack [63]. In this thesis, we have apply both PGD and IFGM attacks.

4.4 Approach

In this section we describe our proposed AdvCL model, which introduces an extended version of contrastive learning method to deal with VQA model robustness across question language variations. First, it re-uses the paraphrases of questions to augment the training set. Given a dataset $D = \left\{(v_i, q_i, a_i)\right\}_{i=1}^{N}$ contains N triplets of image v_i, question q_i and ground-truth answer a_i. Following ([11, 208]), we augment the dataset D with paraphrased samples question $Q^{Para}(q)$ where $q \in D$.

As illustrated in 4.2, the architecture of the proposed VQA model consists of two networks: Contrastive and cross-entropy training to enhance model robustness while keeping discriminative predictions. Given <`image, question, answer`> triplet, we first generate the hard samples by adversarial perturbations to obtain semantically equivalent hard positive triplets. We describe in detail in the following sections.

4.4.1 Preliminaries

VQA is about answering a text question q about an corresponding image v. Classic VQA models [12, 128, 129] pursue the following paradigm.
For every triplet i, image feature and question feature extractors are applied. Multimodal feature fusion network f later is applied to output a joint vision and language representations:

$$\mathbf{f} = f(v_i, q_i) \tag{4.7}$$

The joint representation $\mathbf{f}$ (from Eq.4.7) is then used to answer prediction that drives a softmax classifier $C(\mathbf{f})$ learned by minimizing the cross-entropy loss $\mathcal{L}_{CE}$ as following:

$$\mathcal{L}_{CE} = -log(\frac{e^{f^c(h)[a]}}{\sum_{a' \in a} e^{f^c(h)[a']}}) \tag{4.8}$$

where $f^c(h)[a]$ is the logit corresponding to the answer label a.

Data Augmentation. Due to the risk of affecting semantic structure of original `<image, question, answer>` triplet and answer, we eliminate the visual augmentation. However, as shown in Figure 4.1, we follow works [8, 11] to augment the VQA dataset D with paraphrased samples question $q \in D$. As we mentioned before, the augmented dataset from Eq. 4.9 is annotated as D^{aug}:

$$D^{aug} = D \cup Q^{Para} \tag{4.9}$$

Considering a set of paraphrases for each question q from training set as $Q(q)$, model remains unchanged the triplet corresponding to each paraphrased q' as the original one:

$$Q^{Para} = f(v; q'; a) | q' \in Q \tag{4.10}$$

Figure. 4.1 shows some examples of paraphrasing given questions while they remain stable main semantic meaning that predict the same answers.

4.4.2 Supervised Contrastive Loss

Contrastive learning strategy is originated from computer vision community [4, 160, 167]. Moreover, as we mentioned in Chapter 2.6, overarching goal under contrastive learning is to learn in a way that encoded feature to be similar to positive sample while keeping away from negative ones by minimizing the self-supervised contrastive loss $\mathcal{L}_{NCE}$ using the noise contrastive estimation (NCE) [160]. Given the representation $\mathbf{f}$, then it is fed to the projection network g as:

$$z = g(\mathbf{f}) \in R^{d_z} \tag{4.11}$$

For a mini-batch of size K, the VQA model can be optimised by minimizing self-supervised contrastive loss $\mathcal{L}_{NCE}$:

$$\mathcal{L}^{i}_{NCE} = -log(\frac{e^{\phi((z_i.z_p)/\tau)}}{\sum_{k=1}^{K} \mathbb{I}_{k \neq i}.e^{\phi(z_i.z_k)/\tau}}) \tag{4.12}$$

Given a mini-batch of size K, $\mathcal{L}_{NCE}$ operates on a pair of positives, which are indicated by (z_i, z_p) and $K-1$ negative pairs (z_i, z_k), $i, p, k \in [1, K], k \neq i$. $\phi()$ computes similarity between two representations and τ is a scalar temperature parameter which is bigger than 0. The self-supervised contrastive loss $\mathcal{L}_{NCE}$ [160]

doesn't introduce the label and there is only one positive by default. However, supervised contrastive loss $\mathcal{L}_{SCL}$, introduced by [163], takes samples with labels and also contrasts the reference with more than one positive.

$$\mathcal{L}^{i}_{SCL} = -\sum_{p=1}^{|X^{+}(x_i)|} log(\frac{e^{\phi((z_i.z_p)/\tau)}}{\sum_{k=1}^{K} \mathbb{I}_{k\neq i}.e^{\phi(z_i.z_a)/\tau}}) \tag{4.13}$$

Given a reference sample x, $\mathcal{L}_{SCL}$ uses given ground-truth label information (ground-truth answer a) to make a set of intra-class positives $X^{+}(x)$ that contains samples with the same class label as reference. The calculation of $\mathcal{L}_{SCL}$ is shown in Eq. 4.14 that indicates the total loss of a minibatch K is the sum of the loss from Eq. 4.13 of each sample i.

$$\mathcal{L}_{SCL} = \sum_{p=1}^{|X^{+}(x_i)|} (\frac{\mathcal{L}^{i}_{SC}}{|X^{+}(x_i)|}) \tag{4.14}$$

In this work, we would like to catch the constant prediction as well as discrimination for VQA.

The goal of our proposed model is to explore unbiased learning by exploring the present learning paradigm of consistency in input modalities to overcome some limitations in robustness. Besides, it should be robust against the negative bias effects that caused by dataset distribution. We use contrastive learning methodology to learn the unbiased representation by contrasting the original and paraphrased features with the negative ones, as inspired by recent works such as [8]. However, opposite of the naive contrastive learning, we generate hard positive samples from embedding space that using random intra-class (positive) and random non-target negative samples. To this end, we add adversarial perturbations to both visual and language embedding space. According to very recent works [168, 209], a simple solution that can increase the robustness, is replacing the cross-entropy loss with a loss that is robust to imbalanced data distribution. However, due to the nature of contrastive loss (CL), Cross-Entropy (CE) loss can be replaced by CL to boost the generalization performance of the model in semi-supervised learning approach. Due to the robustness of VQA models (described in Chapter 2.3.1), we engaged the $\mathcal{L}_{SCL}$ to locate anchor and its paraphrasings close. Besides, using the advantages of augmented training set, Supervised Contrastive Loss (SCL) [163] has been applied to learn better representations from anchor and its paraphrased question [196] (described in 4.2.1). In this work, we argue that in multi-modal

tasks the choice of positive/negative samples depends on the variations of input modalities. Selecting "hard" negatives can lead a supervised learning method to correct its mistakes quickly [169, 170].

4.4.3 Multi-modal Contrastive Learning with Adversarial samples

Naive contrastive learning-based methods [163], concretely, apply random composition of intra-class (positives) and non-target (negatives) samples for given a batch of reference samples as positives and negatives. However, we propose a revised method to give the advantages of contrastive loss by initializing hard (true) positives and negatives instead of random sampling strategy to mitigate negative bias. Given a reference samples v_i and q_i from image and question inputs, respectively, `ADVCL` uses the given ground-truth answer label information to generate a set of intra-class positives that contains samples with the same class label as reference.

$$v_{adv} = v + \delta_v^*, \quad where \delta_v^* = \arg\max_{\delta} {}_v\mathcal{L}(v + \delta_v) \tag{4.15}$$

$$q_{adv} = q + \delta_q^*, \quad where \delta_q^* = \arg\max_{\delta} {}_q\mathcal{L}(q + \delta_q) \tag{4.16}$$

From the augmented data Q^{Para}, described above, four positive samples have been generated for every input q_i as ${q_{para}}_p{}_{p=1}^{4}$. Additionally, according to Figure **??**.b, AdvCL generates visual and textual adversarial examples on-the-fly to obtain semantically equivalent samples for two modalities, separately and jointly, as a set of $ADV(x) = \Big\{(v_{adv}, q),\ (v, q_{adv}),\ (v_{adv}, q_{adv})\Big\}$. 4 paraphrased questions *e.g.* (v, q_{para}) and generate (v_{adv}, q_{adv}) adversarial pair from Eq. 4.15 and Eq. 4.16. All of these pairs are equivalent, as they maintain the same ground-truth answer and are generated based on the same triplet in embedding space.

