Analysis of circuit aging on accuracy degradation of deep neural network accelerator

Liu, Wenye; Chang, Chip-Hong

2019

Liu, W. & Chang, C.-H. (2019). Analysis of circuit aging on accuracy degradation of deep neural network accelerator. 2019 IEEE International Symposium on Circuits and Systems (ISCAS). doi:10.1109/ISCAS.2019.8702226

<https://hdl.handle.net/10356/136843>

<https://doi.org/10.1109/ISCAS.2019.8702226>

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at:
<https://doi.org/10.1109/ISCAS.2019.8702226>

Downloaded on 26 Apr 2023 13:52:28 SGT

Analysis of Circuit Aging on Accuracy Degradation of Deep Neural Network Accelerator

Wenye Liu and Chip-Hong Chang

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Email: wliu015@e.ntu.edu.sg, echchang@ntu.edu.sg

Abstract—Deep neural networks have achieved phenomenal successes in vision recognition tasks, which motivate the deployment of deep learning in portable and smart wearable devices. To overcome the fundamental challenges of power and resource limitation, application-specific integrated circuit accelerators have emerged to compact the model and use lower precision arithmetic to increase the throughput of computation with reduced power consumption. Although very high energy efficiency has been achieved by removing redundant weights, compressing data and even sacrificing timing margin, such trend in hardware acceleration that pushes the deep learning systems to the error threshold can be disastrous for the tasks they performed due to failure or degraded performance of circuit components. Concerned by the lack of attention on the evolving unreliability effects in artificial intelligent accelerators implemented by the continuously scaled CMOS technology, this paper is the first to evaluate the effect of circuit aging on performance degradation of deep learning accelerator. Our findings indicate that DNN system running at their peak throughput rate can experience up to 84% accuracy drop after a year of aging and the accumulation of errors aggravates with the depth of learning. It is also found that relaxation of throughput rate can slow down the loss of classification accuracy considerably.

I. INTRODUCTION

Deep learning brings significant breakthroughs in artificial intelligence, for vision recognition problems in particular [1]. Deep Neural Networks (DNNs) require intensive computations on each forward pass operation [2]. Performance improvements of problems solved by advanced DNN models are accomplished with significantly higher demand of computing power than standard signal processing algorithms and techniques for the same [3] [4]. To reduce the dependency on power hungry and costly programmable processors (CPU or GPU), dedicated hardware accelerators have been proposed [5] [6] [7] [8] for DNNs. All of the proposed deep learning accelerators trade various possible design margins and slacks of DNNs to achieve area and energy efficiency without compromising or even increasing the throughput. Data quantization and compression [9] [10] that utilize the redundancy in the parameterization of DNNs have been widely exploited to squeeze every last drop of energy out of the positive slacks from the critical paths at design time [8] [11].

Recent hardware implementations of DNN accelerator [5] [6] [7] [8] manifested the use of aggressive precision and voltage scaling, data compression, tolerance of timing error rate to improve the throughput and energy efficiency. These DNN accelerators are implemented on advanced CMOS technology, which are susceptible to device aging [12]. Threshold voltage shift of stressed CMOS transistors reduces switching speed of logic gates, which in turn limits the maximum operating frequency of the integrated circuit. Two main types of aging mechanism are hot carrier injection (HCI) and negative bias temperature instability (NBTI), which have been well-studied

to be the major causes of performance unreliability of NMOS and PMOS devices, respectively [12] [13]. The progressive timing degradation on combinational circuits due to these two aging phenomena has been demonstrated in [14]. Unfortunately, DNN accelerators rely heavily on MAC units which consist of combinational circuit multipliers, adders and accumulators. It is thus reasonable to conjecture that accuracy of DNN computations will reduce with time for these modern energy-efficient accelerators that are designed to work on the verge of the timing margin.

So far no analysis nor investigation has been targeted on the circuit reliability of aggressively scaled energy-quality accelerator for DNN computations. Ignorance of the extent and rate of influence of aging induced errors on the eventual inference results over the lifespan of the DNN applications can lead to catastrophe. This paper attempts to evaluate how sensitive does the computation result of each convolution layer affected by aging, how rapidly the aging induced errors propagate and accumulate down the layers and the extent of degradation in inference accuracy over time. Our study focuses on the MAC array of DNN hardware accelerator, which is arguably the most timing critical stage and the most common target of energy-quality scaling [6] [8]. We choose the popular and classic DNN, AlexNet [3], for our experiment. Based on our evaluation, we confirm the prominent accuracy loss caused by aging effect. The classification accuracy can experience as large as 80% degradation over a year of operation. This preliminary finding is important to raise the alarm on potential life-threatening failures for DNN applications in autonomous car and biomedical applications such as critical disease diagnosis.

