A Studious Approach to Semi-Supervised Learning

Sahil Khose* **Shruti Jain*** **V Manushree***
Manipal Institute of Technology, Manipal
{sahil.khose, shruti.jain, manushree.v}@learner.manipal.edu

Abstract

The problem of learning from few labeled examples while using large amounts of unlabeled data has been approached by various semi-supervised methods. Although these methods can achieve superior performance, the models are often not deployable due to the large number of parameters. This paper is an ablation study of distillation in a semi-supervised setting, which not just reduces the number of parameters of the model but can achieve this while improving the performance over the baseline supervised model and making it better at generalization. After the supervised pretraining, the network is used as a teacher model, and a student network is trained over the soft labels that the teacher model generates over the entire unlabeled data. We find that the fewer the labels, the more this approach benefits from a smaller student network. This brings forward the potential of distillation as an effective solution to enhance performance in semi-supervised computer vision tasks while maintaining deployability.

1 Introduction

Deep learning has achieved remarkable success in many visual and linguistic tasks as a result of recent research advancements. On computer vision applications like classification, larger, deeper models trained in a supervised learning framework have produced the state of the art results. Despite their superior performance, these models are not feasible for real-time deployment due to their computational and memory requirements and the unavailability of large labeled datasets.

In most deep learning problems; more parameters, larger datasets, and more compute results in better accuracy. This is a result of the model's capacity to learn more complex functions, thereby increasing the performance. Distillation with KL Divergence Loss is one of the ways being actively explored for transferring this information obtained by these models to much smaller models. It also acts as an excellent regularizer, especially when there is less labeled data.

This paper is an empirical study of distillation based semi-supervised learning to overcome overfitting, a common problem in semi-supervised setup and bettering performance when limited with small deployable models. We experimented using three architectures: Efficient Net-b5 [10], ResNet-18 [4], and MobileNet-V3-Large [6] to demonstrate the benefit of model compression and four types of label split, highlighting the semi-supervised advantage and model optimization.

2 Related Work

Some of the early works that influenced distillation were by Bucila et al. [2], who used a single neural network that learns by trying to mimic the output of an ensemble of models. The work by Ba and Caruana [1] compresses larger and complex ensembles into small, faster models without much loss in performance using logits. Hinton et al. [5] introduced distillation of knowledge in neural networks

*Authors have contributed equally to this work and share first authorship

I (Still) Can't Believe It's Not Better Workshop at NeurIPS 2021.

using the soft target predictions of the teacher model to train the student network. Intelligent teachers that provide additional privileged information to the students to accelerate the learning process were introduced by [11].

There has been recent advanced research exploring distillation, such as Fitnet [8], where the student learns by mimicking the feature maps of the teacher instead of using the output distribution. Net2Net [3] uses function-preserving transformations to accelerate the transfer of knowledge from smaller neural networks to significantly larger ones. The paper [9] proposes a noise-based regularizer to improve the performance of the student network guided by the teacher. The paper [7] introduces a novel knowledge transfer method in which the distributions of neuron selectivity patterns are matched between the teacher and student models by minimizing the Maximum Mean Discrepancy (MMD) between them. These works have paved the way for enhancing knowledge transfer in neural networks and help in model compression for real-time deployability.

3 Methodology

In knowledge distillation [5], a model is trained first on the dataset; then it is used as the teacher to transfer its knowledge to another model, the student. This information transmission is accomplished through training students on probability distributions of the teacher's predictions rather than hard labels. Instead of the Cross-Entropy Loss, Kullback–Leibler (KL) Divergence Loss is employed. It is a measure of distance between continuous distributions, in this case, the probability distributions of the teacher and student predictions. The KL Divergence Loss 1 for two distributions, P and Q, is calculated as follows:

$$KL\left(P \mid\mid Q\right) = -\sum P\left(x\right) * \log\left(Q\left(x\right)/P\left(x\right)\right) \tag{1}$$

The probability distribution of the output classes represents the measure of the teacher's uncertainty about the prediction, providing additional information to guide the student. To compress the models, we perform knowledge distillation, where the teacher models are larger than the student models. In the case of self-distillation, the student model and the teacher model are the same architecture. Here, we experiment on both knowledge distillation and self-distillation in a semi-supervised setting to obtain better performing and more generalized models.

