Smart(er) Machine Learning for Practitioners

Prabhu Pradhan*
Feroze Gandhi Institute of Engineering and Technology (FGIET),
Dr. APJ Abdul Kalam Technical University (AKTU),
Raebareli, Uttar Pradesh, India - 229316
prabhupradhan@pm.me

Abstract

Convolutional Neural Networks (CNNs) and Recurrent Neural Network (RNNs) and their variants in many conditions have produced super-human performance (i.e. better than human experts). Although, existing deep network models are incompatible with low power devices or mission-critical applications with crucial latency requirements due to either high computational cost or memory storage. There is a growing focus on model compression and acceleration techniques with very promising results. However, less effort has been made in making the architectural improvements modular. In this paper, we discuss about the techniques for efficient designs including some structural modifications. A brief section on the societal impact of these methods especially in the developing regions of the world is presented. Finally, we conclude this paper by further discussion and proposing possible challenges in these areas.

1 Introduction

Currently, the driving intuition in network designs are for beating records in very specific tasks regardless of modularity or efficiency. State-of-the-art models delivering superb performance often contain billions of parameters and such requirements are a critical hindrance in low-resource applications. While model quality has been shown to scale with model and data-set size (Hestness et al., 2017), the heavy resources required to train them can be prohibitive, especially in regions with low research budgets. Various methodologies have been used to reduce the architectural complexity of such models (Simard, Steinkraus, and Platt, 2003). Residual Networks (ResNets) (He et al., 2016) and SqueezeNet (Iandola et al., 2017) achieve better results despite very small parameter count.

2 Methods for Compact Models

In case of mobile devices (ex. Robots, IoT/Edge) which have limited computational capabilities, resource intensive deep neural networks cannot be readily applied, experiments (Sarkar, Pradhan, and Ghose, 2019), (Lane, Georgiev, and Qendro, 2015), (Haffari et al., 2018) revealed a major hurdle to wide spread use. We summarize four types of compression methods (Pruning, Quantization, Knowledge Distillation and Tensor Vectorization/Low-Rank Factorization) in Table 1 (2).

*This work was done as part of an AI Research Intern at GCDSL, Indian Institute of Science (IISc Bangalore)

6th CiML Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

Technique	Pruning	Quantization	Distillation	Vectorization
Description	Sharing or removing redundant parameters	Reducing the model's numerical precision	Learning smaller models from big ones (mimic)	Approximate a weight matrix by sum minimization
Application*	Conv/FC-layer. (on synapses and/or neurons)	Post-Training (or quantization-aware training)	Conv & FC-layer (across network)	Conv/FC-layer
Pros	Reduces size, complexity and over-fitting	Faster Inference (Even better when Activations are quantized too)	Significant reduction in computational-cost and size	High compression and speed-up
Cons	Longer Training, more hyper-tuning	Hardware dependant benefits	Overfitting, works only for (Softmax) classifications	Rigorous Retraining, Small ranks may hurt model

Table 1: Summary of common model-compression techniques
(**Conv= Convolutional Layer, FC=Fully-Connected*)

3 Structure-based improvements for Applied ML

It is well known in practice that Data Augmentation can be very beneficial for model performance. Cyclic Learning Rates have been shown to enhance training as well. Moreover, it has been shown empirically that DNNs can tolerate high levels of sparsity (Narang et al., 2017), and this property has been leveraged to significantly reduce the deployment cost (van den Oord et al., 2016). Although, here we will be shedding light at other emerging techniques.[2]

Swish
Swish (Ramachandran, Zoph, and Le, 2018) is an activation function:

$$f(x) = x\,\sigma(\beta x)$$

where σ(z) = $\frac{1}{1+exp(-z)}$. Also, β can be defined as a constant (generally, 1) or a trainable parameter (especially helpful when encountering many dead ReLUs). Its advantages are observable in deeper networks since it better handles vanishing gradients.

Regularizer: Swapout
Its (Singh, Hoiem, and Forsyth, 2016) a stochastic training method that shows stable improvements using efficient parameter utilization. It can be seen as a clever merge of the two regularization techniques i.e. dropout and stochastic depth, outperforming both in stand-alone comparisons. Also, linear decay of parameters (less dropping on early layers, more on later ones) significantly improves its results. Relatively shallow Swapout networks give similar performance to extremely deep ResNets.