According to Adversarial Training Strategy [63], $\mathcal{L}$ denotes the loss function that controls the relative weight and it mostly is cross-entropy, depends on attackers, different loss have been used. TRADES proposed by [200] maximized the KL divergence between clean and adversarial logits, as opposed to the adversarial cross-entropy loss, during attack (inner maximization).

$$\mathcal{L} = \mathcal{L}_{CE}(\theta;\ x_{clean};\ y_{true}) + \mathcal{L}_{CE}(\theta;\ x_{adv};\ y_{true}) \tag{4.17}$$

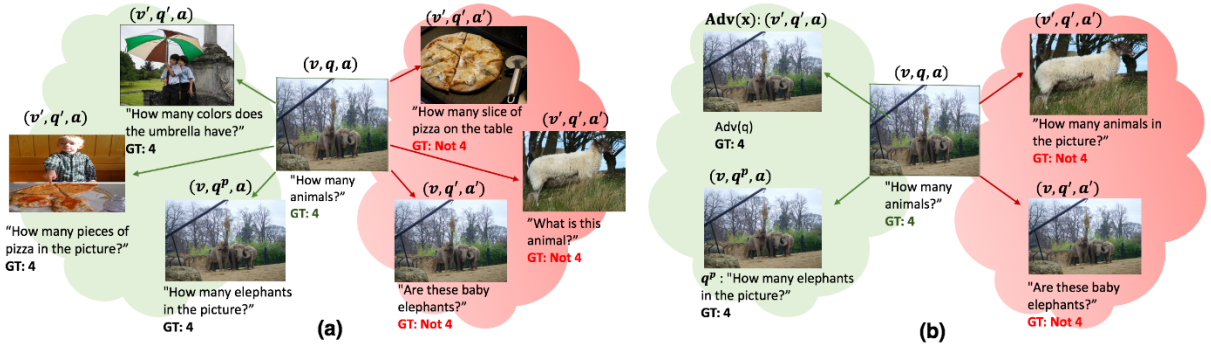

FIGURE 4.3: **Overview of proposed adversarial contrastive learning (b) VS. contrastive learning proposed by [8] (a). Our model alleviates the biases from feature by ignoring non-sense random intra-class but adding adversarial sample.**

$$\mathcal{L}_{TRADES} = \mathcal{L}_{CE}(\theta;\ x_{clean};\ y_{true}) + \beta\mathcal{L}_{KL}(\theta;\ x_{clean};\ x_{adv}) \tag{4.18}$$

β is an hyper-parameter that revisions the relative weight of adversarial samples [148]. However, to make sure the generated representations are robust enough, we apply the $\mathcal{L}_{SCL}$ instead:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{SCL}(\theta;\ v;\ q;\ a_{true}) + \beta\mathcal{L}_{SCL}(\theta;\ v;\ q_{adv};\ a_{true}) \\ +\beta\mathcal{L}_{SCL}(\theta;\ v_{adv};\ q;\ a_{true})\end{aligned} \tag{4.19}$$

We employ an efficient gradient-based attacker Iterative Fast Gradient Sign Method (IFGSM) proposed by [202] to generate visual and language adversarial hard samples:

$$\begin{aligned}q_{adv}^{t+1} &= q_{adv}^{t} + \alpha.sign(\bigtriangledown_{q}\mathcal{L}(\theta; v;\ q_{adv}^{t};\ a_{true})) \\ v_{adv}^{t+1} &= v_{adv}^{t} + \alpha.sign(\bigtriangledown_{v}\mathcal{L}(\theta; v_{adv}^{t};\ q;\ a_{true}))\end{aligned} \tag{4.20}$$

where IFGSM is the extension of FGSM [199] from Eq.4.21, for multi step of t for first step ($t = 1$), we have:

$$\begin{aligned}q_{adv}^{1} &= q + \alpha.sign(\bigtriangledown_{q}\mathcal{L}(\theta; v;\ q;\ a_{true})) \\ v_{adv}^{1} &= v + \alpha.sign(\bigtriangledown_{v}\mathcal{L}(\theta; v;\ q;\ a_{true}))\end{aligned} \tag{4.21}$$

We present our adversarial training scheme in Algorithm.1. For simplicity, we denote the process only for visual input.

According to Figure. 4.3 (b), the proposed model generates optimal samples by attacking to visual and language embeddings compared with using random intra-class samples as shown in Figure. 4.3 (a).

Algorithm 1: Pseudo code of our iFGSM Adversarial Attack Contrastive Training Batch

Inputs: Dataset D, contains a set of visual and textual examples x_{img}, x_{txt} and answer a; Num of samples N_r, ascent steps K encoder f, MModalTransformer g ;
Output: Batch for Contrastive Loss ;
Initialize $X_i = x'_i$, and $z_i = g(f(X_i))$;
Initialize $B = \emptyset$;
Initialize $B_A = \emptyset$
$\alpha = \varepsilon/\mathrm{N}$;
forall $i \in \{1, ..., N\}$ **do**
  $z_i = g(f(x_i))$
  $y_{clean} \leftarrow g(f(x_i))$
  for $t \in \{0, ..., K\}$ **do**
    $z'_i \leftarrow g(f(x'_i))$;
    compute gradient $G_{x't} = \nabla[\mathcal{L}_{SCL}(g(f_\theta(x)),\ y) + \mathcal{L}_{SCL}(g(f_\theta(x+x')),\ g(f_\theta(x))]$
    Update the perturbation x'_t *via sign gradient:*
    $x'_{t+1} = x'_t + \alpha.sign(G_{x'_{t+1}})$
  return $B_A \cup X_{ADV} = x + x'$
forall $i \in \{1, ..., |B_A|\}$ **do**
  $x_i = D^{aug}$;
  $x_i^+ = X_{ADV}^+$
  $x_i^- = X_{txt}^-$
  append $B = B \cup \{x_i, x_i^+, x_i^-\}$
return B

Our method is structurally similar to [8] which uses contrastive learning to train VQA model to leverage informative information from similar representations of joint $V + L$ With modifications for supervised learning. The supervised contrastive loss is computed on the outputs of the projection network. To use the trained model for classification, we train a linear classifier on top of the frozen representations using a cross-entropy loss.

Different from self-supervised learning tasks that the contrast occurs between the anchor and anchor augmentations for positive samples, here, we select positive samples from both augmented and generated data. As a negative sample we explore

the random and .

$$X^{-}(x) = \left\{(v^{-}, q^{-}, a^{-})|a^{-} \neq a\right\} \tag{4.22}$$

$$X^{-}(x) = \left\{(v^{-}, q^{+}, a^{-})|a^{-} \neq a\right\} \tag{4.23}$$

where

$$v^{-} = v + \delta_{v}^{*} \tag{4.24}$$

To create mini-batches for $\mathcal{L}_{SC}$, similar to [8], we start by filling our batch with triplets of reference x_i, a intra-class positive and a non-target negative sample. This procedure is repeated for specified number of times N_r to create a batch B. Given a triplet of image I, question Q, and ground truth answer A, generally VQA model can be formulated as a transformation $H : (Q, I) \rightarrow A^{'}$, where the joint representation H is then used to predict a probability distribution over the answer space A with a $softmax$ classifier learned by minimizing the cross-entropy loss ($\mathcal{L}_{CE}$):

To strengthen the visual and textual joint representations robustness w.r.t semantic linguistic variations in the questions, we apply contrastive loss to leverage better information. The contrastive loss encourages the encoded instance features to be similar to positive keys while keeping away from negative ones. Moreover, we explore the importance of choosing true positive and negative samples For multimodal task by generating different pairs in dual contrastive training paradigm contains two branches. By utilizing the rephrasing questions with the similar answer as anchor, question with different answers as anchor, and random images as anchor as positive question sample; Q^{+}, positive question sample; Q^{-}, and negative image sample; I^{-} respectively, we augment original dataset D.

$$D \cup (v^{+}, v^{-}, q^{+}, q^{-}) \tag{4.25}$$

Different strategies to define positive and negative samples for reference sample, are qualified based on the quality of generated representations.

Hard samples and Batch Creation- Inspired with supervised contrastive loss (SCL) proposed by [163], we utilize question rephrasings and introduce intra-class generated hard positive samples from multi-modal image-question representations. Moreover, we ignore and random positives and negatives to lead model learn useful

consistent and discriminative representation per answer class. Using true positive and negative samples for referenced sample depicted in Figures. 4.2 and 4.3.