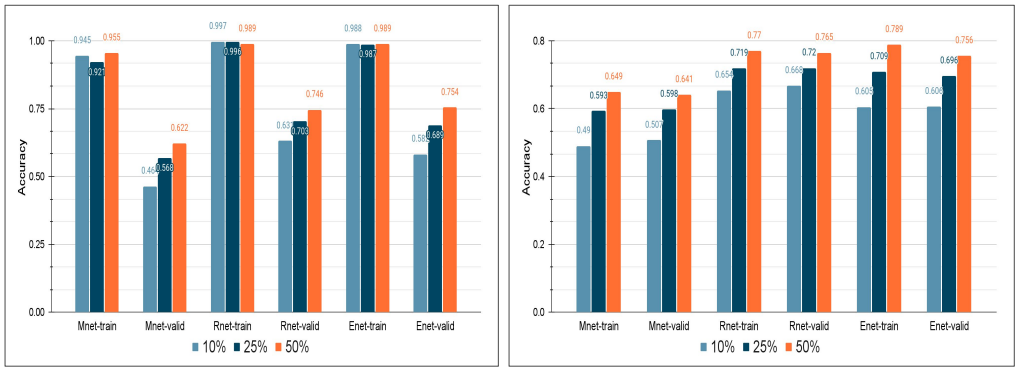
We use the CIFAR-10 dataset that contains 10 classes of natural images with a total of 50000 training and 10000 testing samples, each of which is a 32×32 RGB image. We preprocessed the data by normalizing and augmenting it by using random crops and horizontal flips. The CIFAR-10 dataset is first divided into training, validation, and test sets in the ratio 4:1:1, respectively. Our training data is further divided into two categories: labeled data (X_{lab}) for which we have labels (Y_{lab}) available and unlabeled data (X_{unlab}) whose labels (Y_{unlab}) will not be used for training the teacher model. We will, however, use the entire training dataset for evaluating the student network.

We have used three models for our study: MobileNetV3-Large [6], ResNet-18 [4] and Efficient Net-b5 [10]. Efficient Net-b5 contains roughly 28 million parameters, while ResNet-18 has approximately 11 million parameters, one-third of the former. MobileNetV3-Large has about 4 million, half the times of ResNet-18 and one-seventh of Efficient Net-b5. All the models used were pretrained on ImageNet [**?**]. We perform knowledge distillation from Efficient Net to ResNet and MobileNet; and from ResNet to MobileNet. Along with this, we perform self-distillation on all three models. We experiment on three different label percentages: 10, 25, 50 for a semi-supervised approach, along with 100 percentage labels as the supervised benchmark for comparison.

Implementation details: The teacher model on the labeled data (X_{lab}, Y_{lab}) is trained for 30 epochs. We use Stochastic Gradient Descent for all three models with a learning rate of 0.001 for MobileNet and 0.01 for Efficient Net and ResNet; and momentum of 0.9 with weight decay of $5e-4$. The criterion for training the teacher is Cross-Entropy Loss.

The entire dataset (X_{lab}, X_{unlab}) and the probability distributions of the teacher outputs are passed to the student model. We then train the student model by minimizing the distance between the probability distributions of both the teacher and student model using KL Divergence Loss for 30 epochs. The same optimizer and learning rate as the teacher model are used.

4 Experiments and Evaluation



(a) (b)

Figure 1: Model Accuracies on different splits before distillation (fig. a), after distillation (fig. b). 'Mnet', 'Rnet', 'Enet' represents MobileNetV3-Large, ResNet-18, Efficient Net-b5.

