Knowledge Distillation via Neuron Selectivity Transfer and Teaching Assistant

Sreeja Gaddamidi
sgaddamidi@umass.edu

Pranjali Ajay Parse
pparse@umass.edu

Abstract

Deep neural networks have proved successful in both industry and research in recent years. It's enormous success can be attributed to its capacity to encode vast amounts of data and manipulate billions of model parameters. However, not only because of the high computational complexity, but also because of the massive storage needs, it is difficult to install these cumbersome deep models on devices with low capabilities, such as embedded devices and mobile phones. To address the latency, accuracy, and computational needs at the inference time, different model compression techniques like model quantization and pruning were proposed. One such model compression method widely used and researched is Knowledge Distillation (KD). In this project, we combine knowledge distillation using Neuron Selectivity Transfer(NST) [4] with teaching assistant technique [3] to improve the performance of a student model that is around 10 times smaller than the teacher. We compare the performance of this approach to the performance of a model learned from scratch and NST-KD (explained in section 3), and show that our model achieves better accuracy.

1 Introduction

Deep learning has formed the foundation of numerous artificial intelligence accomplishments in recent years, including a variety of applications in computer vision, reinforcement learning, and natural language processing. It is now possible to train incredibly deep models with thousands of layers on strong GPU or TPU clusters using a variety of new techniques, such as residual connections and batch normalization. For example, training a ResNet model on a common image recognition benchmark with millions of images takes less than ten minutes; training a sophisticated BERT model for language understanding takes less than one and a half hours. Large-scale deep models have had a lot of success, but their high computational complexity and substantial storage requirements make them difficult to use in real-time applications, especially on devices with restricted resources, like video surveillance and self-driving cars.

Our project focuses on Knowledge Distillation via Neuron Selectivity Transfer and Teacher Assistant techniques. Knowledge Distillation is a technique where a large complex model distils its knowledge and passes it to train a smaller network (suitable for deployment) to match the output. It has gotten a lot of attention in recent years from the scholarly community. Since over parameterization enhances generalization performance when fresh data is evaluated, big deep neural networks have achieved amazing success with good performance, especially in real-world scenarios with large-scale data. However, due to the limited processing capacity and memory of mobile devices and embedded systems, deploying deep models is a significant problem. To overcome this problem, researchers originally proposed using model compression to transport data from a large model or an ensemble of models into training a tiny model with minimal accuracy loss. For semi-supervised learning, information transfer between a fully-supervised teacher model and a student model utilizing unlabeled data is also incorporated. Knowledge distillation refers to the process of learning a tiny model from a larger one. A tiny student model is usually supervised by a large teacher model in knowledge distillation.

The primary premise is that in order to achieve a competitive or even superior performance, the student model imitates the teacher model. The main issue is transferring knowledge from a large teacher to a small group of students [2]. According to prior research, when the difference between student and teacher is large, student network performance declines [3]. Given a student network, an arbitrarily large teacher cannot be employed, or, in other words, a teacher can only effectively transfer information to students of a specific size, not smaller. We present multi-step knowledge distillation

to address this problem, which uses an intermediate-sized network (Teacher Assistant) via Neuron Selectivity Transfer (NST) to bridge the gap between the student and the teacher. We also look at the impact of teaching assistant size and expand the approach to include multi-step distillation. The success of our suggested approach is supported by theoretical analysis and extensive experimentation on the Chest X-Ray Images for Classification dataset.

This paper is organized as follows: Section 2 describe works related to Knowledge distillation. Different approaches explored here are given Section 3. Section 4 describes the proposed approach in detail. Experiments are described in Section 5. Finally Section 6 concludes the paper.

2 Related Work

Generally, a teacher-student strategy is used in distillation-based algorithms, in which a big deep network trained for a certain task instructs shallower student network(s) on the same task. The fundamental notions of knowledge distillation or transfer technique have been around for quite some time. [7] demonstrate that the information in an ensemble can be compressed into a single network. By simulating deep neural networks, [8] expand this approach to analyze shallow yet wide, fully linked topologies. The authors introduce the concepts of learning on logits rather than the probability distribution to make learning easier.

