Published in Towards Data Science | Follow

Prakhar Ganesh | Follow
Aug 12, 2019 · 4 min read

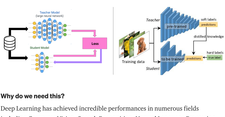
Knowledge Distillation : Simplified

Take a peek into the world of Teacher Student networks

What is Knowledge Distillation?

Neural models in recent years have been successful in almost every field including extremely complex problem statements. However, these models are huge in size, with millions (and billions) of parameters, and thus cannot be deployed on edge devices.

Knowledge distillation refers to the idea of model compression by teaching a smaller network, step-by step, exactly what to do using a bigger already trained network. The 'soft labels' refer to the output feature maps by the bigger network after every convolution layer. The smaller network is then trained to learn the exact behavior of the bigger network by trying to replicate it's outputs at every level (not just the final loss).



Why do we need this?

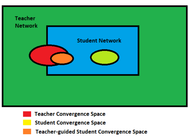
Deep Learning has achieved incredible performances in numerous fields including Computer Vision, Speech Recognition, Natural Language Processing etc. However, most of these models are too expensive computationally to run on devices like mobile phones or embedded devices. To understand more about the need of Model Compression and common techniques involved, visit the blog below.

Deep Learning — Model Optimization and Compression: Simplified
Take a peek into the domain of compression, pruning and quantization of state-of-the-art Machine Learning models
towardsdatascience.com

How is this different from training a model from scratch?

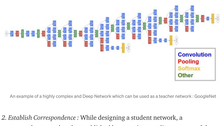
Obviously, with more complex models, the theoretical search space in larger than that of a smaller network. However, if we assume that the same (or even similar) convergence can be achieved using a smaller network, then the convergence space of the Teacher Network should overlap with the solution space of the student network.

Unfortunately, that alone does not guarantee converge for the student network at the same location. The student network can have a convergence which might be hugely different from that of the teacher network. However, if the student network is guided to replicate the behavior of the teacher network (which has already searched through a bigger solution space), it is expected to have its convergence space overlapping with the original Teacher Network convergence space.



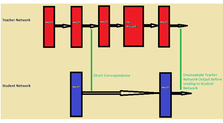
Teacher Student networks — How do they exactly work?

1. *Train the Teacher Network* : The highly complex teacher network is first trained separately using the complete dataset. This step requires high computational performance and thus can only be done offline (on high performing GPUs).



An example of a highly complex and Deep Network which can be used as a teacher network - GoogleNet

2. *Establish Correspondence* : While designing a student network, a correspondence needs to be established between intermediate outputs of the student network and the teacher network. This correspondence can involve directly passing the output of a layer in the teacher network to the student network, or performing some data augmentation before passing it to the student network.



An example of establishing correspondence

3. *Forward Pass through the Teacher network* : Pass the data through the teacher network to get all intermediate outputs and then apply data augmentation (if any) to the same.

4. *Backpropagation through the Student Network* : Now use the outputs from the teacher network and the correspondence relation to backpropagate error in the student network, so that the student network can learn to replicate the behavior of the teacher network.

What's next?

There have been a lot of new modification suggested to the traditional student teacher described above, like introducing multiple teacher (i.e. converting an ensemble into a single network), introducing a teaching assistant (the teacher first teaches the TA, who then in turn teaches the student) etc. However, the field is still pretty young and is quite unexplored in many dimensions.

. . .

This blog is a part of an effort to create simplified introductions to the field of Machine Learning. Follow the complete series here

Machine Learning : Simplified
Know it before you dive in
towardsdatascience.com

Or simply read the next blog in the series

Growing your own RNN cell : Simplified
Take a peek into the 'deep' world of a single RNN cell
towardsdatascience.com

. . .

References

[1] Wang, Junpeng, et al. "DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation." IEEE transactions on visualization and computer graphics 25.6 (2019): 2168–2180.
[2] Mirzadeh, Seyed-Iman, et al. "Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher." arXiv preprint arXiv:1902.03393 (2019).
[3] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
[4] Liu, Xiaodong, et al. "Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding." arXiv preprint arXiv:1904.09482 (2019).