Knowledge Distillation in Computer Vision

Maobin Guo, Xinyi Pan, Yuwei Zhang
Duke University

Background

- **Knowledge distillation** is a **model compression** method with the goal of deploying deep networks in low computation required and storage limited devices without significant decrease in accuracy [1].
- In knowledge distillation a small model is trained to imitate a pre-trained, cumbersome model or an ensemble of models [2].
- This training process is analogous to the student distils knowledge from the teacher [3].
- The cumbersome model is therefore referred to as the **teacher model**, while the lightweight model is called the **student model** [4].

Objective & Experiments

This project is aimed to classify handwritten digits on **MNIST** and detect objects on **CIFAR10** with knowledge distillation.

Our experiments include

- **Normal Knowledge Distillation** on MNIST:
  - Use the **soft labels** generated by the larger model to train the small network on MNIST.
  - Finetune the **temperature** parameter used to generate soft labels.
  - Test model robustness by **omitting one digit** from the training set of the teacher model and repeating the distillation process.
- **Reversed Knowledge Distillation** on CIFAR-10:
  - **Reverse** the operation of normal knowledge distillation by letting the student model teach the teacher model.
  - Use the soft labels generated by the smaller model to train the larger model on CIFAR-10.
- **Self-distillation** on CIFAR-10:
  - Let the model **learn by itself** to improve performance.
  - Train the student model to obtain a pre-trained model and use this pre-trained model to generate soft labels to train itself again.

Datasets

Our datasets include MNIST and CIFAR-10.

- The **MNIST** dataset is a large collection of **handwritten digits**.
  - The dataset has a training set of 60,000 examples, and a test set of 10,000 examples.
  - The digits are grayscale and centered in a 28 × 28 image.

Figure 1: MNIST Dataset Examples

- The **CIFAR-10** dataset is an established dataset used for **object recognition**.
  - The dataset consists of 60,000 32 × 32 color images containing one of 10 object classes, with 6000 images per class.
  - The images are labelled with one of 10 **mutually exclusive** classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.
  - There are 50000 training images and 10000 test images.

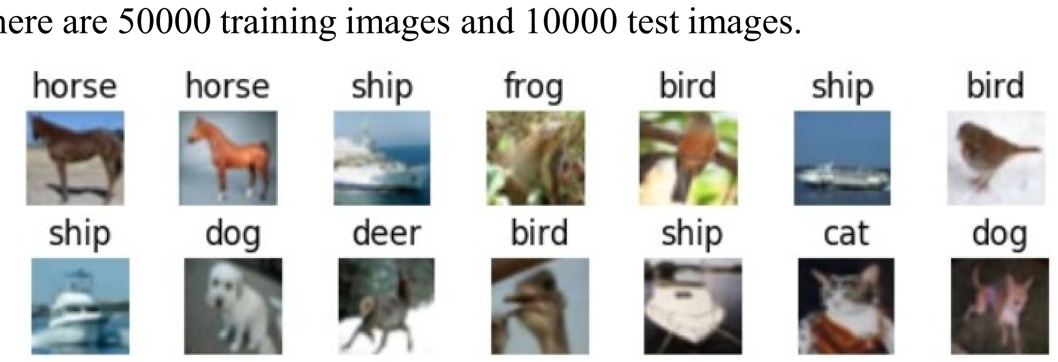


Figure 2: CIFAR-10 Dataset Examples

Methods

To learn about learning abilities of different model distillation methods, we designed different scenarios and corresponding models to conduct a series of experiments.

Firstly, we checked the performance of distilling smaller linear models (student) from larger linear models (teacher) (**Normal KD**).