In [5], the authors were able to reduce the model size by 5 times with only around 1-3% reduction in the model accuracy. Unlike other feature map comparison based knowledge distillation models, Neuron Selectivity Transfer allows the model to understand how and what the CNN layer learns and also with different kernel function we can also train student model not only with feature based but also relation based knowledge. This model was proved to give better accuracy on many other knowledge distillation techniques. In order to account for the limitation on learning capacity of the student model from teacher when the network size gap between them is very huge, teaching assistant based Knowledge distillation model was proposed. Inspired by these approaches, we combined the Neuron Selectivity Transfer model with teaching assistant technique to improve the performance of a student that has around 10 times less number of parameters as compared to the teacher.

3 Approaches for Knowledge Distillation

Quantization: Quantization is the process of approximating a neural network that employs floating-point values by a neural network with low bit width numbers. The memory requirements and computing costs of employing neural networks are drastically reduced as a result of this. Quantization process, in general, starts off with defining a scaling function $sc : R_n \rightarrow [0, 1]$, which converts vectors with values from any range to vectors with values in the range $[0, 1]$. There are different types of quantization methodologies which are as follows:

1. **Scaling:** Different kinds of scaling functions such as min-max scaling and absolute scaling can be used to scale the weight parameters.

2. **Bucketing:** We apply the scaling algorithm to buckets of consecutive values of a fixed size independently. We get improved quantization precision for each bucket, but we have to store two floating-point scaling factors for each bucket as a trade-off.

3. **Uniform Quantization:** The number of quantization levels used is described by the parameter s>=1. Uniform quantization, on the surface, considers s + 1 equally spaced points between 0 and 1. (including these endpoints).

4. **Non-uniform quantization:** Non-uniform quantization accepts a collection of s quantization points $\{p1, ..., ps\}$ as input and quantizes each element v_i to the point that is closest to it.

Neuron Selectivity Transfer: The Neuron Selectivity Transfer model aligns the distribution of neuron selectivity pattern between student models and teacher models. Each channel in feature map is a sample of selective knowledge of the CNN layer. The probability distribution given by those samples is over the patterns extracted by the CNN layer. If a neuron is activated in certain regions or samples, that implies these regions or samples share some common properties that may relate to

the task. It provides an explanation to what the CNN layer is doing instead of just trying to mimic its output as done when only comparing feature maps. The knowledge of KD is the distribution of selective knowledge of the CNN layer.

Teacher Assistant technique: In knowledge distillation via teacher assistant, in order to reduce the gap between the model architectures to account for capacity limitation while learning, a teacher assistant model is first trained using teacher and the student model is later trained from teacher assistant model.

4 Proposed Approach

We carried out a performance analysis on all the above mentioned approaches on the Chest X-Ray Images for Classification dataset. For the initial experiments, we used the network architectures and distillation methods as mentioned in the original papers. Further, as we studied and understood the importance and role of each of these models in improving the distillation, we implemented the student network architecture based on quantization of a pretrained network that is proved to work well with the above dataset, and calculated distillation using Neuron Selectivity Transfer.

