Using Model Distillation for On-Demand Fine-Tuning of Deep Models

Milad Alizadeh and Nicholas D. Lane

University of Oxford

In recent years Deep Neural Networks (DNNs) have found great success across various domains such as computer vision, speech recognition and machine translation [7, 5, 1]. However, best-performing models are often compute and memory hungry and deploying them to mobile devices where there are stringent requirements in terms of available compute power, memory footprint, latency and energy consumption is challenging. Emergence of computing infrastructures such as cloudlet and edge computing promises new opportunities for on-demand customised fine-tuning of deep models before deployment to mobile devices. In this study we investigate how this limited compute capacity can influence the way we design and derive models that can be quickly adopted for mobile deployment with predictable consequences.

We show how distillation techniques [6] that are often used to shrink the size of model can instead be used to derive a mobile-friendly variants of the same architecture. Examples of such solutions are networks with quantised parameters and activations; binary networks, ternary networks and their sparse variants are examples of quantised architectures with suitable properties for mobile platforms. Initial attempts [4] to derive these networks through post-training steps suffered from significant loss of accuracy but it has been shown [2, 3, 8] that much better performance can be achieved by training binary networks *end-to-end*. The optimisation process however is biased, noisy and very slow. In this work we show how one can apply knowledge distillation techniques where a full-precision network acts as the teacher to derive a binary/ternary network as the student network in small number of steps. For CIFAR-10, our approach reaches 87% accuracy in just 4 epochs compared to 80% accuracy in the 30 epochs when trained with conventional methods.

References

[1] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Courbariaux, M., Bengio, Y., and David, J.-P. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems* (2015), pp. 3123–3131.

[3] Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830* (2016).

[4] Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[5] Hinton, G., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine 29*, 6 (2012), 82–97.

[6] Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

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

[8] Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision* (2016), Springer, pp. 525–542.