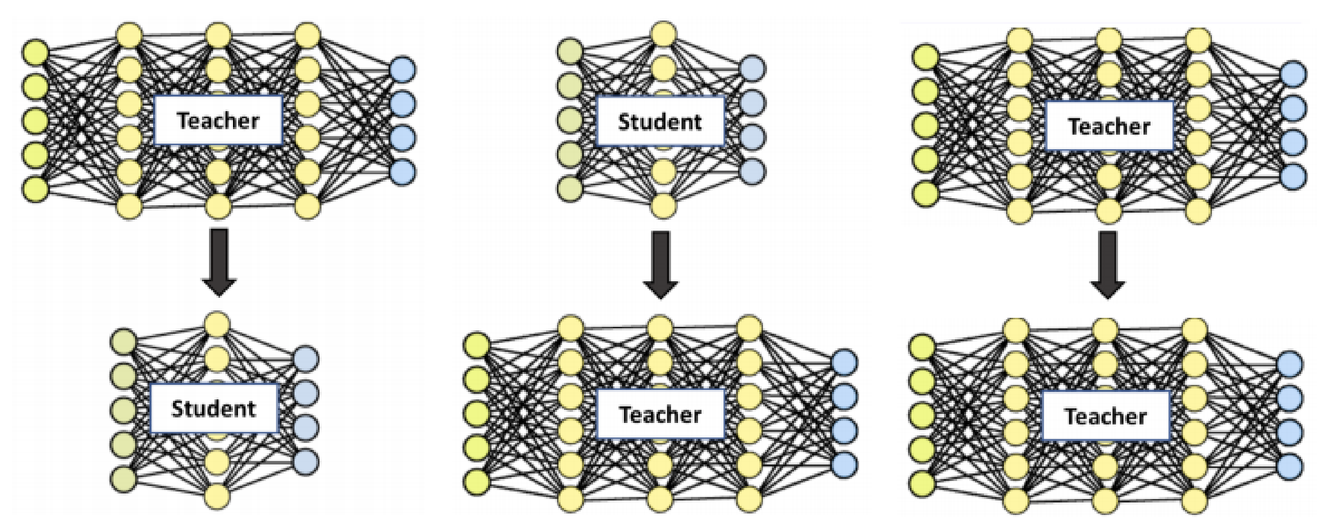
- Compare performance and efficiency of the larger model trained on true labels with that of the smaller model
- Train a smaller model on a combination of true labels and soft labels from teacher model, then compare it with teacher model as well as the small model trained only on true labels
- Experiment with adding temperature and the effect of different temperatures on the performance of distilled student model
- Use distilled models' prediction on the omitted digit to evaluate the generalization ability learned from teacher model

Then we let large model (teacher) learn from small model (student) (**Reversed KD**).

- Train the large model on a combination of true labels and soft labels from the small model. Then compare the accuracy of the large model learned from the small model with a large model learned from the ground truth.

At last, we let models learn from themselves (**Self KD**).

- Train a student model on a combination of true and soft labels from a teacher model. The teacher model shares the same network structure as the student model. Then compare performance between teacher and student model.



(a) Normal KD (b) Reversed KD (c) Self KD

Figure 3: Diagrams of experiments we conduct

Discussion

1. Knowledge distillation does not destroy and even **improve** the model performance.
   - The distilled model trained from **ground truths and soft labels** generated by teacher performs significantly better than the same model solely trained from ground truths.
   - The distilled classifier trained on **partial classes** can still achieve high testing accuracy on all the classes, indicating that distillation from soft labels empowers the model with **greater generalization** ability.
2. Knowledge distillation is not restricted to "**teacher teaches student**" but can also be used as "**student teach teacher**."
   - The larger model (teacher) can improve by learning from the small model (student).
   - It suggests that knowledge distillation works not because teacher is larger than student in model size but due to the combination of soft labels and ground truths.
3. Knowledge can be learned between **heterogeneous networks**.
   - Two networks with different structures, VGG and ResNet. Still, the distillation achieves high classification accuracy (>90%).
4. Our **self-distillation** further confirms that knowledge distillation outperforms the traditional training process due to soft labels.
   - In self-distillation the student and teacher share the same structure, training hyperparameter, and training data. The only difference comes from the soft labels used in self-distillation, which greatly boosts the classification performance.

Results

Results from the series of experiments on distilling small student model from large teacher model:

- Large Model vs Small Model

Model Type	Epoch	Valid Accuracy
Large Model	25	98.77%
Small Model	25	98.73%

- Distilled Model vs Small Model

Model Type	Labels	Valid Accuracy
Distilled Model	true + soft	98.97%
Small Model	true	98.76%

- Relationship between temperature and student performance is not linear, the recommended 20[2] leads to the highest validation accuracy of 99%
- Distilled Model vs Student Model with omitting digit 3

Model Type	Data	Valid Accu on 3	Overall Valid Accu
Distilled Model	omit 3	95.74%	96.63%
Small Model	omit 3	0.00%	88.6%

Results from the series of experiments on distilling a large student from a small teacher model and teacher modes learn from themself.

- The large model validation accuracies rise from 90.23% to 91.26% after learning from the small model whose accuracies is 88.23%
- Both small and large models can benefit from self-learning, which raises their validation accuracies by 1.0% and 2.98%.

