Selective Teacher Reliance for Knowledge Distillation

Hasan Khan HK3550@NYU.EDU
Department of Computer Science
New York University

Sarvani Nadiminty SN2884@NYU.EDU
Department of Computer Science
New York University

Divya Amin DSA8036@NYU.EDU
Department of Computer Science
New York University

Abstract

Knowledge distillation (KD) is a powerful, well established model compression technique that can face performance limitations when student models attempt to mimic large teacher models on high dimensional tasks like image classification (Cho and Hariharan, 2019). Motivated from ideas in curriculum learning, we explore the idea of selective reliance on the task of image recognition, where a student model relies more heavily on teacher guidance for data samples deemed difficult by a teacher generated curriculum. Experimental results show minimal effect of curriculum setting and selective reliance techniques on student accuracy and convergence.

Keywords: Knowledge Distillation, Curriculum Learning, Dynamic Loss functions

1. Introduction

Knowledge Distillation (Hinton et al., 2015) is a model compression technique commonly used in applied settings where large models are difficult to store and run. KD comprises of a dual model student-teacher framework, where a small capacity student model aims to mimic the performance of a larger capacity teacher model by learning the distribution of the output labels generated from the teacher model trained on the same dataset. More specifically, the student model uses a bipartite loss function $L_{student}$ that incorporates both L_{kd} (the KD loss measured by the KL divergence between the softmax of the student output logits P_S and the softmax of the teacher output logits P_T, scaled by the temperature parameter τ) and L_{ce} (the standard cross entropy training loss using the true labels y_{true}). The parameter λ controls the weight given to each component loss. The student loss and its component losses are defined below:

$$L_{ce} = CE(y_{true}, P_S)$$

$$L_{kd} = \tau^2 KL(P_T, P_S)$$

$$L_{student} = (1 - \lambda)L_{ce} + \lambda L_{kd}$$

However, KD has been shown to provide minimal to no performance gains for certain tasks such as image recognition on ImageNet (Zagoruyko and Komodakis, 2016). Cho and Hariharan (2019) examine the reasons behind failures in this context, noting that large differences in student teacher model capacities may limit the student model's ability to minimize both training loss and KD loss, forcing the student to minimize the KD loss over the train loss.

From this finding, we propose *selective reliance*, a technique for dynamically changing the student's reliance on KD loss by updating λ in $L_{Student}$ at the sample level, where λ is determined by the difficulty of the sample. Difficulty is defined in the context of Curriculum Learning (CL), a training strategy for improving model convergence speed and accuracy that involves training on easily learnable samples before difficult ones.

Difficulty rankings for samples can be determined using the confidence scores of additional models, as outlined by Weinshall and Cohen (2018). In the context of knowledge distillation, we use the teacher model to be distilled as the scoring function (Hacohen and Weinshall, 2019) for each datapoint in the the student's curriculum. We build closely on work conducted by Zhao et al. (2021) where curriculum learning is also used to improve knowledge distillation on image recognition tasks, but aim to employ a distinct curriculum generation scheme from theirs, and a novel mechanism for its effect on KD loss utilization. We hypothesize that relying on the teacher is only in the student's best interest when the sample being evaluated is difficult, and that accuracy yielded from KD w/ CL (KD-CL) can be improved upon with selective reliance (KD-CL-SR) techniques built into KD.

2. Related Works

2.1 Efficacy of Knowledge Distillation

Cho and Hariharan (2019) test the notion of generality in Knowledge Distillation, which states that any student can learn from any teacher by experimenting with various student and teacher network architectures on CIFAR and ImageNet datasets. They conclude that while larger teacher models have increased accuracy, student models trained on larger teachers may become less accurate due to large capacity gaps between student and teacher models. Larger models do not make inherently better teachers, and matching student and teacher capacities must be taken into consideration for proper knowledge distillation. In exploring ways around this capacity gap, they find that early stopping of teacher model training can help mitigate the effect of mismatched capacities.