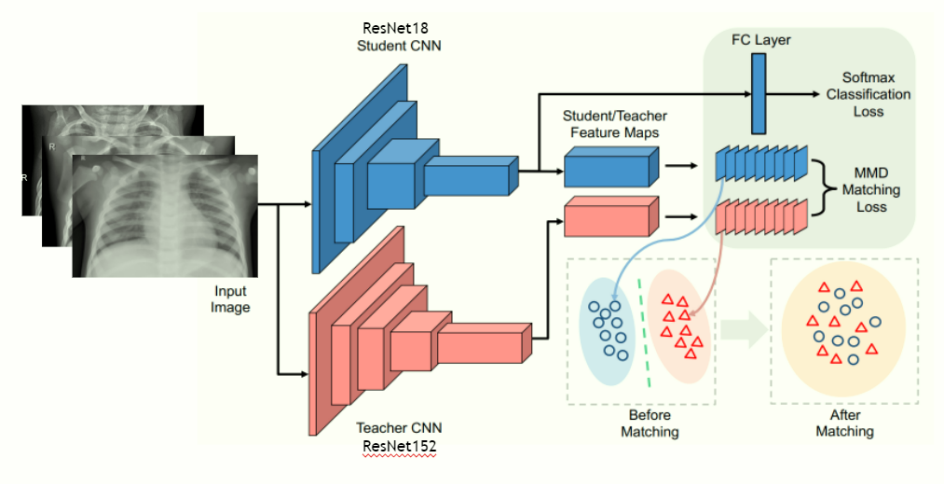

Figure 1: Neuron Selectivity Transfer model architecture

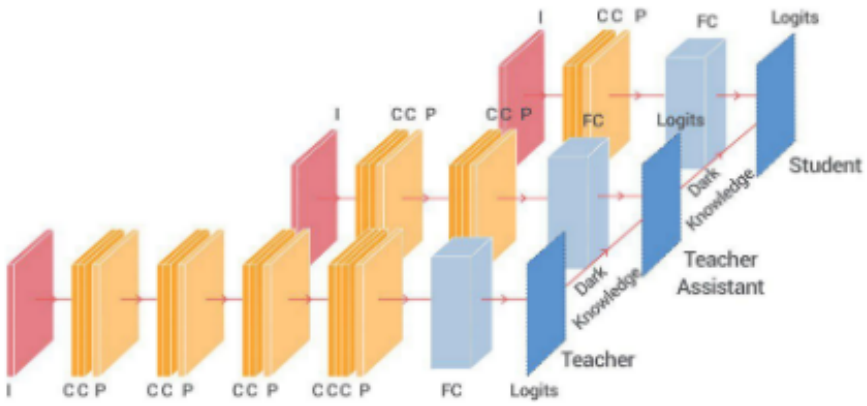

Figure 2: Teacher assistant model architecture

Finally, we improved the performance of the student network using various TAs (we performed the experiments within our resources limitation) in between teacher and the student network. We train the student model ResNet101 with the teacher model which is ResNet152, and then the trained ResNet101 as the Teacher model with ResNet50 as the student model, further trained the ResNet50 (as Teacher) with ResNet34 and so on until ResNet18 as seen in Fig. 1 and Fig. 2.

5 Experiments and Results

5.1 Dataset

We use Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification [1] dataset for all our experiments. The dataset consists of frontal chest X-ray images. The radiographic images were from 5863 pediatric patients one to five years old from the Medical Center in Guangzhou.All images in dataset (Kermany, 2018) underwent a treatment in order to remove all low-quality scans, as well as being classified by two specialist physicians and by a third party specialist, in order to prevent any misclassification. The images in the dataset are varying resolutions such as 712x439 to 2338x2025. There are 1583 normal case, 4273 pneumonia case images in the dataset as shown in Table 5.3.

	Virus Infected Cases	Bacteria Infected Cases	Normal Cases
Training Dataset	1345	2539	1360
Testing Dataset	234	148	242

5.2 Evaluation Metrics

We used accuracy as performance metric for evaluating all the models.

$$Accuracy = \frac{(TP + TN)}{(TP + FN + TN + FP)} \tag{1}$$

TP, TN, FN, FP represents the number of true positive, true negative, false negative, false positive respectively.

5.3 Discussion

We conducted all our experiments on google colab (Intel(R) Xeon(R) CPU @ 2.20GHz with 13GB RAM and Tesla K80 GPU with 12 GB memory). We experimented with different KD parameters, batch sizes, and optimizer parameters and presented the results for the best set of hyperparameters. For the NST-poly TA model, the ResNet152 model achieved a test accuracy of 98.67%, ResNet101 of 97.54, ResNet50 of 94.79, ResNet34 of 92.91.

Model	Base	NST Linear	NST Poly	NST-Poly TA
Test Accuracy	85.28	86.47	89.75	91.63

From Figure 3 and Figure 4, we observe that the NST-poly with TA model has better convergence in terms of accuracy and loss as compared to both the base model that is learnt from the scratch and model learnt using NST-poly from the same teacher. Table 5.3 shows that the NST-poly TA model achieves around 6% better test accuracy as compared to the baseline model and 2% more as compared to the NST-Poly model. This happens because the model is able to learn the features and activations learnt by the teacher using NST loss faster as compared to the baseline with no access to additional knowledge, and also teacher assistant technique reduces the gap and improves the generalization capacity of the student model.