The rest of the paper is organized as follows. Background on deep learning accelerator and two major CMOS transistor aging mechanisms are presented in Section II. The design of aging evaluation experiments is described in Section III. Section IV discusses the simulation results. Finally, the conclusion is given in Section V.

II. PRELIMINARIES

A. Deep Neural Network Accelerator

State-of-the-art DNN algorithms [3] are compute-intensive. Over 90% of the computations are contributed by the convolution operations. A convolutional layer extracts feature from the input feature maps (fmaps) by high-dimensional convolutions and generates output feature maps in a higher abstraction level. The computational complexity of output fmaps depends on the filter size and the dimensions of output fmaps. Fig. 1 illustrates the computations of a convolution layer. The upper part of Fig 1 shows the single step convolution. It requires element-wise multiplications between each element of the receptive field and the corresponding filter. All the elements of the generated

product array are accumulated to generate an element y of the output fmaps as follows,

$$y = \sum_{i=1}^K \sum_{j=1}^K \sum_{c=1}^C w_{i,j,c} \times x_{i,j,c} + bias \quad (1)$$

where w is the filter weight and x is the input element.

Thus, the number of MACs required to compute each output element is determined by the size of the filter ($K \times K \times C$). The bottom part of Fig 1 shows the complete computation of the convolution layer where a full set of filters is used to extract the features by sliding through the whole input fmaps. The computation is given by:

$$Z = f([Y_1, Y_2, \dots, Y_{nC}]) \quad (2)$$

where $Y = [y_{11}, y_{12}, \dots, y_{nW \times nH}]$ and $f(\cdot)$ is the activation function (normally a rectified linear unit (ReLU) in modern Convolutional Neural Networks (CNNs)). Consequently, the number of MACs required in one convolution layer is $K \times K \times C \times nW \times nH \times nC$.

Fig. 1. Convolution computation. Upper part: a single-step convolution of a receptive field and one filter. Bottom part: Entire convolution of a full set of filters sliding through the whole input fmaps

Hardware acceleration of DNNs involves both algorithmic and circuit-level optimizations [2]. Data quantization and voltage scaling are two primary methods used to improve the energy efficiency for dedicated DNN accelerator design. The energy consumption and circuit area can be reduced with smaller network size and shorter bit width operators. Ristretto [10], which is a DNN quantization tool, can be used to obtain an 8-bit fixed point implementation of AlexNet [3] within 1% accuracy loss compared with its 32-bit floating point counterpart. An alternative method to reduce the network size is by network compression [9]. Quantizing the data from 32 bits to 8 bits can save three quarters energy in memory access, and an 8-bit fixed point multiplier consumes 18.5x less energy than a 32-bit floating point multiplier [2]. In addition to uniform quantization, layer-wise quantization can also be applied by using precision scalable arithmetic circuit to adapt to different bit widths for different layers [6] [15]. For example, a 16-bit scalable multiplier can be reconfigured to two 8-bit multipliers or four 4-bit multipliers. The positive slack caused by the reduced precision operands can be traded for dynamic power efficiency through scaling down the supply voltage.

More aggressive voltage scaling has been proposed by time-borrowing in datapath logic [8].

Approximate computing or voltage-accuracy scaling for energy efficiency pushes the circuit to the error threshold of computation. Error compensation of DNNs under voltage over-scaling have been explored in [8] [16]. The error performance was only demonstrated on shallow network for simple task (e.g., 10 Digit classification). What has been omitted in most of these studies is the effect of aging. The aging induced degradation of the MOS transistor's threshold voltage V_{th} leads to decreased drive current and increased propagation delay of circuit paths [13] [17]. Although timing degradation due to aged arithmetic circuitry has relatively low computation error rate, the intensive accumulations in (1) of modern DNNs increase the chance of error produced on each output element. Due to the regularization in training phase, the weight distribution of DNN model is sparse and close to zero [18]. This intrinsic property of DNNs leads to the small values in a majority of intermediate products. Due to the critical path, the higher order bits of these intermediate products are more frequently affected by the timing violation, resulting in large magnitude errors. In two's complement number representation, an error in the most significant bit will also cause a large magnitude change. As a consequence, the error outputs overwhelm the intermediate products during accumulations.

B. Aging of CMOS transistor

Two major drift-related transistor aging mechanisms have been identified in nanoscale VLSI circuits. Negative-bias temperature instability (NBTI) is a physical phenomenon that affects PMOS devices most while hot carrier injection (HCI) impacts prominently NMOS devices.