The performance of teacher models trained on different splits of data are shown in Figure 1(a). It is observed that the teacher models are overfitting on all splits displaying substantial variations in the training and validation accuracy, with the larger models showing higher accuracy. We explore knowledge distillation of ResNet to MobileNet; and Efficient Net to ResNet and MobileNet to demonstrate that the models can be compressed with little to no harm to the performance. We further explore self-distillation to show that distillation also acts as a regularizer, as shown in [5].

4.1 Self-distillation in semi-supervised setup

Decrease in overfitting: It is observed that self-distilling the models using KL Divergence Loss enhanced the validation accuracy while simultaneously acting as a regularizer, resulting in a model that generalizes well. The difference in training and validation accuracy has decreased considerably for self-distilled models, as shown in Figure 1(b).

The training accuracy for MobileNet on 10% labels dropped from 0.9450 to 0.4904, and its validation accuracy increased from 0.4638 to 0.5073. Similar behavior can be observed for the other models and other data splits, which is summarized in Figure 1(b).

This is because the student learns richer information from the teacher's uncertainties while making a prediction. As the student attempts to mimic the behavior of the teacher model, it learns not just the correct label for an example but also the probability that the teacher assigns to the other classes. Since the teacher model is not perfect and overfits the dataset, it adds noise to the dataset and helps in the generalization of the student model.

Increase in accuracy: It is observed that students that learned from teachers and trained on data splits with fewer labels had greater improvement in validation accuracy. MobileNet had a 9.37% increase in validation accuracy compared to the teacher model on 10% labeled data and had a 3.02% increase with 50% labeled data. Another inference drawn from the experiments is that models with fewer parameters showed more significant gains in validation accuracy as compared to the larger models. The inference, lesser the labeled data, and smaller the model; greater is the increase in accuracy during distillation is summarized in the Figure 2 (a) and (b) respectively.

This is due to the fact that when the teacher has more labels or has more parameters, it is more confident in its predictions, which results in the soft labels being similar to the hard labels. As a result, the KL Divergence Loss yields fewer benefits since there is comparatively less information to learn from the probability distribution.

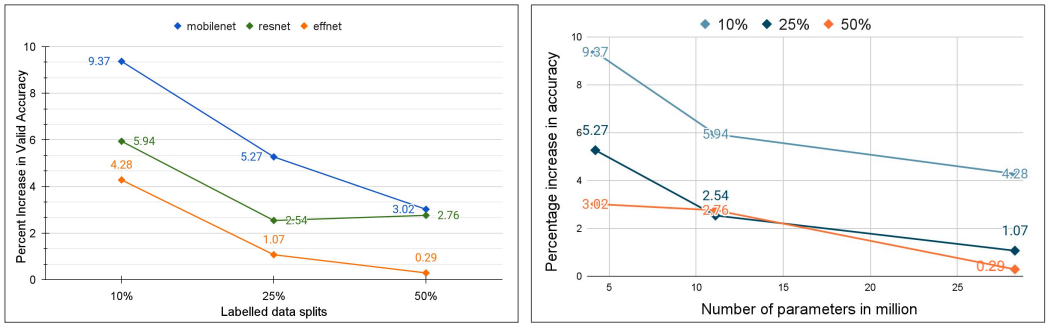


Figure 2: Relative percentage increase in valid accuracy after self-distilling the models on different splits.