Algorithm 2: Pseudo code of our iFGSM Adversarial Attack contrastive training Batch using cross-entropy

Inputs: data D= $\{x_{img}, x_{txt}, y\}$; Num of samples N_r, ascent steps K encoder f, MModalTransformer g ;
Output: Batch for Contrastive Loss;
Initialize $X_i = x'_i$, and z_i=$g(f(X_i))$;
Initialize $B = \emptyset$;
Initialize $B_A = \emptyset$
forall $i \in \{1, ..., N\}$ **do**
  $z_i = g(f(x_i))$
  $y_{clean} \leftarrow g(f(x_i))$
  for $t = 0, ..., K$ **do**
    $z'_i = g(f(x'_i))$
    compute gradient $G_{x'_t} = \nabla[\mathrm{L}_{cl}(g(f_\theta \mathrm{CosSim}[(\mathrm{x}+ \mathrm{x}'), x]), y)$
    Update the perturbation x'_t via sign gradient:
    $x'_{t+1} = x'_t + \alpha.\mathrm{sign}(\mathrm{G}_{x'_{t+1}})$
return B_A
forall $i \in \{1, ..., |B_A|\}$ **do**
  $x_i = D_{aug}$;
  $x_i^+ = X_{ADV}$;
  $x_i^- = X_{rand}$;
  append $B = B \cup \{\mathrm{x}_i, x_i^+, x_i^-\}$;
return B

4.5 Experiments on Visual Question Answering

We conduct several experiments to study whether the introduced method is able to effectively learn the high-level representations with respect to consistency and discriminative.

4.5.1 Robust VQA Datasets

As we mentioned in Chapter 2, some robust datasets have been collected to analyse VQA model robustness from different aspects. Due to this thesis goal, we evaluate the proposed VQA model on VQA-Rephrasings [11] which is made to examine

model vulnerability cross input language variations. Besides we apply the proposed model on balanced but challenging benchmark: VQA2.0 [143]. In comparison with some other datasets like VQA1.0 [12], VQA2.0 contains much more question-image pairs for train and validation set. Question-answer pairs are categorized with answer types: *i.e.* "Yes/No", "Number", "Other". For fair comparison, we perform experiments on both validation and test set (provided in Table. 4.4).

4.5.2 Metrics

From [12], the standard metric for evaluating VQA is accuracy:

$$accuracy = \min(\frac{\#\text{Humans votes}}{3}, 1) \tag{4.26}$$

As we mentioned earlier, overall accuracy is not enough to evaluate model performance with robustness. Since, it doesn't indicate the consistency in model predictions. Shah *et al.* [11] also proposed a comprehensive metric called Consensus Score (CS) as an evaluation metric to control the model robustness again top$k$ rephrasings of each testing question. For a set of paraphrases $Q(q)$ questions for reference question q, the consensus score CS(k) is defined as in Eq.4.28 [11]:

$$CS(k) = \sum_{Q' \subset Q, |Q'|=k} \frac{S(Q')}{{}^{n}C_{k}} \tag{4.27}$$

$$S(Q') = \begin{cases} 1 & \text{if } \forall q \in Q' \phi(q) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.28}$$

${}^{n}C_{k}$ denotes number of subsets k from a set of rephrasings with n number of samples. Notably, $CS(k)$ is becoming zero if the model answers at least k questions correctly.

4.5.3 VQA Model Architecture

Inspired with [130], we use a multimodal transformer (MMT) as function f in Figure 4.2. It takes as input two visual and language modalities; image-question pair.

Question Feature Embedding- Question are embedded using a pre-trained three layer BERT encoder [142] that is fine-tuned along with the MMT [130].

Image Embedding- Image embedding uses features by extracting features from ResNeXT [16] based Faster R-CNN model [17].

Multi-modal Fusion- We adopt a popular attention model from [210, 211] works to calculate the multi-modal features $\boldsymbol{m} \in \mathbb{R}^{1024}$ given each pair of image-question i:

$$\boldsymbol{m} = F(\hat{\boldsymbol{v}}, \boldsymbol{q}), \tag{4.29}$$

$\hat{\boldsymbol{v}} = \sum_{i=1}^{N} \alpha_i \boldsymbol{v}_i$ is the attentive image feature from the input feature set $\{\boldsymbol{z}_i\}$. F denotes the multi-modal feature function that we use the fuse model from in [174] as in Eq. 4.30:

$$F(\boldsymbol{u}, \boldsymbol{v}) = \text{ReLU}(\boldsymbol{W}_1\boldsymbol{u} + \boldsymbol{W}_2\boldsymbol{v}) - (\boldsymbol{W}_1\boldsymbol{u} - \boldsymbol{W}_2\boldsymbol{v})^2, \tag{4.30}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ project embedding input representations into the same dimension to implement Eq. (4.29).

Classifier- Similar to straight-forward models, classifier maps function h into a score vector contains scores over answer candidates follow by softmax classifier. Instead, we add non-linear layer $\boldsymbol{g}$ after getting h. Finally, model passes the fused $\boldsymbol{g} \cdot [\boldsymbol{m}]$ into the softmax classifier.

4.5.4 Implementation Setting

We generated hard positives and negatives that enable contrastive learning to contrast useful input variations from useless representations. All models are trained using the AdamX Optimizer [212], with learning rate set as 1×104 and β set as 0.98. Experiments have been conducted on 3 GTX 1080TI GPU with 60 and 128 batch sizes for contrastive and cross-entropy learning due to the limitations of memory. The learning rate lr and β are both initialized with 0.1. ResNet [15] backbone is used for all models except for the black-box experiments.

Model	CS (VQG)		CS (BT)		VQA Scores		
	$k=3$	$k=4$	$k=3$	$k=4$	val	test-dev	test-std
BUTD [129]	40.5	34.5	-	-	63.3	-	-
BUTD+CC [11]	44.7	42.5	-	-	-	-	-
Pythia [187]	45.9	39.5	-	-	65.8	68.4	-
Pythia+CC [11]	50.9	44.3	-	-	66.0	68.9	-
ConClaT [8]	55.3	52.3	**70.4**	**70.0**	67.0	**68.5**	70.0
baseline	52.8	50.0	68.2	67.7	64.1	65.1	65.3
Ours-ADVCL	**56.4**	**53.3**	69.6	68.7	**68.2**	67.7	**70.3**

TABLE 4.1: **Proposed method vs existing methods/baselines on VQA-Rephrasings and VQA v2.0.** For test-dev and test-std, we train our model on train+val set of VQA v2.0.

Model	CS				VQA Scores	
	$k=1$	$k=2$	$k=3$	$k=4$	Orig	Rep
MUTAN [214]	56.7	43.6	38.9	32.7	59.1	46.8
BUTD [129]	60.5	46.9	40.4	34.5	61.5	51.2
BUTD+CC [11]	61.7	50.8	44.7	42.5	62.4	52.6
Pythia [187]	63.4	52.0	45.9	39.5	64.1	54.2
Pythia+CC [11]	64.4	55.4	50.9	44.3	64.5	55.6
Ours-ADV	66.8	59.5	55.1	51.9	-	66.0
Ours-ADVCL	**67.1**	**59.7**	**55.3**	**52.2**	-	**65.7**

TABLE 4.2: **Consensus performance on VQA-Rephrasings dataset using VQG** Baseline results are copied from [11]

4.5.5 Comparison with state-of-the-arts

4.5.5.1 Comparing Methods

We compare ADVCL with classic: PYTHIA [187], Bottom-Up-Attention and Top-Down (BUTD) [129] and state-of-the-art settings: VQA+CC [11] and Contrast and Classify Training (CONCLAT) [8]. Besides, we used two baselines: first one is MMT model without any additional training branch and using cross-entropy loss ($\mathcal{L}_{CE}$), that is denoted as BASELINE. Another one which is denoted as OURS-ADV is MMT with adversarial attack training that uses $\mathcal{L}_{CE}$ for both clean and noisy status. OURS-ADV is very close to the model proposed by VILLA [213]. For fair comparison, the have searched for methods using the same Faster-RCNN features [17] similar to ours.