Fig. 2. MOS aging mechanisms. (a) HCI in NMOS device. (b) NBTI in PMOS device

Fig 2(a) shows the phenomenon of HCI in NMOS device, where the particles under strong electric field between the source and drain terminals are accelerated and become hot carriers with high kinetic energy [12] [19]. These hot carriers injected into the gate oxide give rise to the generation of interface states. These defects results in an increase of V_{th} . In early CMOS device, HCI suffers from high supply voltage. With the reduction of supply voltage and introduction of high-k metal gate (HKMG) technology, the effect of HCI is mitigated in advanced technology node [12]. NBTI effect in PMOS device is shown in Fig 2(b). This aging mechanism can be described by reaction-diffusion (R-D) model [13] [20]. During the reaction phase, the silicon-hydrogen bond located at oxide/substrate interface will break due to holes induced by negative bias stress. This dissociation of the chemical bond leads to interface charges increases the threshold voltage. At the diffusion stage, the released hydrogen atoms can combine to generate the hydrogen molecule. Due to the gradient of H_2 density, hydrogen species diffuse towards the gate terminal and

are trapped permanently [19]. If the negative bias is removed, the hydrogen species diffuse back and recover the Si+ bond [14]. This reduction of interface charges due to the recovery of Si+ bond is called the annealing process of NBTI. With the sustained scaling of device dimensions, NBTI becomes the dominant limiting factor of speed degradation [21].

Other than voltage overscaling, aging effect has different impact on individual instances even for the same type of gate. Timing degradation on specific instance is determined by its own stress condition (i.e., signal switching activity). Aging effect is inevitable after the chips have been operating long enough in the field. For deployed chips, timing errors caused by temporary overrated voltage can be recovered by returning back to nominal voltage. However, reinstating aged transistor to initial state is impossible. Fresh chips passed production tests may still suffer from the risk of computation errors after deployment, especially for the compute-intensive tasks performed by DNN accelerator.

III. DESIGN OF EXPERIMENTS

Fig 3 presents a typical deep learning accelerator architecture that is evaluated in this study. The data path circuitry consists of an array of 256 (16 rows $\times$ 16 columns) MAC units. The buffered inputs of features and weights will be broadcast to the MAC units. All the MAC units from the same row will share the same weight and one input feature will be broadcast to all the units in the same column. According to the data flow of [6] [22], the input features are fetched on the sliding sequence of pixels from one input channel to different channels. The input weights utilize parallel computations from different filters to eliminate partial sum movement between each MAC unit and avoid the requirement of large adder tree for summing all the partial sums in one clock cycle [22]. Separate power domains are commonly applied on the MAC array and other circuits (memory and control logics), as voltage scaling can be used to increase energy efficiency during MAC operations [6] [8].

Fig. 3. A typical deep learning accelerator with separate voltage domains at MAC array

We synthesis and construct the 256 MAC array in 45nm characterized standard cell library. Other portions of the deep learning accelerator are integrated in higher level Verilog test-bench. The simulation is performed on system level of the accelerator while timing information is specifically annotated in gate level netlist of the MAC array under aging criteria.

The details of the experimental setup and evaluation flow are shown in Fig 4. After synthesis, the gate level netlist of the MAC unit can be obtained. This gate level netlist is transformed into fresh (non-aging) transistor level netlist

Fig. 4. Simulation flow for aging evaluation on MAC array

for aging simulation. Before aging, all the transistors in the netlist share the uniform NMOS/PMOS device model. We use Cadence RelXpert simulator, a commercial aging simulation tool with built-in HCI and NBTI model, to simulate the aging of MAC unit over days, months and years of operation. As it is impractical to simulate the entire DNN computation in transistor level, we capture a section of real input data as stimuli to represent the workload of a DNN application for aging simulation. The output of RelXpert is an aged transistor level netlist. The circuit structure of this aged netlist is identical as the fresh netlist, while each transistor of the aged netlist has its own degraded MOS model based on various stress conditions from different instances. Once the aged netlist is obtained, we can characterize the delay of every aged instances through their individual dedicated MOS models. In this case, each instance will be treated as a specific circuit with independent timing information. The final gate-level simulation is conducted on the entire MAC array with aged timing library back annotated. Although the gate-level simulation saves significant run time compared with the transistor-level simulation, it is still infeasible to evaluate the whole imagenet [23] validation dataset. Therefore, we shuffle the dataset and evaluate one batch of images (50 images) on the AlexNet.