2.2 Curriculum Learning for Deep Networks

Hacohen and Weinshall (2019) analyzes the effect of curriculum learning on deep network training by examining the challenges of sorting datapoint difficulties and computing mini-batches such that they exhibited increasing levels of difficulty. They sort the difficulty of training examples by using confidence scores for each training image, computed using two methods. The *transfer scoring function* method consists of using the confidence scores from a classifier trained on feature vectors from the penultimate layer activations of a pre-trained ImageNet inception network to rank the data. The second method consists of

training the deep network with the help of uniformly sampled mini-batches, then defining a scoring function by computing the confidence score for each image, also called the *self-taught scoring function*. They found that both methods, combined with pacing functions meant for mini-batch computation, had similar performance.

2.3 Improved Knowledge Distillation via Teacher Assistant

Another way to mitigate KD performance drops caused by model capacity gaps are explored by Mirzadeh et al. (2019) through Teacher Assistant Networks. Mirzadeh et al. (2019) also examines why a teacher-student capacity mismatch caused a drop in student performance, and concludes that the drop in performance is mainly because of two factors arising from enlarging the teacher. The first is that the teacher network starts becoming too complex for the student to mimic their behavior with their current capacity. The second factor is that the teacher's data certainty increases, which weakens the knowledge transfer to students as the logits become less soft and the students learn from the teacher network by matching the logits. They show that introducing a teacher assistant network between the teacher and student resulted in significantly better performance than baseline Knowledge Distillation. They tried improving the performance of a teacher assistant network by inserting an intermediate teacher assistant network between the teacher and the teacher assistant. They repeated the process to form *distillation paths* and demonstrated that the incorporation of intermediate teacher assistants still maintained teacher assistant knowledge distillation's better performance over baseline knowledge distillation.

3. Method

To test the effects of curriculum learning and selective reliance on knowledge distillation in difficult contexts, we conduct various experiments on an image classification task using the CIFAR-10 and CIFAR-100 image datasets (Krizhevsky, 2009), utilizing appropriately sized ResNet (He et al., 2015) architectures for all models. All the code and results are available on our GitHub.

As part of the knowledge distillation framework, a ResNet-152 model pre-trained on ImageNet is fine-tuned on the full size of our training dataset (50,000 samples[1], one model for each CIFAR dataset) to serve as our teacher. Next, a curriculum is generated on the same training set by passing the training set through the teacher model at inference, and deriving confidence scores for each sample by checking its classification status (i.e whether it was correctly classified or not) and its Top-1 score (the highest log probability value). Once confidence scores have been assigned to each sample, samples are sorted into an appropriate easy-to-hard curriculum. Samples correctly classified by the teacher come first, and are secondarily sorted from highest confidence scores (easiest) to lowest confidence scores (hardest). Incorrectly classified samples are placed after all correctly classified points; here, the secondary sort is flipped, and samples with low confidence scores are placed first, followed by incorrectly classified samples with high confidence scores. Care

1. We also attempted to split the CIFAR-100 dataset in 3:2 ratio, the larger part of the dataset for training the teacher model and the smaller part for the teacher to generate inference on and then generate curriculum. This was done as an attempt to avoid the teacher inferencing the same data that it was trained on. Please refer Appendix A for the results.

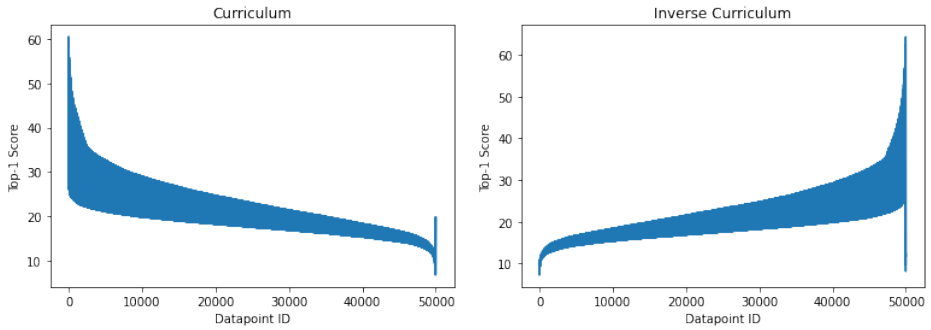

Figure 1: Left: Sample confidences from the teacher generated curriculum. Samples are fed into the student model on the order of left to right. The spike at the toward the end of the curriculum represents the samples misclassified by the teacher model, which are placed at the end of the curriculum and ordered in the reverse order of confidence. Right: Reversed curriculum

is taken to stratify by class so as not to imbalance the curriculum with any particular class concentrated on one end of the curriculum. An inverse curriculum (hard-to-easy) is also generated in a similar fashion for comparison purposes. See Figure 1 for curriculum ordering visualizations.