Data Split	MUTAN	BUTD	BUTD+CC	Pythia	Pythia+CC	BAN	BAN+CC	UNITER	VILLA	Ours
Original	59.1	61.5	62.4	64.1	64.5	65.0	65.9	70.3	71.2	-
Rephrasing	46.9	51.2	52.3	54.2	55.7	55.9	56.6	64.6	65.4	**65.7**

TABLE 4.3: Result on VQA-Rephrasings. Baseline Results are copied from [11]

	VQA-CP v2 test(%)				VQA-v2 val (%)			
Model	Yes/No	Num.	Other	Overall	Yes/No	Num.	Other	Overall
GVQA [144]	-	-	-	31.3	72.0	31.2	34.6	48.2
UpDn [129]	42.3	11.9	46.0	39.7	81.2	42.1	55.6	63.5
S-MRL [215]	42.8	12.8	43.2	38.5	-	-	-	63.1
NSM [216]	-	-	-	45.8	-	-	-	-
BUTD [129]	42.3	11.9	46.0	39.7	81.2	42.1	55.6	63.5
RUBi [215]	42.8	12.8	43.2	38.5	-	-	-	63.1
MUREL [217]	42.9	13.2	45.0	39.5	-	-	-	65.1
LXMERT [135]	42.8	18.9	55.5	46.2	83.3	46.2	56.9	65.3
methods based on data-augmentation and training strategy								
CSS [218]	84.4	49.4	48.2	58.9	73.3	39.8	55.1	59.9
CL-VQA [168]	86.9	49.9	47.2	59.2	67.3	38.4	54.7	57.3
Loss-Rescaling [219]	72.8	48.0	44.5	53.3	68.2	36.4	52.3	56.8
MUTANT [214]	88.9	49.7	50.8	61.7	82.1	42.5	53.3	62.6
RandImg [220]	83.9	41.6	44.2	55.4	76.5	33.9	48.6	57.2
Unshuffling [221]	47.7	14.4	47.2	42.4	78.3	42.2	52.8	61.1
ADA-VQA [222]	87.4	53.0	46.8	59.6	78.8	42.2	54.4	61.9
Ours-ADV	-	-	-	-	83.0	48.8	57.3	66.0
Ours-ADVCL	**88.3**	**56.0**	**62.8**	**62.1**	**85.2**	**51.2**	**60.1**	**67.3**

TABLE 4.4: **Comparison to task-specific state-of-the-arts on VQA-CP v2 test VQA v2.0 validation split**.

4.5.5.2 Quantitative Results

Table 4.1 reports the comparison of our model performance with various state-of-the-art VQA methods on test-std, test-dev and consensus score (CS(k)) for $k = 3$, 4 on VQA-Rephrasings [11] and VQA Accuracy on VQA v2.0 [143] datasets. For fair comparison, we provide CS(k) performances on both augmented data by VQG and BT. We compare our model with the BUTD [129], Pythia, VQA+CC [11] ConclaT [8] settings. Our method outperforms Conclat gains of 1.2%, on validation (CS(k)) for $k = 3$ and $k = 4$ respectively.

Table 4.2 also provides further comparison between our proposed model and state-of-the-arts due to original dataset and augmented dateset using positive question rephrasing on consensus score (CS(k)) for k = 1, 2, 3, 4. Following Table 4.2, Table 4.3 provides further comparison between our proposed model and the state-of-the-arts due to the original dataset and augmented dataset using positive question rephrasing.

FIGURE 4.4: **Qualitative Examples.** Predicate of ADVCL and our BASELINE on several image-question pairs and their corresponding rephrased questions.

Table 4.3 further provides the performance based on different question types of some other state-of-the-art methods [8, 11, 129, 215, 218]. In summary, our proposed model achieves state-of-the-art robust performance on the VQA v2.0 dataset by using robust metrics that show the robustness of ADVCL across language variations.

4.5.6 Qualitative Analysis

Figure 4.4 visualizes the ability of ADVCL model to learn more consistent representations in comparison with the baseline . According to these examples, ADVCL improves the consistency in predictions across the rephrasings (**Q1**, **Q2**, **Q3**) of the original question **Q**. Note that these qualitative results were produced from the data augmented via back-translation (BT).

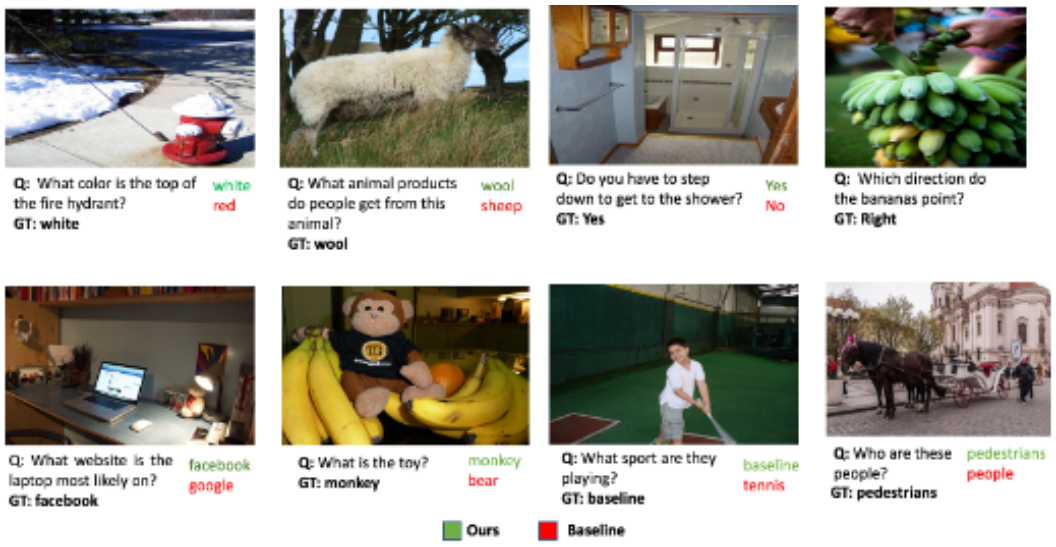

FIGURE 4.5: **Qualitative Examples.** Visualization of examples collected from ADVCL predictor for complicated questions and unbiased samples in compare with our second baseline OURS-ADV.

We further show some qualitative samples in Figure 4.5 comparing the baseline and ADVCL on other aspects of VQA robustness: deal with complicated questions (upper row) and biased samples (lower row). For example, the ground-truth answer to "what color is the hydrant?" in most cases is "red" due to the imbalanced datasets. However, the correct answer for the given image and question "what color is the top of hydrant?" in Figure 4.5 is "while". Hence, the proper model has to provide high-level representations to tackle the negative effects of language bias as well as complicated reasoning to distinguish between "top of hydrant" and "hydrant" in visual input.

4.5.7 Ablation Studies

For simplicity and conciseness, we omitted CS(1) and CS(2) scores in the comparison part (Section. 4.5.5.2). We provide them in Table 4.5.

	CS				CS (BT)			
Model	$k=1$	$k=2$	$k=3$	$k=4$	$k=1$	$k=2$	$k=3$	$k=4$
ConclaT [8]	-	-	55.3	52.3	68.6	61.4	57.1	54.0
baseline	64.0	58.9	54.8	51.8	67.5	60.0	55.4	52.3
Ours-ADVCL	67.1	59.7	55.3	52.2	72.1	70.7	70.0	69.6

TABLE 4.5: Ablations Study. experiments are run with Back Translation data.

Ablation on Adversarial Attackers- We have applied some attackers to see the effects of different attackers on network performance. Specially, we have tried to PGD[63] and FGSM [199] on feature embedding space. PGD is the simplest and effective method that adds the random noise to original sample representation. Thee structure in our proposed vQA training paradigm are shown in Algorithm. 1 and Algorithm. 3.

Ablation on training loss- In addition to the structure introduced in Section 4.4, we experiment with various schemes of combining different learning schemes and losses. Figure. 4.6 shows one model which alternatively trains with three models of using $\mathcal{L}_{CE}$, $\mathcal{L}_{AT}$ and $\mathcal{L}_{SC}$. Note that, for adversarial training, we have applied $\mathcal{L}_{CE}$. However, we denoted it as $\mathcal{L}_{AT}$, to emphasize learning paradigm. Comparing to ADVCL, JOINT-VQA framework, is not optimal with respect to unbiased representation learning.

Table 4.6 shows the parameters that have been applied to get best performance.