Model	#param	FLOPs
LinearTeacher	2.39×10^6	2.39×10^6
LinearStudent	1.27×10^6	1.27×10^6

Model	#param	FLOPs
ResNet18	11.17×10^6	1.8×10^9
VGG13	9.41×10^6	229.61×10^6

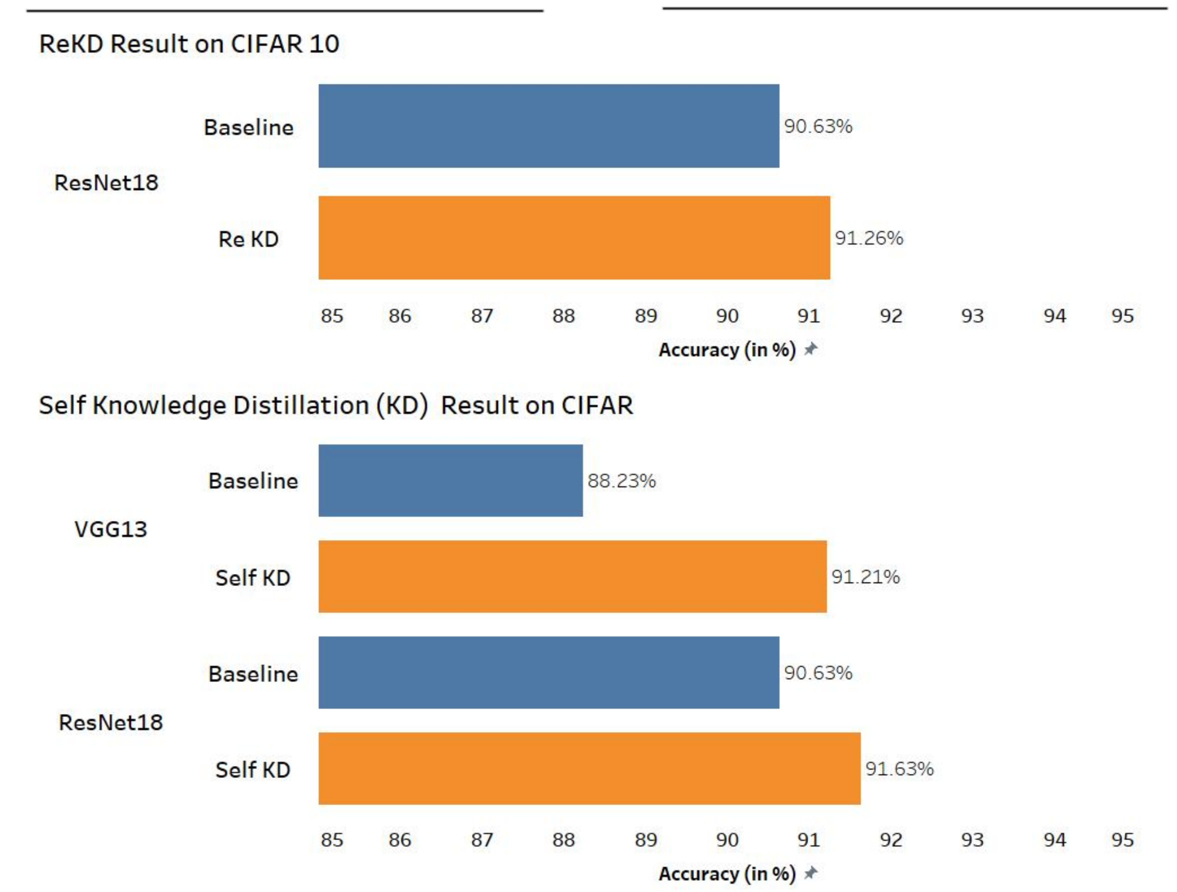


Figure 4: Results of Reversed KD and Self-KD on CIFAR-10

References

[1] Balamurali Murugesan, Sricharan Vijayarangan, Kaushik Sarveswaran, Keerthi Ram, and MohanasankarSivaprakasam. Kd-mri: A knowledge distillation framework for image reconstruction and image restoration n mri workflow. *ArXiv, abs/2004.05319, 2020.*

[2] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. *Distilling the knowledge in a neural network.ArXiv,abs/1503.02531,* 2015.

[3] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via labelsmoothing regularization. *In2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition,CVPR 2020*, Seattle, WA, USA, June 13-19, 2020, pages 3902–3910. IEEE, 2020.

[4] Zhang, Linfeng, et al. "Be your own teacher: Improve the performance of convolutional neural networks via self distillation." *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2019.