IV. SIMULATION RESULTS AND ANALYSIS

In this section we present the results and analysis of aging effect on deep learning accelerator. We perform two sets of aging evaluation. One is on system level of the DNN models across all the convolution layers. Another is layer-wise evaluation under various aging time. The nominal frequency is set to 920MHz based on the critical path reported from the synthesis. The results are normalized based on the top5 accuracy.

Fig 5(a) shows the accuracy drop of Alexnet after 30 days and 1 year of simulated circuit aging. The figure contains two groups of lines. The dash lines correspond to aging results of individual layer for which the input of the layer is accurate. The solid lines include the propagation of aging-induced errors where the input of the current aged layer is the aged results from the previous layer. For aging on individual layer, minor accuracy drops are observed on the first four convolution layers. 33% and 46% drops of accuracy are found on the 5th convolution layer after 30 days and 1 year stresses, respectively. Consider the effect of error propagation, the entire convolution computation suffers more severe accuracy degradation. The

solid lines indicate that the accuracy loss of each aged convolution layer aggregates as the errors propagate and accumulate down the layers. Consequently, the final layer has 60% accuracy drop after 30 days and 84% accuracy drop after 1 year aging. The solid lines are emblematic of the actual aged classifier performances as the inputs will be computed through all the layers.

Fig. 5. Accuracy drop of Alexnet due to circuit aging. (a) accuracy degradation at each layer due to aging-induced errors of independent layer and with aging errors propagated from previous layers for 30 days and 1 year of aging. (b) accuracy degradation after 1 year aging by lowering the operation frequency.

Despite varying degrees of sensitivity to aging effect are observed in Fig 5(a) among all the layers, the rapid increase in accuracy loss at the final layer due to accumulated errors from previously aged layers suggests that more computation layers of DNN model lead to worse classification performance as its devices age, which offsets the gain in accuracy of deeper network architecture after certain period of operation.

Fig 5(b) presents the aging results of operating the DNN at a lower frequency. Compared with operating at the nominal frequency, relaxing the frequency by 10MHz to 910MHz can slow down the aging effect and reduce the accuracy drop. By further relaxing the operating frequency to 900MHz, there is no discernible performance degradation of the DNN after 1 year of operation.

Fig. 6. Aging evaluation on specific layers for 1 day, 7 days, 1 month, 3 months, 1 year, 3 years, 5 years and 10 years

Fig 6(a) shows aging results of conv1 and conv5 layers of AlexNet for different aging time. The conv1 layer is insensitive to aging effect until after 10 years of stress. In contrast, the accuracy of conv5 layer drops monotonically with stress until 1 year and begins to recover part of the lost accuracy after 3 years of aging. This self-healing effect is caused by the flipping of multiple internal signals, which increases the probability of coincidentally correcting the output bit errors as illustrated by an example in Fig 7. The effect of circuit aging will lead to timing degradation on individual gates and may results in bit flip. Considered the internal signal n8 and n9, after one year aging, n9 may have a high probability of flipping from a correct output state (assume logic 0) to an erroneous state (logic 1) due to the longer path delay while n8 with shorter path delay may remain in the correct output state (assuming logic 1). Thus the output bit n11 will flip to the wrong state (i.e., from logic 1 to logic 0). However, with longer stress (more than 3 years), n8 may also flip from the correct output to logic 0. Hence n11 will have higher probability to stay at the correct output even though both its inputs are incorrect. This phenomenon can mitigate some of the computation errors caused by aging but only to a limited extent. Before this phenomenon takes effect, the performance of system has already been severely degraded.

Fig. 7. An example of multiple internal bit flips that restore the "correct" output

Adding timing guard band on deep learning accelerator can reduce the aging effect as shown in Fig 6(b). Relaxing the operation to 910MHz can postpone the accuracy degradation but will still suffer from more than 33% loss of accuracy after 3 months. With more guard band (900MHz), the accelerator can maintain the performance for 3 years. Adding guard band will defeat the original purpose of aggressively increasing or preserving the high throughput by eliminating hardware redundancy for improved energy efficiency.

V. CONCLUSION

In this paper, we design experiments to evaluate if circuit aging will lead to conspicuous performance degradation on deep learning accelerators that have been designed with aggressive energy-quality scaling. It is found that device aging, which is an inevitable issue of the dense integration of nanoscale transistors, can result in 84% drop on classification accuracy after the DNN has been deployed for one year. Our study also indicates that different layers of DNN have different sensitivities to aging effect, but accuracy degradation of DNN increases as the data passes through more computation layers. This work alerts the potential catastrophe of artificial intelligence applications due to fatigue deep learning accelerator. Our future work will explore aging aware energy-quality scaling of deep learning accelerator.