Algorithm 3: Pseudo code of our AdversialAttack contrastive training Batch

Inputs: data D; Num of samples N_r, ascent steps K encoder f, MModalTransformer g ;
Output: Batch for Contrastive Loss;
Initialize $X_i = x_i^{'}$, and $z_i = g(f(X_i))$;
Initialize $B = \emptyset$;
Initialize $B_A = \emptyset$
forall $i \in \{1, ..., N\}$ **do**
  $z_i = g(f(x_i))$
  for $t = 0, ..., K$ **do**
    $z_i^{'} = g(f(x_i^{'}))$
    $Sim = (z_i^{'} z_i)$
    compute gradient $\nabla_{x_i^{'}}$
    Update the perturbation $x_i^{'}$ via gradient ascend:
    gimg $\leftarrow \nabla_{x_i^{'}}[\mathrm{L}_{ce}(f_\theta(x + x^{'}, xtxt), y) + L_{kl}(f_\theta(x + x^{'}, xtxt), \tilde{y})]$
    where $\tilde{y} = g(f(X_i)$
    Update the perturbation $x_i^{'}$ via sign gradient:
    append $B = B \cup \{\mathrm{x}_{img}^{'} i, \mathrm{x}_{txt}^{'} i\}$
return B_A
forall $i \in \{1, ..., |B_A|\}$ **do**
  $\mathrm{x}_i = D$
  $\mathrm{x}_i^{+} = X_{cls}$
  $\mathrm{x}_i^{-} = X_{rand}$
  append $B = B \cup \{x_i, x_i^{+}, x_i^{-}\}$;
return B

#	Hyperparameters	Value	#	Hyperparameters	Value
1	Maximum question tokens	23	**8**	Maximum object tokens	101
2	Embedding size	768	**9**	Number of TextBert layers	3
3	Similarity Threshold	.95	**10**	Number of Multimodal layers	6
4	Vocabulary size	3129	**11**	Number of attention heads	12
5	Warm-up learning rate factor	.1	**12**	Multimodal layer dropout	.1
6	Learning rate decay steps	10665, 14931	**13**	Optimizer	ADAM
7	Projection Dimension (R^{d_z})	128	**14**	Warm-up iterations	4266

TABLE 4.6: Hyper-parameters choice for the proposed model

4.6 Conclusion

In this chapter, we demonstrated the importance of learning stability feature to enhance robustness of VQA model to linguistic variations in questions. In particular, ADVCL preserves model consistency, or robustness, as well as model correctness, or discrimination, by jointly applying contrastive and cross-entropy losses. ADVCL deals with model consistency and achieves improvement in the comprehensive

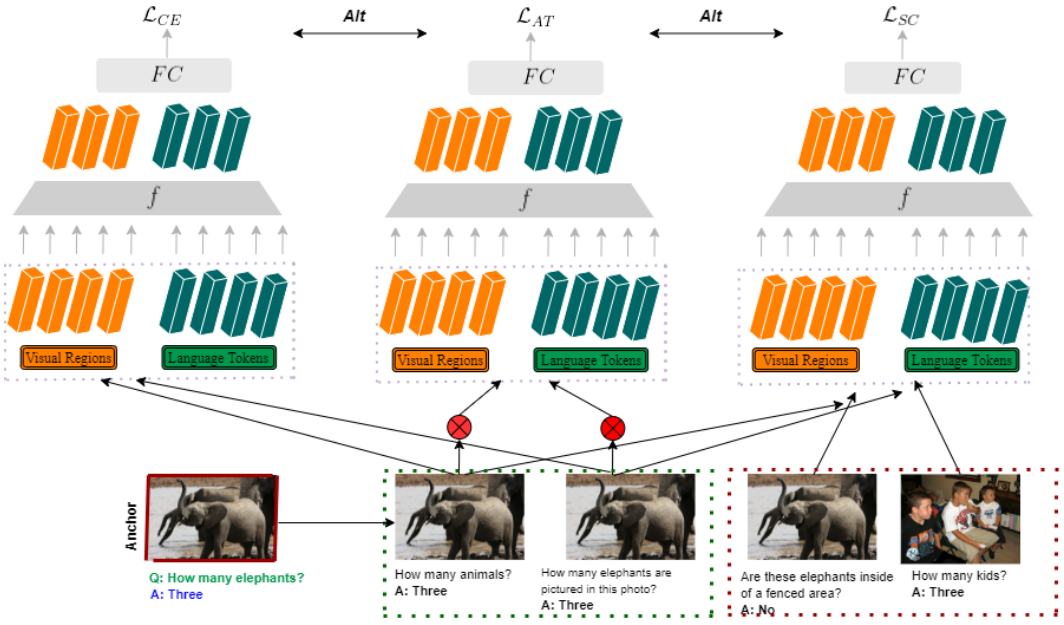

FIGURE 4.6: **Overview of joint robust VQA**

metric [11], @K (CS@K). On the VQA-Rephrasings benchmark, `ADVCL` improves consensus score by 1.57% over a baseline and improves the state-of-the-art model ([11]) score from 48.2 to 53.3. In addition, on the standard VQA v2.0 benchmark, our model improves VQA accuracy by 0.78% overall.

Chapter 5

Summary and Future work

5.1 Summary

In this thesis, we have proposed different feature representation and learning approaches to enhance robustness of deep neural networks for relation annotation and question answering. Chapter 2 goes over the related work for various semantic representation learning and deep representation learning models. Our work presented in this thesis is inspired by those methods.

Before introducing how to incorporate consistency in visual representation learning in Chapter 3, we explore the importance of learning feature alignment for dynamic object feature concatenation in relation prediction, where objects may not only have large overlaps in their spatial locations but also have a lot of commonalities in their semantic concepts. To this end, we introduced a meta-architecture — learning-to-align — for dynamic object feature concatenation, named ALIGN R-CNN framework with an object context encoding alignment module to explicitly learn complementary object alignment for predicting each object pair relation. After generating the complementary aligned features, the generated pairwise features and the region features are effectively combined as the final relation feature. ALIGN R-CNN transfers information between the pairwise nodes using multi-head hard-attention in a differentiable manner to infer the best alignment parts per pair. Subsequently, it re-aligns object features by gathering features from the top-K semantic-rich portions of the paired objects, thereby facilitating with alignment

accuracy. This enables the scene graph generator to satisfy the discriminant criterion for pairwise region-based relation representation by generating descriptive and human-like annotations. After introducing this scene graph generator for extracting scene graphs from images, we further exploited the trained re-aligned features to recognize unseen components with no labelled instances in the training set.

In Chapter 4, we seek the robust visual reasoning to present consistence learning paradigm for multi-modal reasoning task such visual question answering to overcome some limitations toward robustness. Unsupervised visual representation learning has shown encouraging progress recently, thanks to the introduction of contrastive learning [160, 163] framework. Moreover, we pointed out the naive contrastive positive and negative samples those are randomly picked are sub-optimal. Hence, to overcome this issue, we generate hard positive samples from adversarial perturbation samples. Experimental results on two benchmarks, VQA2.0 and Praphrased VQA, show that AdvCL is robust to question variations and less-frequent data.

5.2 Recommendations for Future Work

The role of feature representation learning to improve the robustness of deep neural networks is becoming more and more important. However, there is still lots of work to explore in robust representation learning. For instance, data augmentation [223] and generative training [63, 202] have shown improvements in the robustness of DNNs. However, this does not address the ability of the model to handle unseen manipulations.

One solution is using integrated multiple biased strategies, which makes the model deal with complex samples like rare or unseen ones. This integrated unbiased model makes the biased strategies overfit the dissimilar data distribution, which makes the model to learn unbiased and fine-grained feature representation to solve examples that are hard to solve by biased models.

In the following section, some recommendations about how to further develop the network will be discussed.

5.2.1 Causality

As we mentioned before, one of the widespread problems in DNNs is the negative effect of bias that is caused by imbalanced datasets. To date, training paradigms mostly rely on correlations rather than predictions. Given input X, and prediction Y, by training a model based on correlation $P(Y|X)$, the model misses the true causal effect from X to Y [224, 225] and cannot eliminate the negative effect of the confounding bias [226].

The confounder bias from correlation $P(Y|X)$ is defined as a common cause of X and Y that induces spurious correlation even if X and Y have no direct causation [227]. Confounder bias manifests differently in different tasks. For instance, in visual relation annotation, frequent predicate categories, *e.g.* `riding` for (`person-horse`) pair in the training set, denote that by detecting objects `person` and `horse`, predict `riding`, even if there is no such visual relationship between `person` and `horse` [2].