With the teacher trained and curriculum generated, we conduct six experiments on ResNet-18 student models: Student with no teacher (Student-Unguided) which serves as the baseline for all experiments student w/ standard knowledge distillation (Student-KD), student w/ knowledge distillation using curriculum (Student-KD-CL), student w/ standard knowledge distillation using an inverse curriculum (Student-KD-CL'), student w/ selectively reliant knowledge distillation based on curriculum (Student-KD-CL-SR), and student w/ selectively reliant knowledge distillation based on inverse curriculum (Student-KD-CL'-SR). All teacher and student model hyperparameter values can be found in Appendix A. See figure 2 for a full diagram of the experiment procedures.

Since the curriculum is by definition ordered on difficulty, selective reliance is implemented by scaling up λ from 0 through the duration of the epoch. One lambda value is selected per batch. Values are based on the average confidence score of the batch $\bar{C}_b$, normalized by the maximum global confidence score C_g and shifted by a small scaling factor s (this is to set the initial λ to zero). We define lambda in the context of selective reliance as follows:

$$\lambda = \frac{(1 - \bar{C}_b)}{max(C_g)} - s$$

4. Results

We report the validation accuracy and loss results for both Teacher Models in Table 1, the CIFAR-10 student results in Table 2 and the CIFAR-100 student results in Table 3. We record convergence speeds for selected experiments. This can be found in Figure 3, Appendix A

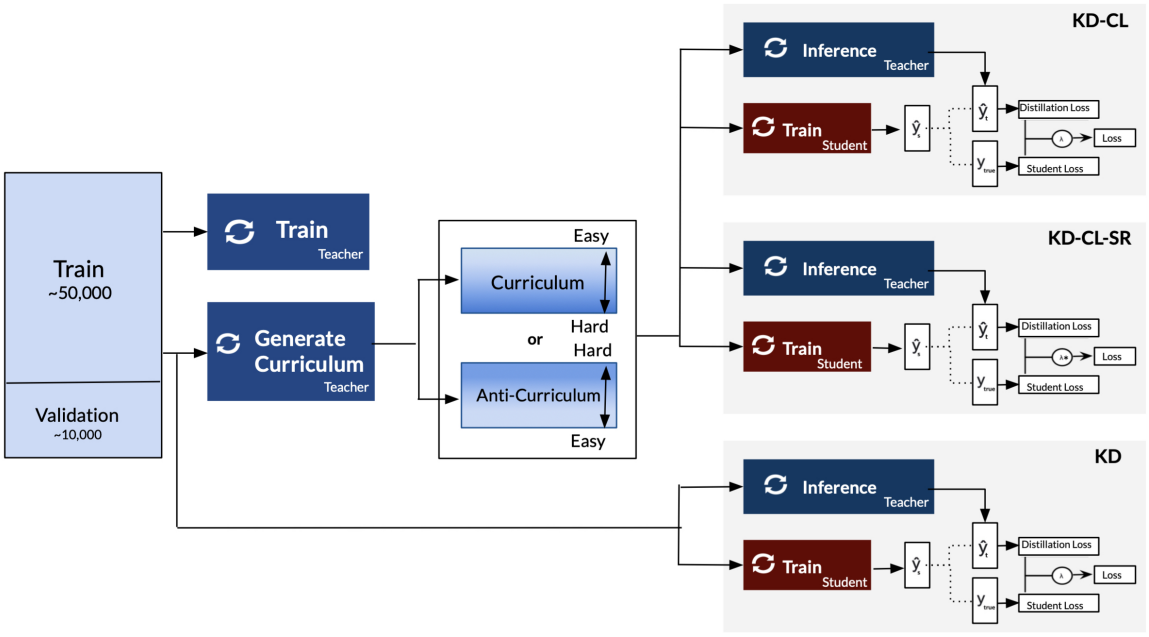

Figure 2: Process flow for teacher training and curriculum generation (left), followed by all experiments (right). $\hat{y}_t$ stands for teacher softmax output, $\hat{y}_s$ stands for student softmax output, and $\hat{y}_{true}$ stands for the true labels. $\lambda*$ denotes dynamic λ based on average batch confidences.

