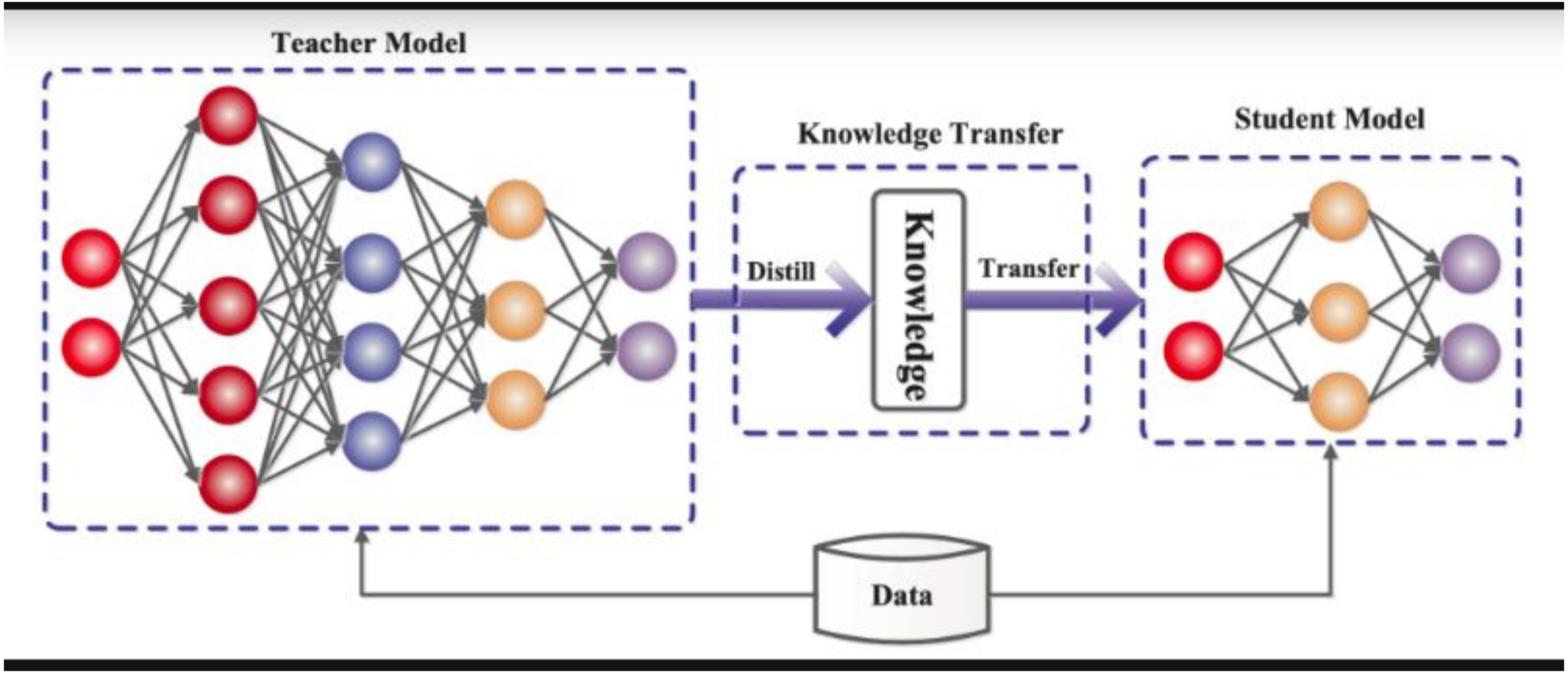
Knowledge Distillation

Ansh Prakash

Overview

1. Knowledge distillation
2. Distillation Loss
3. Architecture
4. Dataset
5. Experiments
6. Conclusion



Source: https://neptune.ai/blog/knowledge-distillation

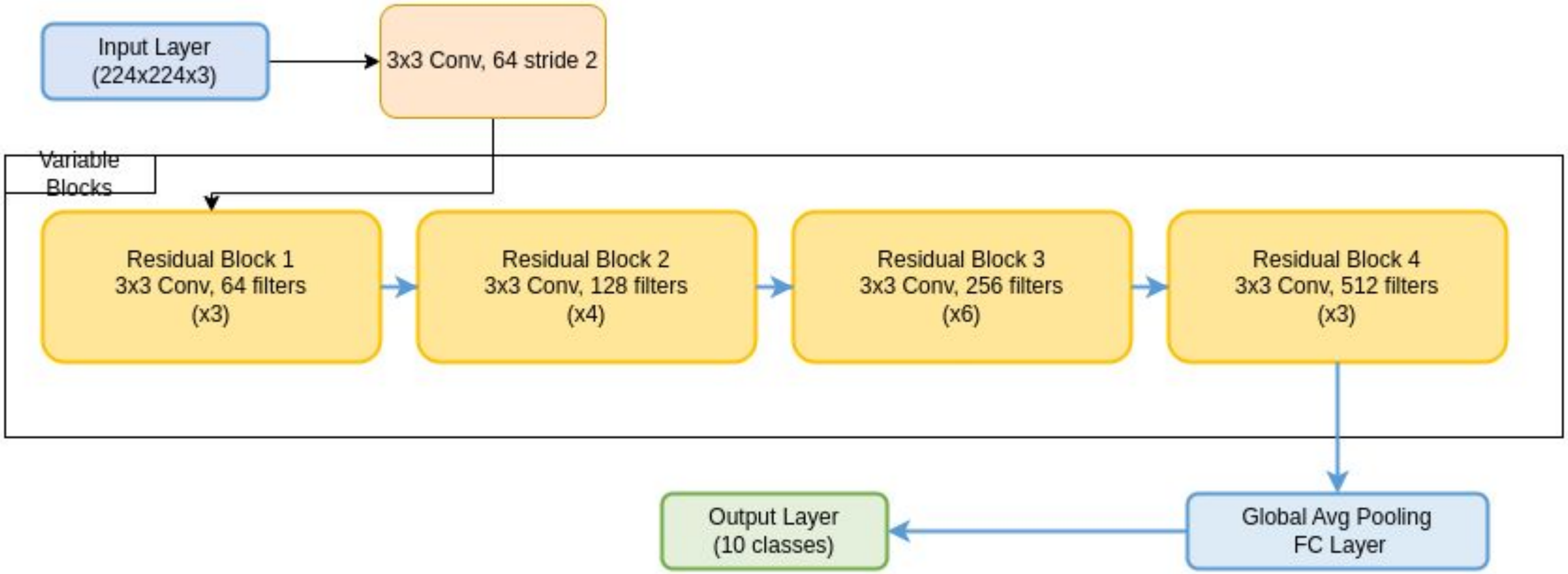
Distillation Loss

$$\mathcal{L}_{\text{distill}} = \alpha \cdot T^2 \cdot \mathrm{KL}(\sigma(z^{\text{student}}/T), \sigma(z^{\text{teacher}}/T)) + (1-\alpha) \cdot \mathcal{L}_{\text{CE}}(y, \sigma(z^{\text{student}}))$$

Where:

- α is a weighting factor.
- T is the temperature.
- $\sigma(z)$ is the softmax function applied to the logits z.
- $\mathrm{KL}(p, q)$ is the Kullback-Leibler divergence.