Besides, common sense can be assumed as a triplet *e.g.* `person-riding-horse`. In this case, the detection of object pairs $O = person, horse$ is biased by [228]: $C \rightarrow O$ path. Given a test image with "`person-sitting on- horse`", predictor that has been effected by *confounding* bias in a causal view: the backdoor [229]: $C \rightarrow O; O \rightarrow Y$, considers `person` with `horse` to reason `riding`. Hence, besides data augmentations that help to reduce confounder distributions, building a robust model by taking causal predictions with respect to data has been demanded recently. Compounding *causal inference* [224, 225] into DNNs [230] has been explored in several tasks [146, 227, 231, 232]. Causal inference improves the model robustness by analyzing the interactions between features/regions of interest and identifying the confounding bias. By replacing $P(Y|X)$ with $P(Y|do(X))$, the true causal effect between X and Y has been calculated. This model leads to inferring the true prediction by ignoring the *negative* or *confounding* bias using backdoor adjustment [229] that removes backdoor confounding direction and keeps the true causal direction by adding *mediator* Z causal effect through: $X \rightarrow Z \rightarrow Y$. Finally, the direct causal effect : $X \rightarrow Y$ is considered as the final prediction logits [226].

$$P(Y|do(X)) = \sum P(X = x)P(Y|X = x, Z), \tag{5.1}$$

In general, the counterfactual is the ability to infer negative bias from training samples to increase the unbiased prediction in a model. By adding this module to the proposed model in Chapter 3, we hope to reason about the contributions of different regions of objects via performing counterfactual causality on a causal graph.

5.2.2 Causal Attention

In Chapter 3, we explored visual attention to refine and generate feature alignment for learning less biased but more informative feature representation. Moreover, we used hard-attention to alleviate the weak approximation issue of soft-attention that widely used in different machine learning and computer vision tasks. For instance, `on` is a more "frequent but the less meaningful" predicate that can be replaced with the "meaningful" predicate *e.g.* `sitting on`.

Due to recent demand to identify *true* causal effects from input to output by incorporating causal inference [224, 225], causal attention [227] is an adjustment to apply to any attention-based architecture. In order to improve the proposed network in Chapter 3, we can apply causal attention to solve the mentioned issue, and we could utilize it for any visual-language tasks that are driven by attention. Consider refined pairwise features as Z, the correlation training could be re-implemented by:

$$P(Y|X) = \sum P(Z = z|X)P(Y|Z = z), \tag{5.2}$$

Hence, inspired by [227], we would like to extend our attention-based model by adding causal inference as:

$$P(Y|do(X)) = \sum P(Z = z|X)P(X)P(Y|Z = z), \tag{5.3}$$

where, Z is a multi-head attention applied to a set of object pairs to refine pairwise features X. For VQA model in Chapter 4, input X will be considered as a multi-modality set for images and corresponding questions. And the causality module explores the true casual effect of the input question and its corresponding image regions for re-weighting and to generate an answer.

5.2.3 Long-Tailed Distribution

So far, we have argued about the imbalanced dataset distribution problem and proposed a solution to alleviate the lack of model robustness by generating discriminative and consistent feature representations. However, the achieved performances are still limited by the real-word data long-tailed datasets. The imbalanced data distribution and, simultaneously, the deficiency in sample numbers are two issues induced by long-tailed datasets [233]. Despite recent work on long-tailed distribution issues [234–237], there are many unanswered questions and unsolved problems for various tasks. If we analyse the overall accuracy (*e.g.* SGG performance and VQA accuracy in Chapter 3 and Chapter 4, respectively) and robust accuracy (*e.g.* comprehensive metrics in Chapter 3 and Chapter 4), per class label, we would observe that overall performance drops from head to tail (frequent to rare samples). For instance, our proposed method in Chapter 3, by generating fine-grained representations and by zero-shot learning could explore the unevenly distributed predicates. However, the gap between overall and robust performances ($ACC_{overall}$ and ACC_{robust} respectively) indicates the model vulnerability of tail predicates:

$$E_{rror} = ACC_{overall} - ACC_{robust} \tag{5.4}$$

As we discussed before, robust training paradigms like Adversarial Training (AT) [63, 199] are widely adopted in developed algorithms as they increase both overall and robust performances against adversarial attacks [32]. However, a recent study [233] on long-tailed problems argues that adding adversarial perturbations in embedding space damages the overall performance achieved by tail classes significantly [233]. AT has been well used for re-scaling data and lacks of fundamental theory that is re-imbalancing data.

Inspired by these observations, for future work, we would like to continue our research to alleviate the vulnerability of tail label predictions by adding causal inference (Section 5.2.1) with robust adversarial training to explore and ignore the negative effect of robustness from general performance, reducing the *error* gap from Eq. 5.4 [200] while increasing the performance.

1

Journal Articles

- **Mitra Tajrobehkar**, Kaihua Tang, Hanwang Zhang, Joo-Hwee Lim, "Align R-CNN: A Pairwise Head Network for Visual Relationship Detection," *IEEE Transactions on Multimedia, 2021.*

[1]The superscript * indicates joint first authors

Bibliography

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. xix, 7

[2] Mitra Tajrobehkar, Kaihua Tang, Hanwang Zhang, and Joo Hwee Lim. Align r-cnn: A pairwise head network for visual relationship detection. 2021. xix, 15, 18, 79

[3] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv*, 2018. xix, 6, 25

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. xix, 6, 25, 26, 60

[5] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6619–6628, 2019. xx, xxi, 10, 14, 17, 18, 29, 30, 31, 38, 39, 40, 41, 43, 44, 45, 51

[6] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. 2017. xx, 10, 14, 18, 29, 30, 31, 38, 40, 41, 42, 43, 44, 45, 46, 47, 51

[7] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3107–3115, 2017. xx, xxi, 14, 30, 31, 39, 40, 41, 44, 52

[8] Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust vqa models, 2021. xx, 22, 26, 54, 60, 61, 63, 64, 65, 69, 70, 71, 73

[9] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 3097–3106, 2017. xxi, 14, 17, 18, 29, 30, 39, 40, 41, 42, 43, 44, 45

[10] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722, 2020. xxi, 3, 24, 30, 38, 40, 41, 43, 44

[11] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6649–6658, 2019. xxi, 22, 23, 53, 54, 55, 59, 60, 66, 67, 69, 70, 71, 75

[12] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1, 2, 21, 22, 53, 59, 67

[13] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. 1, 2, 22

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1026–1034, 2015. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6, 7, 68

[16] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 43, 68

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6), 2017. 1, 16, 22, 32, 43, 68, 69

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision(ECCV)*, 2016.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 5, 16, 31, 32, 38, 43, 44

[20] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *arXiv preprint arXiv:1801.00868*, 2018. 1

[21] Russell Epstein and Nancy Kanwisher. A cortical representation the local visual environment. *Nature*, 392:598–601, 1998. 1

[22] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73:415–434, 2012.

[23] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111:8619–8624, 2014. 1

[24] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 2013. 1, 5

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012. 1, 2, 5, 7, 55

[26] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[27] Wenfeng Zheng, Xiangjun Liu, Xubin Ni, Lirong Yin, and Bo Yang. Improving visual reasoning through semantic representation. *IEEE Access*, 9: 91476–91486, 2021. 1, 2

[28] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, 10:2137–2155, 2009. 2

[29] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *NIPS*, 2016. 3, 9

[30] Safa Cicek, Alhussein Fawzi, and Stefano Soatto. Saas: Speed as a supervisor for semi-supervised learning. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206, pages 152–166, 2018.

[31] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. KE-GAN: knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5237–5246, 2019. 3

[32] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*, 2013. 3, 53, 81

[33] Jürgen Schmidhuber. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249*, 2015. 4

[34] Honglak Lee, Roger B. Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning (ICML)*, pages 609–616, 2009. 4

[35] Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29:82–97, 2012. 5

[36] Dan C. Ciresan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3642–3649, 2012. 5

[37] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision(ECCV)*, pages 818–833, 2014. 5

[38] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005. 5, 6

[39] Chiranjibi Sitaula, Sunil Aryal, Yong Xiang, Anish Basnet, and Xuequan Lu. Content and context features for scene image representation. *Knowl. Based Syst.*, 232:107470, 2021. 6, 7, 8

[40] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006. 6

[41] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.*, 42(3): 145–175, 2001. 6

[42] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision(ECCV)*, volume 8695, pages 392–407, 2014. 6, 7

[43] Ilja Kuzborskij, Fabio Maria Carlucci, and Barbara Caputo. When naïve bayes nearest neighbors meet convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 2100–2109, 2016.