Octave Convolution (OctConv)
There are a lot of spatial redundancies in CNN frameworks, thus OctConv (Chen et al., 2019) enhances efficiency by leveraging low/high-frequency features independently. It is a plug-and-play, orthogonal unit to substitute regular convolutions (2D and 3D) without any modifications to the network structure. To the best of our knowledge, OctConv has been used to stabilize GAN training (Durall, Pfreundt, and Keuper, 2019) and also reacts well with model compression (Zhou et al., 2019).

4 ML Efficiency for Good

With all the modern technology available to humankind, humble farmers across the world are still at the mercy of the environment for their livelihood. Remote Health Diagnostics in parts of developing and under-developed countries still hasn't hit a critical point. It is quite evident that these problems already have baseline solutions using AI and Machine Learning, although from a realistic perspective these are still out of reach from the populace. Neural Networks which *won't be a luxury* to deploy can empower local officials and regional scientists. AI can thus be truly transformative if it receives contributions from all over the globe, its significance as a tool depends on the adoption scale.

[2]Extended Version available at- https://openreview.net/forum?id=H1l7mN6AwH

5 Discussion

As evident in recent literature, there's a growing interest in Efficient/Compact Networks. It'll be exciting to see challenges invested in real-world applications. Some topic proposals are:

- Deployable Efficient Networks (Ex. Remote Health Diagnostics). (Kouw et al., 2017)
- Cloud-less (on-device) execution of time-critical tasks (ex. Robotics, Edge devices).
- Model-Agnostic (modular) improvements and Best Practices in Neural Networks.
- Interpretable and Explainable Compact Network Architectures. (Zhou et al., 2017)

Acknowledgments

Author acknowledges EPSRC-GCRF for partial funding (EP/P02839X/1), DeepMind for TFRC TPU credit award, and CiML Organizers for complimentary registration. PP also expresses special thanks to Debasish Ghose (IISc) for insightful feedback, and R.P. Sharma (FGIET) for guidance and support.

References

Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; and Feng, J. 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proc.. of International Conference on Computer Vision*, volume abs/1904.05049.

Durall, R.; Pfreundt, F.-J.; and Keuper, J. 2019. Stabilizing gans with octave convolutions. *ArXiv* abs/1905.12534.

Haffari, R.; Cherry, C.; Foster, G.; Khadivi, S.; and Salehi, B. 2018. Workshop on deep learning approaches for low-resource nlp. Melbourne, Australia: Association for Computational Linguistics (ACL).

He, K.; Zhang, X.; Ren, S.; and Sun., J. 2016. Deep residual learning for image recognition. In *Proc.. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G. F.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; and Zhou, Y. 2017. Deep learning scaling is predictable, empirically. *CoRR* abs/1712.00409.

Iandola, F. N.; Moskewicz, M. W.; Ashraf, K.; Han, S.; Dally, W. J.; and Keutzer, K. 2017. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *ArXiv* abs/1602.07360.

Kouw, W. M.; Loog, M.; Bartels, L. W.; and Mendrik, A. 2017. Mr acquisition-invariant representation learning. *ArXiv* abs/1709.07944.

Lane, N. D.; Georgiev, P.; and Qendro, L. 2015. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *UbiComp*.

Narang, S.; Diamos, G.; Sengupta, S.; and Elsen, E. 2017. Exploring sparsity in recurrent neural networks. In *International Conference on Learning Representations, (ICLR)*, volume abs/1704.05119.

Ramachandran, P.; Zoph, B.; and Le, Q. V. 2018. Searching for activation functions. In *(Workshop Track) International Conference on Learning Representations, (ICLR)*, volume abs/1710.05941.

Sarkar, M.; Pradhan, P.; and Ghose, D. 2019. Planning robot motion using deep visual prediction. In *7th Workshop on Planning & Robotics (PlanRob), ICAPS 2019*. Berkeley, USA: CoRR, abs/1906.10182.

Simard, P. Y.; Steinkraus, D.; and Platt, J. C. 2003. Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* 958–963.

Singh, S.; Hoiem, D.; and Forsyth, D. 2016. Swapout: Learning an ensemble of deep architectures. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*, 28–36. Curran Associates, Inc.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A. W.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. In *SSW*.

Zhou, B.; Bau, D.; Oliva, A.; and Torralba, A. 2017. Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41:2131–2145.

Zhou, D.; Jin, X.; Wang, K.; Yang, J.; and Feng, J. 2019. Deep model compression via filter auto-sampling. *ArXiv* abs/1907.05642.