VI. ACKNOWLEDGEMENT

This research is supported by the Singapore Ministry of Education Tier 2 Grant MOE2015-T2-2-013 and Tier 1 grant MOE2018-T1-001-131 (RG87/18).

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [2] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Information Processing Syst. (NIPS)*, Lake Tahoe, Nevada, USA, Dec. 2012, pp. 1097–1105.
- [4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," presented at Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, May 2014. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [5] Y. H. Chen, T. Krishna, J. Emer, and V. Sze, "14.5 eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Jan. 2016, pp. 262–263.
- [6] B. Moons and M. Verhelst, "A 0.3-2.6 TOPS/W precision-scalable processor for real-time large-scale convnets," in *Proc. IEEE Symp. VLSI Circuits (VLSI-Circuits)*, Honolulu, Hawaii, USA, Jun. 2016, pp. 1–2.
- [7] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient inference engine on compressed deep neural network," in *Proc. 43rd Int. Symp. Comput. Arch. (ISCA)*, Seoul, Republic of Korea, Jun. 2016, pp. 243–254.
- [8] P. N. Whatmough, S. K. Lee, H. Lee, S. Rama, D. Brooks, and G. Wei, "14.3 a 28nm SoC with a 1.2ghz 568nj/prediction sparse deep-neural-network engine with >0.1 timing error rate tolerance for IoT applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Feb. 2017, pp. 242–243.
- [9] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," presented at Int. Conf. Learning Representations (ICLR), Caribe Hilton, San Juan, Puerto Rico, May 2016. [Online]. Available: <http://arxiv.org/abs/1510.00149>
- [10] P. Gysel, M. Motamedi, and S. Ghiasi, "Hardware-oriented approximation of convolutional neural networks," presented at Int. Conf. Learning Representations (ICLR), Caribe Hilton, San Juan, Puerto Rico, May 2016. [Online]. Available: <http://arxiv.org/abs/1604.03168>
- [11] B. Moons and M. Verhelst, "An energy-efficient precision-scalable convnet processor in 40-nm cmos," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 903–914, Apr. 2017.
- [12] E. Maricau and G. Gielen, *Analog IC reliability in nanometer CMOS*. New York: Springer Science & Business Media, 2013.
- [13] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao, "Compact modeling and simulation of circuit reliability for 65-nm CMOS technology," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 4, pp. 509–517, Dec. 2007.
- [14] W. Wang, S. Yang, S. Bhardwaj, S. Vrudhula, F. Liu, and Y. Cao, "The impact of NBTI effect on combinational circuit: Modeling, simulation, and analysis," *IEEE Trans. VLSI Syst.*, vol. 18, no. 2, pp. 173–183, Feb. 2010.
- [15] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Feb. 2017, pp. 246–247.
- [16] Y. Lin and J. R. Cavallaro, "Energy-efficient convolutional neural networks via statistical error compensated near threshold computing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Florence, Italy, May 2018, pp. 1–5.
- [17] C. Q. Liu, Y. Cao, and C. H. Chang, "ACRO-PUF: A low-power, reliable and aging-resilient current starved inverter-based ring oscillator physical unclonable function," *IEEE Trans. Circuits Syst. I*, vol. 64, no. 12, pp. 3138–3149, Dec. 2017.
- [18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, Massachusetts: MIT Press, 2016.
- [19] H. Dadgour and K. Banerjee, "Aging-resilient design of pipelined architectures using novel detection and correction circuits," in *Proc. Design, Autom. Test Eur. Conf. (DATE)*, Dresden, Germany, Mar. 2010, pp. 244–249.
- [20] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "An analytical model for negative bias temperature instability," in *Proc. IEEE/ACM Int. Conf. Comput.-aided Design (ICCAD)*, New York, NY, USA, 2006, pp. 493–496.
- [21] N. Kimizuka, T. Yamamoto, T. Mogami, K. Yamaguchi, K. Imai, and T. Horiuchi, "The impact of bias temperature instability for direct-tunneling ultra-thin gate oxide on MOSFET scaling," in *Proc. IEEE Symp. VLSI Technology*, Kyoto, Japan, Jun. 1999, pp. 73–74.
- [22] Y. Ma, Y. Cao, S. Vrudhula, and J. Seo, "Optimizing the convolution operation to accelerate deep neural networks on FPGA," *IEEE Trans. VLSI Syst.*, vol. 26, no. 7, pp. 1354–1367, Jul. 2018.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Comput. Vision Pattern Recogn. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255.