Indian Institute of Science
Bangalore, India
भारतीय विज्ञान संस्थान
बंगलौर, भारत

Department of Computational and Data Sciences

Zero-Shot Knowledge Distillation in Deep Networks

Gaurav Kumar Nayak[1], Konda Reddy Mopuri[1,2], