XNOR-Net: Classificação Utilizando redes Neurais Binárias

Lucas Thom Ramos
Departamento do Informática
UFES
Vitória, Brasil
lucas.ramos@motoratech.com

Cézar A. G. Passamani
Departamento do Informática
UFES
Vitória, Brasil
cezar.gobbo@motoratech.com

Resumo*—Este trabalho consiste em executar experimentos utilizando *XNOR-Networks, cuja proposta é aproximar os valores de input e pesos para valores binários garantir uma economia de 32x em memória, além de serem 58x mais rápidas, pois as convoluções são baseadas em operações binárias que estão implementadas nas CPUs.

Palavras-chave—classificação; redes neurais binárias; redes neurais; redes neurais embarcadas

I. Introdução

Redes Neurais Profundas (RNP) têm trazido diversos domínios de aplicação, como visão computasional e reconhecimento de fala. Em visão computacional, um tipo específico de RNP chamada Rede Neural Convoloucional (RNC), que tem apresentado resultados no estado da arte para classificação de objetos [2,3,4,5] e detecção [6,7,8]. As Redes Neurais Convolucionais estão trazendo melhorias em diversas aplicações do mundo real. Apesar do principal objetivo da implementação das XNOR-Nets serem os ganhos em armazenamento e processamento utilizando CPU ao invéz de GPU, neste trabalho a implementação não foi feita substituindo as operações de ponto flutuante. Para fins de experimentação, os pesos ainda são representados por pontos flutuantes, porém possuem valores binários atribuídos (0 ou 1), assim como os valores de input. O objetivo deste trabalho é avaliar a acurácia das redes sem a implementação das operações binárias, como o próprio autor original o fez [1].

II. Trabalhos Correlatos

Trabalhos que tratam de redes neurais binárias vem sido desenvolvidos [9,10,11,13]. Em 9, os autores propõem um framework para construção e treinamento deste tipo de rede, que inclusive pode ter a precisão (número de bits) definida para cada peso das camadas da rede. Em 10, é proposto um método otimizado para treinamento e em 11 é proposto embarcar as redes binárias em *chips* embarcados (FPGA). Em 13, um processo de binarização para redes treinadas com precisão padrão de 32 bits.

III. Metodologia

A. Rede

A arquitetura escolhida para a realização dos experimentos foi a AlexNet [12].

B. Experimentos

Os experimentos levaram em consideração a acurácia do treinamento, comparados com resultado da mesma redes (AlexNet), em sua versão não bínária. O dataset utilizado foi o Imagenet de larga escala (1M+ imagens). Ambas treinadas e testadas utilizando o Torch 7.0.

O experimento foi realizado utilizando uma GPU Nvidia Titan X.

IV. Resultados

Os resultados obtidos podem ser Visualizados na Tabela 1. Podemos observar que a diferença na acurácia não é tão drástica, uma vez que se passa de uma precisão de 32 bits para uma precisão de 1 bit apenas

Tabela 1

Rede	**Acurácia de Classificação (%)**		
	Top-1	Top-5	GPU
AlexNet-XNOR	**43.112**	**70.341**	Sim
AlexNet	54.623	78.854	Sim

Bibliografia

[1] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105.

[3] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1–9.

[5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR (2015) .

[6] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587.

[7] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1440–1448.

[8] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. (2015) 91–99.3.

[9] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, Yuheng Zou: DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients (2016).

[10] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, Yoshua Bengio: Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1 (2016).

[11] Yixing Li, Zichuan Liu, Kai Xu, Hao Yu, Fengbo Ren: A GPU-Outperforming FPGA Accelerator Architecture for Binary Convolutional Neural Networks (2017).

[12] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks (2012).

[13] Courbariaux, M., Bengio, Y., David, J.P.: Training deep neural networks with low precision multiplications. arXiv preprint arXiv:1412.7024 (2014)