[44] Chiranjibi Sitaula, Yong Xiang, Sunil Aryal, and Xuequan Lu. Unsupervised deep features for privacy image classification. volume 11854, pages 404–415, 2019.

[45] Chiranjibi Sitaula and Sunil Aryal. Fusion of whole and part features for the classification of histopathological image of breast tissue. *Health Inf. Sci. Syst.*, 8:38, 2020. 7

[46] Chiranjibi Sitaula and Mohammad Belayet Hossain. Attention-based VGG-16 model for COVID-19 chest x-ray image classification. *Appl. Intell.*, 51(5): 2850–2863, 2021. 6, 7

[47] Yanming Guo and Michael S. Lew. Bag of surrogate parts: one inherent feature of deep cnns. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016. 6

[48] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.

[49] Pengjie Tang, Hanli Wang, and Sam Kwong. G-MS2F: googlenet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing*, 225: 188–197, 2017. 7

[50] Chunjie Zhang, Guibo Zhu, Qingming Huang, and Qi Tian. Image classification by search with explicitly and implicitly semantic representations. *Information Sciences*, 376:125–135, 2017. 7, 8

[51] Yanming Guo, Yu Liu, Songyang Lao, Erwin M. Bakker, Liang Bai, and Michael S. Lew. Bag of surrogate parts feature for visual recognition. *IEEE Trans. Multim.*, 20(6):1525–1536, 2018. 6, 7

[52] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv*, 2020. 6

[53] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 766–774, 2014. 6

[54] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1249–1258, 2016.

[55] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision(ECCV)*, pages 791–808, 2016.

[56] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresnet: Environmental sound classification based on visual domain models. In *International Conference on Pattern Recognition, ICPR*, pages 4933–4940, 2020. 6

[57] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 7

[58] Shuang Bai, Huadong Tang, and Shan An. Coordinate cnns and lstms to categorize scene images with multi-views and multi-levels of abstraction. *Expert Syst. Appl.*, 120:298–309, 2019. 7

[59] Chiranjibi Sitaula, Yong Xiang, Sunil Aryal, and Xuequan Lu. Scene image representation by foreground, background and hybrid features. *Expert Syst. Appl.*, 182, 2021. 7

[60] Li-jia Li, Hao Su, Li Fei-fei, and Eric Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems*, volume 23, 2010. 7

[61] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *arXiv*, 2020. 7

[62] Dongzhe Wang and Kezhi Mao. Task-generic semantic convolutional neural network for web text-aided image classification. *Neurocomputing*, 329:103–115, 2019. 8

[63] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2018. 8, 57, 58, 59, 62, 73, 78, 81

[64] Chen Wei, Huiyu Wang, Wei Shen, and Alan L. Yuille. CO2: consistent contrast for unsupervised visual representation learning. In *International Conference on Learning Representations, (ICLR)*, 2021. 9

[65] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Neural Information Processing Systems, (NeurIPS)*, 2020. 9

[66] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6163–6171, 2019. 10, 14, 17, 18, 30, 40, 41, 43, 44, 45

[67] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017. 10, 24, 30, 39, 43

[68] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015. 13, 14

[69] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *European Conference on Computer Vision(ECCV)*, 2016. 13, 14, 17

[70] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, and Petros Maragos. Attention-translation-relation network for scalable scene graph generation. In *International Conference on Computer Vision (ICCV)*, 2019. 14, 29, 39, 40, 42, 44

[71] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 17, 40, 44

[72] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *European Conference on Computer Vision(ECCV)*, 2018.

[73] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8188–8197, 2019.

[74] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. *arXiv preprint arXiv:1704.03114*, 2017. 14, 17, 33, 39, 42

[75] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *International Conference on Computer Vision (ICCV)*, pages 1270–1279, 2017. 14, 17, 18, 29, 39

[76] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *Proc. Eur. Conf. Comput. Vis*, volume 11205, pages 346–363, 2018. doi: 10.1007/978-3-030-01246-5_21. 17

[77] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Proc. Eur. Conf. Comput. Vis*, volume 11207, pages 330–347, 2018.

[78] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17

[79] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision(ECCV)*, 2020. 14, 29

[80] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *European Conference on Computer Vision(ECCV)*, pages 711–727, 2018. 14, 17, 29

[81] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10685–10694, 2019.

[82] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Trans. Multimedia*, 19 (9):2045–2055, 2017.

[83] Longteng Guo, Jing Liu, Shichen Lu, and Hanqing Lu. Show, tell, and polish: Ruminant decoding for image captioning. *IEEE Trans. Multimedia*, 22(8): 2149–2162, 2020. 14

[84] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 14

[85] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3233–3241, 2017. 17, 29

[86] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. In *Proc. BMVC*, 2019.

[87] Z. Huasong, J. Chen, C. Shen, H. Zhang, J. Huang, and X. Hua. Self-adaptive neural module transformer for visual question answering. *IEEE Trans. Multimedia*, pages 1–1, 2020. 14

[88] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2623–2631, 2015. 14

[89] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 14

[90] Georgia Gkioxari, Ross B. Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. *arXiv preprint arXiv:1704.07333*.

[91] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *International Conference on Computer Vision (ICCV)*, pages 1068–1076, 2017. 39

[92] Wentong Liao, Shuai Lin, Bodo Rosenhahn, and Michael Ying Yang. Natural language guided visual relationship detection. *arXiv preprint arXiv:1711.06032*, 2017.

[93] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *International Conference on Computer Vision (ICCV)*, 2017. 17, 21

[94] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *Winter Conference on Applications of Computer Vision, (WACV)*, pages 1568–1576, 2018. 24

[95] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *European Conference on Computer Vision(ECCV)*. 29, 41, 42, 43, 44, 45

[96] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in Neural Information Processing Systems*, 2017. 14

[97] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016. 14

[98] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision(ECCV)*, pages 15–29, 2010.

[99] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rosenberg, and Fei-Fei Li. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1109, 2015.

[100] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1745–1752, 2011.

[101] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiaoou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7244–7253, 2017. 39

[102] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *European Conference on Computer Vision(ECCV)*, 2018. 33

[103] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *Advances in Neural Information Processing Systems*, pages 7211–7221, 2018. 14

[104] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016. 16

[105] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6985–6994, 2018. 17, 29

[106] Ranjay Krishna, Ines Chami, Michael S. Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6867–6876. 17

[107] Yikang Li, Wanli Ouyang, and Xiaogang Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. *arXiv preprint arXiv:1702.07191*, 2017. 21

[108] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4158–4166.

[109] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed M. Elgammal. Relationship proposal networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5226–5234, 2017. 21

[110] Brigit Schroeder, Subarna Tripathi, and Hanlin Tang. Triplet-aware scene graph embeddings. In *International Conference on Computer Vision (ICCV)*, pages 1783–1787, 2019. 17

[111] Seong Jae Hwang, Sathya N. Ravi, Zirui Tao, Hyunwoo J. Kim, Maxwell D. Collins, and Vikas Singh. Tensorize, factorize and regularize: Robust visual relationship learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 17, 21

[112] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Trans. Multimedia*, 22(6):1423–1432, 2020. 17

[113] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN. In *International Conference on Computer Vision (ICCV)*, pages 4243–4251, 2017. 18

[114] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015. 19, 20

[115] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015. 20, 34

[116] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *Proc. Conf. Human Language Technologies, 2015*, pages 153–163. doi: 10.3115/v1/n15-1016. 20

[117] Xu Yang, Hanwang Zhang, and Jianfei Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *European Conference on Computer Vision(ECCV)*, 2018. 21

[118] Ranjay Krishna, Ines Chami, Michael S. Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6867–6876, 2018. 21

[119] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 21, 36

[120] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018. 21

[121] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7239–7248, 2018. 21

[122] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Proc. Adv. Neural Inf. Process. Syst*, pages 558–568, 2018. 21, 40

[123] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2014. 21, 53

[124] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 457–468, 2016. 21, 22, 53

[125] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *International Conference on Computer Vision (ICCV)*, 2017. 21, 54

[126] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1571–1581, 2018. 22

[127] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6639–6648, 2019. 22, 53

[128] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. pages 289–297, 2016. 22, 59

[129] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 21, 22, 53, 54, 59, 69, 70, 71

[130] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *arXiv*, 2019. 21, 22, 54, 67, 68

[131] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[132] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[133] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[134] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 22

[135] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 5099–5110, 2019. 21, 70

[136] Badri N. Patro and Vinay P. Namboodiri. Differential attention for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7680–7688, 2018. 21

[137] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual QA. In *Advances in Neural Information Processing Systems*, pages 361–369, 2016. 22

[138] Qiang Sun and Yanwei Fu. Stacked self-attention networks for visual question answering. In *International Conference on Multimedia Retrieval (ICMR)*, pages 207–211, 2019.