Our experiments show that Knowledge Distillation marginally improves student validation accuracy for both CIFAR-10 and CIFAR-100 datasets over baseline (unguided) students. Curriculum learning does not improve validation accuracy, while inverse curriculum does provide marginal increases. Selective Reliance training for students using the standard curriculum fails to make any noticeable improvement over students without selective reliance; however, selective reliance significantly weakens model performance trained on inverse curriculum for both CIFAR datasets.

Experiment	**Val Acc**	**Val Loss**
Teacher-CIFAR10	88.3%	0.76
Teacher-CIFAR100	66.0%	1.80

Table 1: Teacher Model Scores

	CIFAR-100		CIFAR-10	
Experiment	**Val Acc**	**Val Loss**	**Val Acc**	**Val Loss**
Student-Unguided	45.9%	3.783	77.3%	1.413
Student-KD	47.3%	3.169	79.6%	1.421
Student-KD-CL	46.3%	3.261	80.6%	1.880
Student-KD-CL'	48.0%	3.183	79.7%	2.155
Student-KD-CL-SR	46.9%	3.470	80.0%	2.007
Student-KD-CL'-SR	43.1%	5.485	77.0%	1.10

Table 2: Experiment Accuracy and Loss results

5. Discussion

Our results for Student-KD-CL and Student-KD-CL are in line with findings by Cho and Hariharan (2019) that show marginal improvement by using curriculum in the context of KD. The attempt to advance CL using selective reliance yields little improvement. We hypothesize a number of factors may contribute to curriculum learning and selective reliance's lack of positive effect in our results: the absence of a proper curriculum pacing function, effects of class stratification in curriculum generation, using the teacher model as the curriculum generator (rather than snapshots of the student as done by Cho and Hariharan (2019) or other external models), redefining λ and scaling for selective reliance, among others.

The curriculum in our experiments is generated by the teacher network after fine-tuning on the dataset. In future work, further experiments could be conducted by following other novel methods to generate the curriculum. Other such methods include using multiple teachers to generate the curriculum as proposed by Gong et al. (2016), self-paced curriculum learning where the difficulty is measured during training and the order of the samples is changed dynamically, or using the student network to generate the curriculum during training by using snapshots of the network and changing the order of samples as implemented by Zhao et al. (2021).

Acknowledgments

Thank you to Professor Ranganath for all his guidance towards this project. Thank you to Rohit Bollineni for introductory discussions on knowledge distillation.

References

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. *CoRR*, abs/1910.01348, 2019. URL `http://arxiv.org/abs/1910.01348`.

Chen Gong, Dacheng Tao, Stephen J. Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016. doi: 10.1109/TIP.2016.2563981.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. *CoRR*, abs/1904.03626, 2019. URL http://arxiv.org/abs/1904.03626.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019. URL https://arxiv.org/abs/1902.03393.

Daphna Weinshall and Gad Cohen. Curriculum learning by transfer learning: Theory and experiments with deep networks. *CoRR*, abs/1802.03796, 2018. URL http://arxiv.org/abs/1802.03796.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2016. URL https://arxiv.org/abs/1612.03928.

Haoran Zhao, Xin Sun, Junyu Dong, Zihe Dong, and Qiong Li. Knowledge distillation via instance-level sequence learning, 2021. URL https://arxiv.org/abs/2106.10885.

Appendix A.