4.2 Knowledge distillation in semi-supervised setup

Out of the three models we use, Efficient Net has the largest number of parameters, followed by ResNet and MobileNet. The smaller models can learn and gain knowledge from the bigger models and perform in a similar manner. There is a substantial improvement in accuracy compared to their undistilled counterpart trained directly using hard labels. One of the most significant advantages of knowledge distillation is the ability to compress models without sacrificing accuracy and compromising performance. This is a potential solution for creating deployable models without computation or memory issues. We performed knowledge distillation in the following settings:

Distillation of larger models to MobileNet: We performed knowledge distillation from Efficient Net to MobileNet and from ResNet to MobileNet. The distilled MobileNet model has validation accuracies closer to the teacher models despite having almost 7 times lesser parameters than Efficient Net and 2.6 times lesser parameters than ResNet. When compared to the teacher Efficient Net, the distilled MobileNet model has a validation accuracy gain of 2.68% with 10% labels and a reduction of just 4.09% with 25% labels. Distillation from ResNet also showed validation accuracies closer to the teacher for 10% labels and a small drop in accuracy for 25% and 50% labels. Given the differences in parameters between the two models, this is rather impressive. The distilled MobileNet model consistently outperforms the MobileNet model that is trained without distillation, indicating that even with fewer parameters, performance can be enhanced by transferring information from a larger model. The results of distillation from Efficient Net to MobileNet and ResNet to MobileNet can be seen in Figure 3 (a) and (b) respectively.

Distillation of Efficient Net to ResNet: The distillation of Efficient Net to ResNet showed remarkable results in terms of gain in validation accuracy of the student. Despite lowering the number of parameters to almost half, ResNet's validation accuracy increased by 6.99% compared to Efficient Net's accuracy with 10% labels, 2.68% with 25% labels and by 0.87% with 50% labels. These results are further summarized in Figure 3 (c). The performance is better than the distillation of Efficient Net to MobileNet because of lesser difference in the number of parameters.

It is interesting that, while the validation accuracies of distilled and undistilled ResNet are identical, the distilled ResNet has a lower training accuracy, which makes it better at generalization.

4.3 Distillation in supervised setup

We also carried out the above-stated experiments with fully supervised training. The improvements were not as significant as they were for less labeled data. One reason might be that while using distillation in a semi-supervised setting, KL Divergence Loss is calculated over additional number of datapoints which are unlabeled, thus resulting in increase in performance.

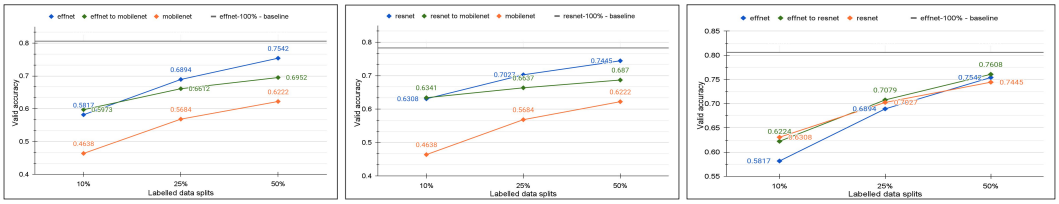


Figure 3: Knowledge distillation of Efficient Net to MobileNet (fig. a), ResNet to MobileNet (fig. b), Efficient Net to ResNet (fig. c).

5 Conclusion

We demonstrate in our study that by distilling semi-supervised models with KL Divergence Loss, we can easily improve their generalization. It has also been shown that a semi-supervised model may be compressed into a smaller model with comparable validation accuracy but greater generalization via knowledge distillation. To summarise, if faced with the challenge of limited labels and memory for deployment, distillation can be a simple method to overcome the problem and make the model deployable. It could also be used as a general practice when performing semi-supervised learning as it is relatively easy to implement with few or no major downsides.

6 Acknowledgement

We would like to thank Research Society Manipal for their valuable inputs and research guidance.

References

[1] L.J. Ba and R. Caruana. Do deep nets really need to be deep? *Advances in Neural Information Processing Systems*, 3:2654–2662, 01 2014.

[2] Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA, 2006. Association for Computing Machinery.

[3] Tianqi Chen, I. Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. *CoRR*, abs/1511.05641, 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 03 2015.

[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[7] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. 07 2017.

[8] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv*, 12 2014.

[9] Bharat Sau and Vineeth Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. 10 2016.

[10] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[11] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. Advances in Neural Networks Research: IJCNN2009.