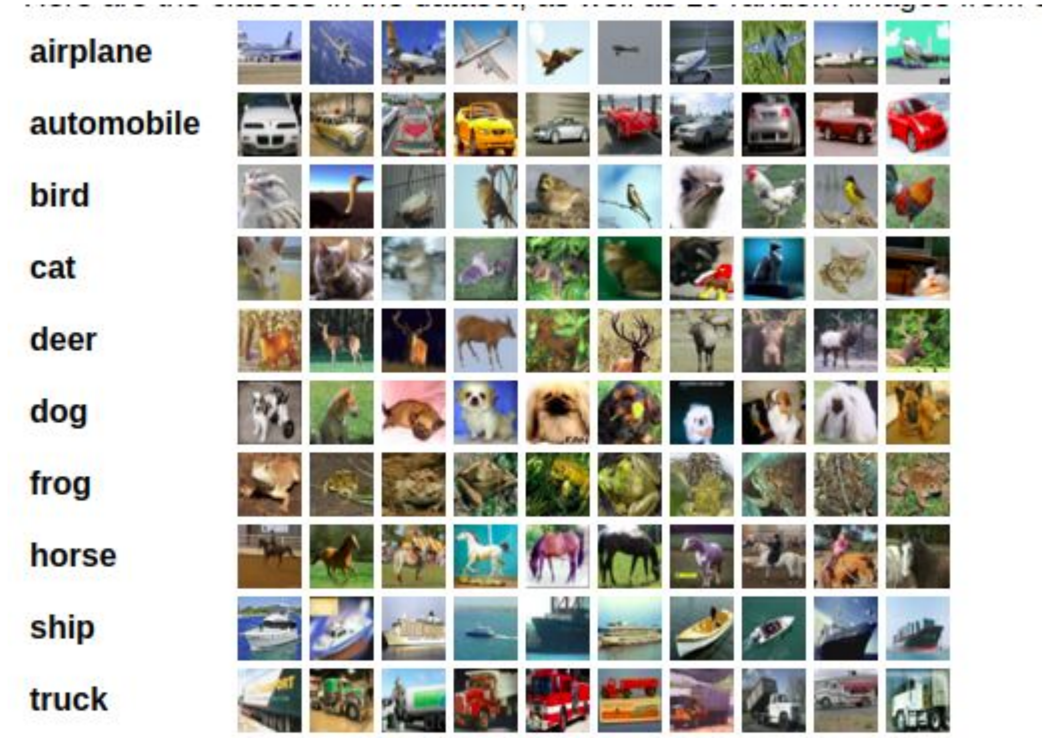
Architecture: ResNet Variants



ResNet Variants

1. Teacher: ResNet-10
2. Students:
    a. ResNet-3
    b. ResNet-4
    c. ResNet-5
    d. ResNet-6

Dataset : CIFAR10



Dataset split

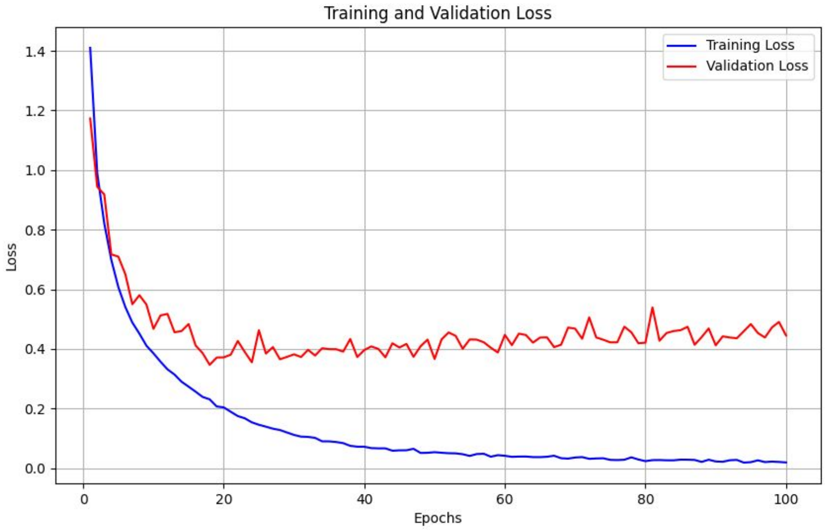
- Generated the validation set from a random split to training set
- 20% used as Validation set and rest 80% for training
- Test set is separately available

Baseline vs KD Results

Models	Baseline test accuracy %	KD test accuracy %
Renet-10[Teacher]	90	-
Resnet-3	64	67(+3)
Resnet-4	70	76(+6)
Resnet-5	82	84(+2)
Resnet-6	87	87(+0)