Model	**Hyperparameters**
Teacher-Resnet-152	epochs-st = 350 learning rate = 1e-3 loss = Cross entropy optimizer = SGD with momentum=0.9

Table 3: Teacher Model Hyperparameters

Model	**Hyperparameters**
Student Resnet-18 (shared parameters)	epochs = 300 learning rate = 1e-3 student loss = cross entropy distillation loss = KLDivLoss(reduction="batchmean") optimizer = SGD with momentum=0.9 temperature = 4
Student-Unguided	lambda = 0.0
Student-KD	lambda = 0.5
Student-KD-CL	lambda = 0.5
Student-KD-CL'	lambda = 0.5
Student-KD-CL-SR	lambda = $\frac{(1-C_b)}{max(C_g)} - s$
Student-KD-CL'-SR	lambda = $\frac{(1-C_b)}{max(C_g)} - s$

Table 4: Student Model Hyperparameters

Experiment	**Val Acc**	**Val Loss**
Teacher - ResNet152	62.2%	2.02
Student-KD	38.3%	4.9
Student-KD-CL	34.6%	5.01
Student-KD-CL'	35.1%	5.8
Student-KD-CL-SR	35.2%	3.7
Student-KD-CL'-SR	35.0%	3.75

Table 5: CIFAR-100 Results for data split in 3:2 ratio for training and inference

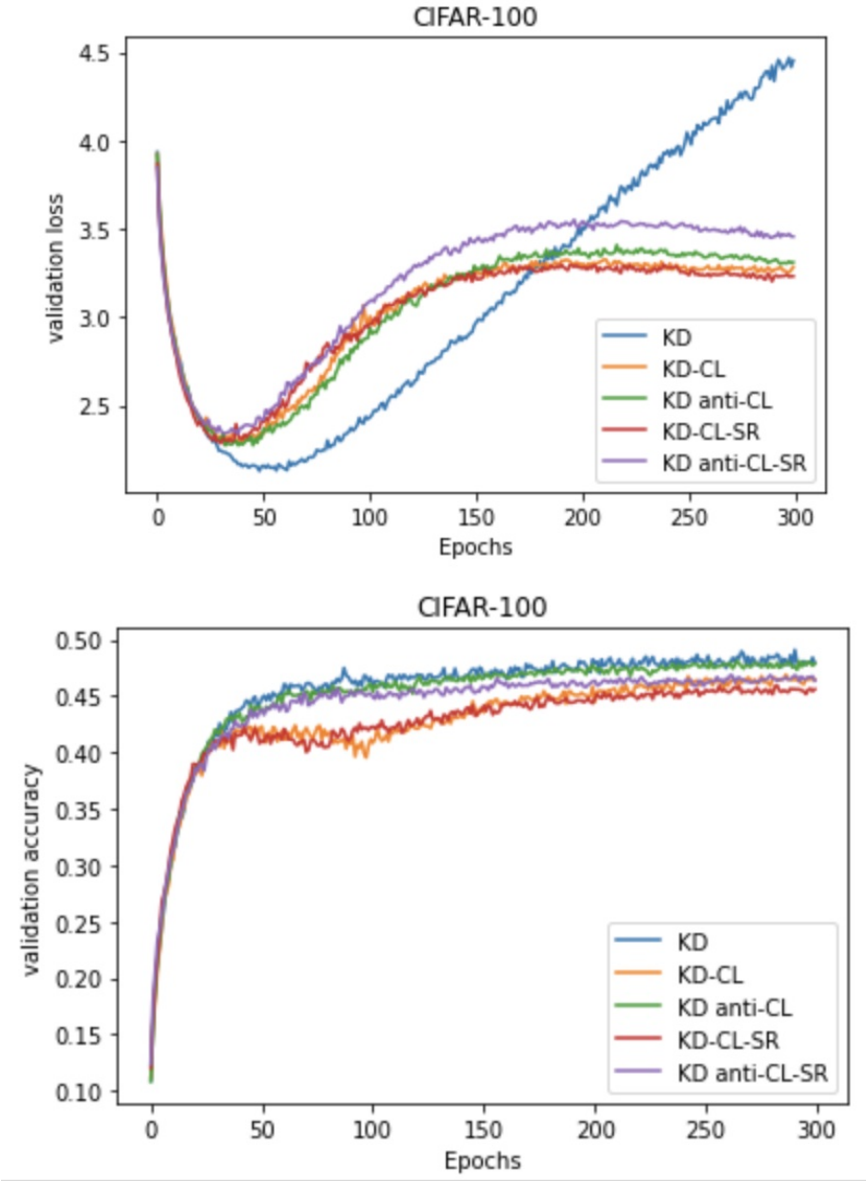

Figure 3: CIFAR-100 validation metrics across epochs.
Validation loss (top) and Validation Accuracy (bottom)