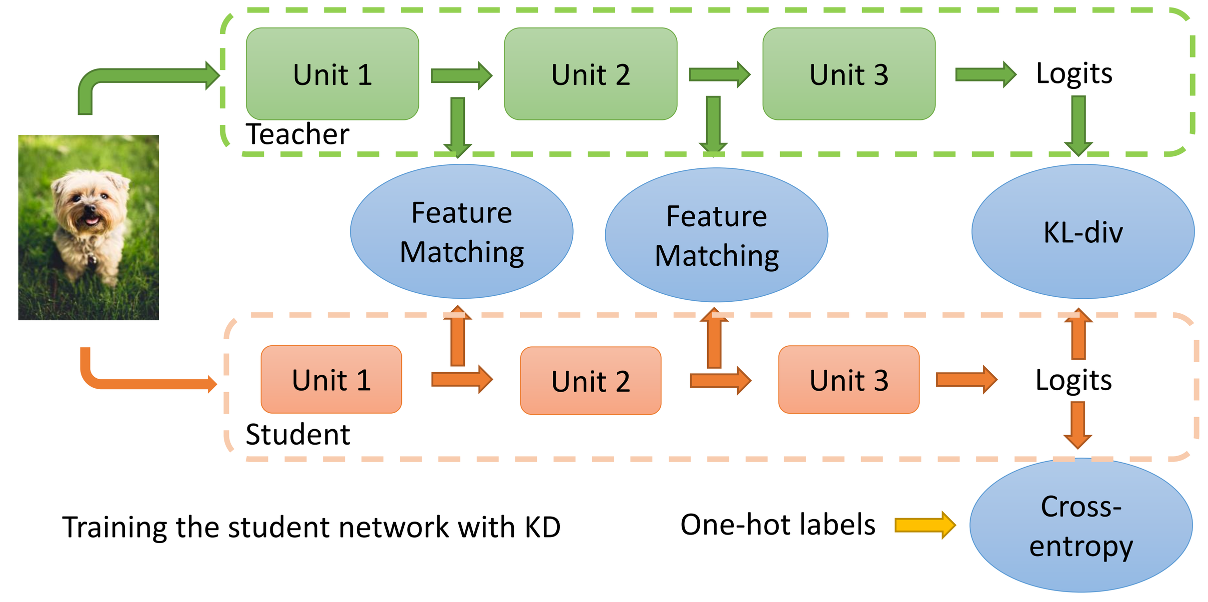
Model Compression Using Optimal Transport

Suhas Lohit and Michael Jones

Mitsubishi Electric Research Laboratories, Cambridge, MA, USA

WAIKOLOA, HI JAN 4-8

Knowledge Distillation (KD)

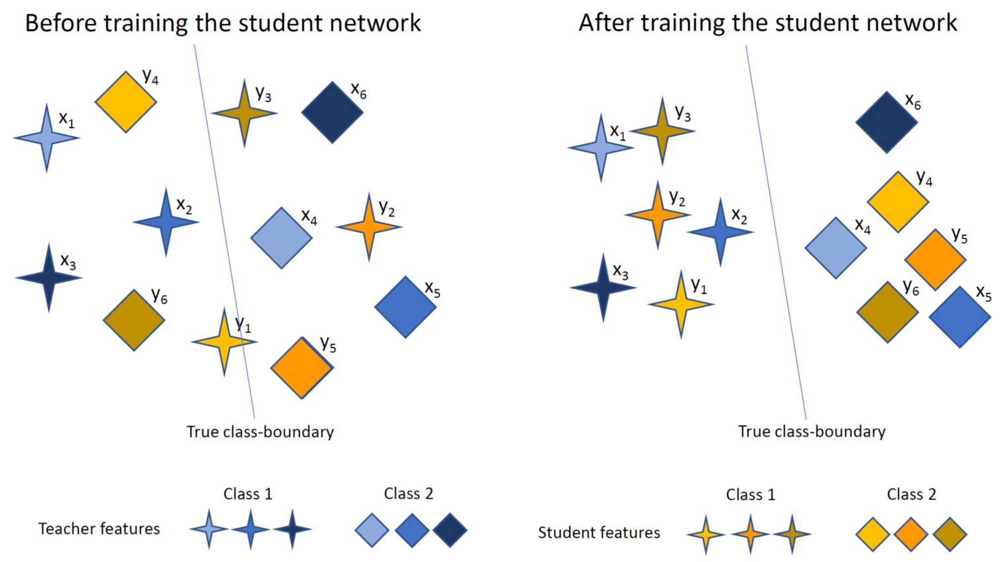


- Accurate deep neural networks for vision are usually very large and cannot be easily deployed in resource-constrained settings
- Model compression is an important research direction to make networks smaller without losing accuracy

- KD is one of the main ways to achieve model compression, by transferring knowledge from a larger, more accurate teacher to a smaller student network.
- In order to train the student, the earliest methods used a combination of the usual cross-entropy loss with the K-L divergence b/w student and teacher outputs
- Student performance can be further improved using supervision at the intermediate layers by adding additional loss terms that encourage matching the teacher and student features. E.g., Fitnets and Relational KD

Using optimal transport (OT) for feature matching

- Optimal transport matches student and teacher feature distributions in a principled way
- Unlike methods like FitNets, it relaxes the unnecessary requirement that teacher and student features need to match one-to-one
- It is a stronger condition than in Relational KD which only matches distance matrices computed in the teacher and student feature spaces



$$L_{OT}(X^{(l)},Y^{(l)}) = \min_{T\geq 0}\sum_{i,j} T^{(l)}_{i,j}C^{(l)}_{i,j}$$

$$\text{s.t.} \sum_i T^{(l)}_{i,j} = \sum_i T^{(l)}_{i,j} = \frac{1}{b},$$

$$L = L_{CE}(\mathbf{c},\hat{\mathbf{c}}_S) + \alpha\sum_{l=1}^{l_{max}} L_{OT}(X^{(l)},Y^{(l)}) + \gamma L_{KD}(\hat{\mathbf{c}}_S,\hat{\mathbf{c}}_T) \qquad C^{(l)}_{i,j} = 1 - \frac{\mathbf{x}^{(l)T}_i\mathbf{y}^{(l)}_j}{\|\mathbf{x}^{(l)}_i\|\|\mathbf{y}^{(l)}_j\|}$$

Relaxations of OT for KD

- We use relaxations of OT in order to solve the OT problems at multiple layers efficiently
- We experiment with
  - Relaxed Earth Mover's Distance (REMD)
  - Inexact Proximal Optimal Transport (IPOT)
- Both can be easily integrated with modern deep learning toolboxes

$$L_{ROT}(X^{(l)},Y^{(l)}) = \min_{T\geq 0}\sum_{i,j} T^{(l)}_{i,j}C^{(l)}_{i,j} + \epsilon h(T)$$

$$\text{s.t.} \sum_i T^{(l)}_{i,j} = \sum_j T^{(l)}_{i,j} = \frac{1}{b},$$

$$R^{(1)}_{OT}(X^{(l)},Y^{(l)}) = \min_{T\geq 0}\sum_{i,j} T^{(l)}_{i,j}C^{(l)}_{i,j} \quad \text{s.t.} \sum_i T^{(l)}_{i,j} = \frac{1}{b}$$

$$R^{(2)}_{OT}(X^{(l)},Y^{(l)}) = \min_{T\geq 0}\sum_{i,j} T^{(l)}_{i,j}C^{(l)}_{i,j} \quad \text{s.t.} \sum_j T^{(l)}_{i,j} = \frac{1}{b}$$

The final relaxed EMD (REMD) is computed using

$$\begin{aligned} &L_{REMD}(X^{(l)},Y^{(l)}) \\ &= \max(R^{(1)}_{OT}(X^{(l)},Y^{(l)}), R^{(2)}_{OT}(X^{(l)},Y^{(l)})) \\ &= \frac{1}{b}\max\left(\sum_i \min_j C^{(l)}_{i,j}, \sum_j \min_i C^{(l)}_{i,j}\right) \end{aligned}$$

Experimental results on image recognition datasets

CIFAR-100

Numbers shown are accuracies (higher is better)

Teacher Student	WRN-40-2 WRN-16-2	resnet110 resnet20	resnet32x4 resnet8x4	vgg13 vgg8	resnet32x4 ShuffleNetV2
Teacher	75.61	74.31	79.42	74.64	79.42
Student (no distillation)	73.26	69.06	72.50	70.36	71.82
KD	74.92	70.67	73.33	72.98	74.45
CRD+KD	75.64	**71.56**	75.46	74.29	76.05
FitNet+KD	75.12	70.67	74.66	73.22	75.15
RKD+KD	74.89	70.77	73.79	72.97	74.55
REMD + KD	**75.79**	70.98	**76.06**	**74.35**	76.66
IPOT + KD	75.63	71.29	75.99	74.29	**76.78**
IPOT + CRD	75.57	71.47	76.06	74.30	76.81
IPOT + CRD + KD	**76.22**	**71.81**	**76.82**	**74.79**	**76.81**

ImageNet

Teacher: Resnet-34, Student: ResNet-18

Numbers shown are error rates (lower is better)

	Teacher	Student	KD	Online KD *	CRD	CRD+KD	AT	SP	CC	IPOT	IPOT+KD
Top-1	26.69	30.25	29.34	29.45	28.83	28.62	29.30	29.38	30.04	29.54	28.88
Top-5	8.58	10.93	10.12	10.41	9.87	9.51	10.00	10.20	10.83	10.48	9.66

Street View House Numbers (SVHN)

Numbers shown are accuracies (higher is better)

T-S pair	Teacher	Student	KD	CRD	CRD+KD	FitNet	Fitnet+KD	RKD	RKD+KD	PKT	PKT+KD	REMD	REMD+KD	IPOT	IPOT+KD
resnet32x4 resnet8x4	94.36	90.39	94.49	**94.96**	**95.47**	91.32	94.48	93.30	94.58	90.77	94.38	89.66	94.49	91.63	94.73
WRN-40-2 WRN-16-2	94.52	93.45	95.22	94.74	95.25	93.93	95.27	95.23	**95.39**	93.68	95.15	93.15	94.94	94.28	**95.41**

Conclusion

- We have presented feature matching methods using optimal transport between teacher and student features at intermediate layers
- We have shown improved performance in knowledge distillation using optimal transport compared to methods like FitNets and RKD

References

1. Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, arXiv 2014.
2. Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation, CVPR 2019
3. IPOT: Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact Wasserstein distance, PMLR 2020
4. REMD: Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances, ICML 2015.