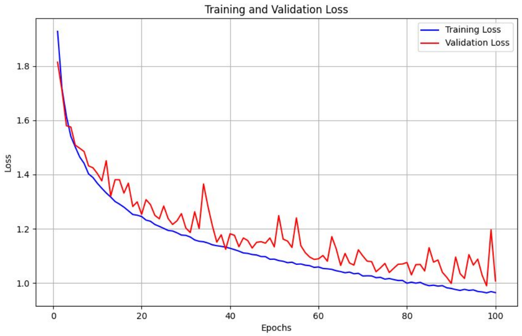
Loss Curves for Baseline Training

Teacher Model: ResNet-10

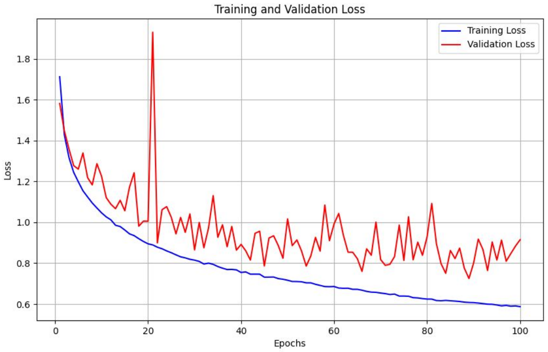


Loss curve for Baseline training

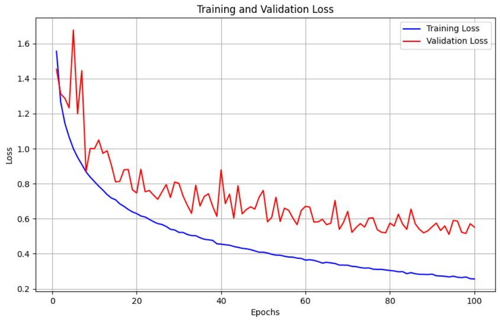
Resnet-3



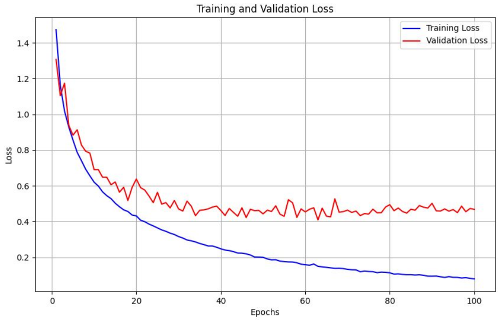
Resnet-4



Resnet-5

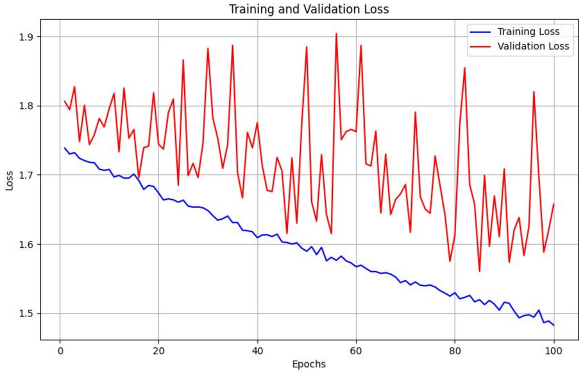


Resnet-6

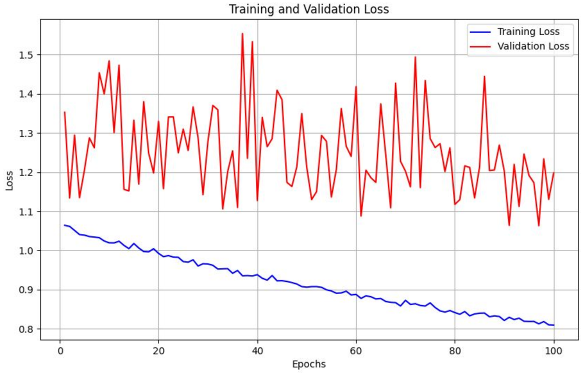


Loss Curves for KD training

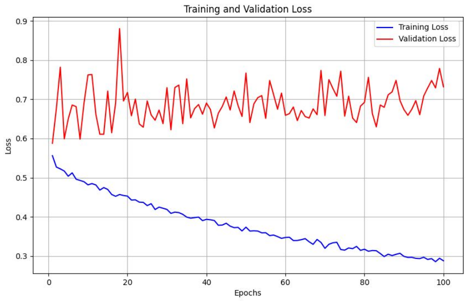
Resnet-3



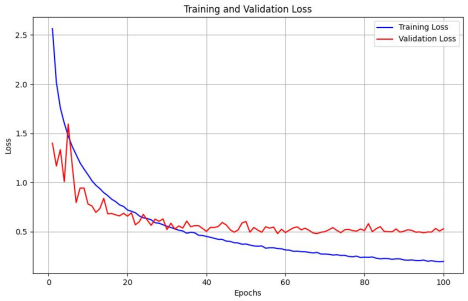
Resnet-4



Resnet-5



Resnet-6



Conclusion

- KD does improve accuracy for small and moderate model sizes
- For relatively larger like resnet-6 there was no improvement
- Effect of KD is best for resnet-4
  - as it has the capacity to learn from teacher model
  - Learning from hard label for smaller model is hard

References

- https://pytorch.org/tutorials/beginner/knowledge_distillation_tutorial.html
- https://arxiv.org/pdf/1503.02531