[139] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik G. Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10264–10273, 2020. 22

[140] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. In *International Conference on Computer Vision (ICCV)*, pages 1300–1309, 2017. 22

[141] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6087–6096, 2018. 22

[142] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 22, 68

[143] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017. 22, 23, 53, 67, 70

[144] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980, 2018. 22, 70

[145] Yiyi Zhou, Rongrong Ji, Jinsong Su, Xiangming Li, and Xiaoshuai Sun. Free VQA models from knowledge inertia by pairwise inconformity learning. In *AAAI*, pages 9316–9323, 2019.

[146] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9687–9695, 2020. 23, 79

[147] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? *arXiv preprint arXiv:2006.05121*, 2020. 22, 23

[148] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *European Conference on Computer Vision(ECCV)*, pages 437–453, 2020. 23, 54, 55, 63

[149] Spencer Whitehead, Hui Wu, Yi Ren Fung, Heng Ji, Rogério Schmidt Feris, and Kate Saenko. Learning from lexical perturbations for consistent visual question answering. *arXiv preprint arXiv:2011.13406*, 2020. 23

[150] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. *arXiv preprint arXiv:1511.05099*, 2015. 23

[151] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. VQA-LOL: visual question answering under the lens of logic. In *European Conference on Computer Vision(ECCV)*, pages 379–396, 2020. 23

[152] Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv*, 2019. 23

[153] Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at VQA models: Interrogating VQA models with sub-questions. *arXiv preprint arXiv:2001.06927*, 2020. 23

[154] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *International Conference on Computer Vision (ICCV)*, pages 2712–2719, 2013. 24

[155] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *in TPAMI*, 36: 453–465, 2014.

[156] Stanislaw Antol, C. Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *European Conference on Computer Vision(ECCV)*, pages 401–416, 2014.

[157] Zhengming Ding, Ming Shao, and Yun Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[158] Chenrui Zhang, Xiaoqing Lyu, and Zhi Tang. TGG: transferable graph generation for zero-shot and few-shot learning. In *Proc. ACM Multimedia Conf.*, pages 1641–1649, 2019. 24, 36

[159] Xu Yang, Hanwang Zhang, and Jianfei Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *European Conference on Computer Vision(ECCV)*, 2017. 24

[160] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018. 25, 26, 60, 78

[161] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision(ECCV)*, 2020. 25

[162] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arxiv*, 2020. 25

[163] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020. 26, 61, 62, 65, 78

[164] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018. 26

[165] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[166] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision(ECCV)*, 2020.

[167] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 26, 60

[168] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 3285–3292, 2020. 26, 61, 70

[169] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 27, 62

[170] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016. 27, 62

[171] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, 2015. 31

[172] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *European Conference on Computer Vision(ECCV)*, 2018. 33

[173] Thomas Zenkel, Joern Wuebker, and John DeNero. Adding interpretable attention to neural translation models improves word alignment. In *Proc. IEEE Int. Conf. Comput. Lang.*, 2019. 34

[174] Yan Zhang, Jonathon S. Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations (ICLR)*, 2018. 38, 68

[175] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3107–3115, 2017. 39, 42

[176] Xiaodan Liang, Lisa Lee, and Eric P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4408–4417, 2017.

[177] Weilin Cong, William Wang, and Wang-Chien Lee. Scene graph generation via conditional random fields. *arXiv preprint arXiv:1811.08075*, 2018.

[178] Yaohui Zhu, Shuqiang Jiang, and Xiangyang Li. Visual relationship detection with object spatial distribution. In *International Conference on Multimedia and Expo, ICME*, pages 379–384, 2017. 39, 42

[179] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed M. Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. *arXiv preprint arXiv:1804.10660*, 2018. 39, 42

[180] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Rethinking visual relationships for high-level image understanding. *arXiv preprint arXiv:1902.00313*, 2019. 39

[181] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian D. Reid. Towards context-aware interaction recognition for visual relationship detection. In *International Conference on Computer Vision (ICCV)*, pages 589–598, 2017. 42

[182] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. Context-dependent diffusion network for visual relationship detection. *arXiv preprint arXiv:1809.06213*, 2018. 42

[183] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 43

[184] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016. 43

[185] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Scene dynamics: Counterfactual critic multi-agent training for scene graph generation. *arXiv preprint arXiv:1812.02347*, 2018. 44

[186] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. *arXiv preprint arXiv:1706.07365*, 2017. 45

[187] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the VQA challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 53, 54, 69

[188] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 53

[189] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2019. 53

[190] Yingzhen Li, John Bradshaw, and Yash Sharma. Are generative classifiers more robust to adversarial attacks? In *International Conference on Machine Learning (ICML)*, 2019. 53

[191] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek V. Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual LSTM networks. In *COLINGInternational Conference on Computational Linguistics, Proceedings of the Conference*, pages 2923–2934, 2016. 55

[192] Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. Paraphrase generation with deep reinforcement learning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing,*, pages 3865–3878, 2018. 55

[193] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. A deep generative framework for paraphrase generation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, pages 5149–5156, 2018. 55

[194] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. Paraphrasing revisited with neural machine translation. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *European Chapter of the Association for Computational Linguistics, EACL*, pages 881–893, 2017. 55

[195] John Wieting, Jonathan Mallinson, and Kevin Gimpel. Learning paraphrastic sentence embeddings from back-translated bitext. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 274–285, 2017.

[196] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL*, pages 116–121, 2018. 55, 61

[197] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *Conference on Learning Representations, ICLR*, 2018. 55

[198] Javid Ebrahimi, Daniel Lowd, and Dejing Dou. On adversarial examples for character-level neural machine translation. In *International Conference on Computational Linguistics, COLING*, pages 653–663, 2018. 55

[199] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 57, 58, 63, 73, 81

[200] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482, 2019. 57, 62, 81

[201] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP*, pages 39–57, 2017. 58

[202] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR)*, 2017. 58, 63, 78

[203] Qian Huang, Isay Katsman, Zeqi Gu, Horace He, Serge J. Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *International Conference on Computer Vision (ICCV)*, pages 4732–4741, 2019.

[204] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations (ICLR)*, 2020.

[205] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, pages 2206–2216, 2020.

[206] Zhengyu Zhao, Zhuoran Liu, and Martha A. Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1036–1045, 2020. 58

[207] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4312–4321, 2021. 58

[208] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. pages 489–500, 2018. 59

[209] Aritra Ghosh and Andrew S. Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2703–2708, 2021. 61

[210] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 68

[211] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 68

[212] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 68

[213] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NIPS*, 2020. 69

[214] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 878–892, 2020. 69, 70

[215] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. *arXiv preprint arXiv:1906.10169*, 2019. 70, 71

[216] Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine. In *NIPS*, pages 5901–5914, 2019. 70

[217] Rémi Cadène, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. MUREL: multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1989–1998, 2019. 70

[218] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10797–10806, 2020. 70, 71

[219] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, and Qi Tian. Loss-rescaling VQA: revisiting language prior problem from a class-imbalance view. *arXiv*, 2020. 70

[220] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. 2020. 70

[221] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020. 70

[222] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 708–714, 2021. 70

[223] Kushal Kafle, Mohammed A. Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of International Conference on Natural Language Generation, INLG*, pages 198–202, 2017. 78

[224] Judea Pearl. Causality: models, reasoning and inference. In *Springer*, volume 29, 2000. 79, 80

[225] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. In *Journal of the American Statistical Association*, pages 322–331, 2005. 79, 80

[226] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv*, 2020. 79

[227] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9847–9857. 79, 80

[228] Judea Pearl. Direct and indirect effects. In *Proceedings of the 17th conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001. 79

[229] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688, 1995. 79

[230] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015. 79

[231] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NIPS*, 2020. 79

[232] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3957–3966, 2021. 79

[233] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8659–8668, 2021. 81

[234] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing*, 2020. 81

[235] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *9th International Conference on Learning Representations, ICLR*, 2021.

[236] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[237] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020. 81