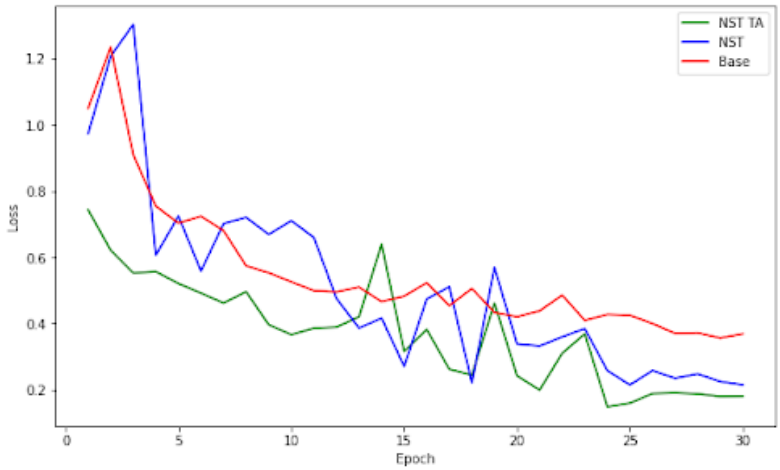

Figure 3: Validation Loss vs Number of Epochs

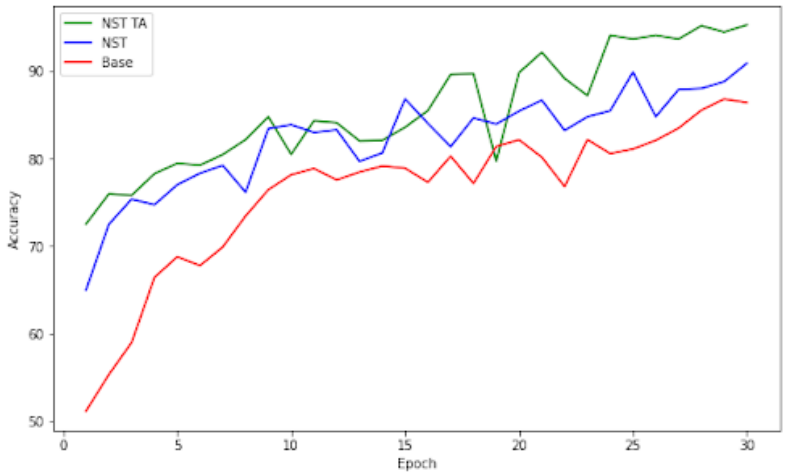

Figure 4: Validation Acc. vs Number of Epochs

6 Conclusion

In this paper, we proposed that the neuron selectivity transfer model when used with teacher assistant technique gives the model accuracy around 2% as compared to KD via NST and around 6% as compared to Base. We observed that NST Poly gives better KD performance as compared to NST Linear, and the TA improves performance of NST model. This approach helps to reduce the parameter size by 10 times (ResNet18 - 6MB and ResNet152 - 58MB) allowing to achieve similar performance as compared to the teacher model.

Due to the limitation on resources our experiments were limited to less training epochs and a small dataset, however, with better computational power (GPUs), we can experiment with larger datasets to evaluate the performance. 3D scenarios can also be explored. This work can also be extended by incorporating Quantized and Differential Distillation and Differential Quantization to quantize weights to a limited set of levels for faster convergence.

References

[1] Kermany, D., K. Zhang, and M. Goldbaum. "Labeled optical coherence tomography (OCT) and chest X-ray images for classification, mendeley data." London, UK (2018).

[2] Gou, Jianping, et al. "Knowledge distillation: A survey." International Journal of Computer Vision 129.6 (2021): 1789-1819.

[3] Mirzadeh, Seyed Iman, et al. "Improved knowledge distillation via teacher assistant." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 04. 2020.

[4] Huang, Zehao, and Naiyan Wang. "Like what you like: Knowledge distill via neuron selectivity transfer." arXiv preprint arXiv:1707.01219 (2017).

[5] Polino, Antonio, Razvan Pascanu, and Dan Alistarh. "Model compression via distillation and quantization." arXiv preprint arXiv:1802.05668 (2018).

[6] Mishra, Asit, and Debbie Marr. "Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy." arXiv preprint arXiv:1711.05852 (2017).

[7] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In ˇ Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Min- ing, KDD '06, pp. 535–541, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150464.

[8] Lei Jimmy Ba and Rich Caurana. Do deep nets really need to be deep? CoRR, abs/1312.6184, 2013.