Exploring Knowledge Distillation of DNNs for Efficient Hardware Solutions

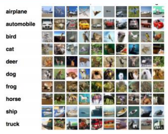
Haitong Li
Department of Electrical Engineering, Stanford University; Email: haitongl@stanford.edu

Motivation and Quick Summary

- **Knowledge distillation (KD): training student models with knowledge from teachers**
  - Utilizing "soft targets" learned by teacher models on training data
  - Beneficial for small models deploying on resource-constrained edge devices
  - Dark knowledge: not fully understood by community; worth "exploring" with experiments and data analysis for future work inspiration
- **What has been done in this project (PyTorch framework):**
  - Explored KD training on MNIST and CIFAR-10 datasets (unlabeled/data-less schemes)
  - Networks: MLP, 5-L CNN, ResNet, WideResNet, ResNext, PreResNet, DenseNet
  - Dark knowledge provides regularization for both shallow and deep models

Datasets and Methodology

- MNIST dataset (60,000/10,000)
- Normalization



- CIFAR-10 dataset (50,000/10,000)
- Normalization
- Augmentation (random crop, random horizontal flip)

Training pipeline and KD loss implementation

Teacher training → 'softened' targets → Student training with KD loss

$$q_i = \frac{exp(z_i \,/\, T)}{\sum exp(z_j \,/\, T)}$$

$$L_{KD}(W_{student}) = \alpha T^2 * CrossEntropy(Q_S^\tau, Q_T^\tau) + (1-\alpha) * CrossEntropy(Q_S, y_{true})$$

Training with Unlabeled MNIST Data: Dark Knowledge

NN architecture & distillation details	Learning rate: 0.01	Learning rate: 0.1
MLP-784-1200-1200-10 (dropout=0.8)	98.30% (as the teacher model)	Learning rate too high
MLP-784-800-800-10	98.10%	97.75%
MLP-784-800-800-10 w/ KD	98.18%	98.50%
MLP-784-800-800-10 w/ KD **(unlabeled training data)**	97.69%	**98.16%**

Shallow and Deep Distillation Experiments with CIFAR-10

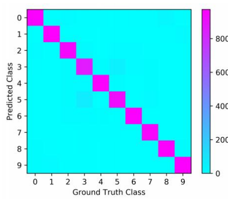
Distill ResNet-18 for 5-layer CNN

	Dropout=0.5	No dropout
5-layer CNN 3 CONV (w/ BN) + 2 FC	83.51%	84.74%
5-layer CNN **w/ ResNet18-KD**	84.49%	**85.69%**
5-layer CNN 5% training data	65.86%	/
5-layer CNN w/ ResNet18-KD 5% training data	66.71%	/

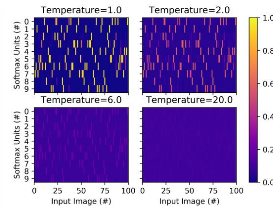
'Deeper' distillation for ResNet-18

	Evaluation accuracy (10k samples)
Baseline ResNet-18	94.175%
+ KD WideResNet-28-10	94.333%
+ KD PreResNet-110	94.531%
+ KD DenseNet-100	94.729%
+ KD ResNext-29-8	**94.788%**

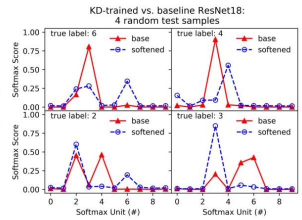
Visualization & Analysis: ResNext-29 → ResNet-18



confusion matrix



Impact of distilling temperature



Test analysis

Discussions and Future Work

- KD provides regularization benefits, even for well-designed state-of-the-art models
- Training with unlabeled data or partial dataset should leverage previous dark knowledge
- As expected, benefits on "easy" dataset are limited. Future work needed on ImageNet

References

[1] Hinton, Geoffrey, et al., arXiv:1503.02531 (2015).
[2] Romero, A., et al., arXiv:1412.6550 (2014).
[3] Lopes, R. G., et al., arXiv preprint arXiv:1710.07535 (2017)
[4] Some pre-trained models: https://github.com/bearpaw/pytorch-classification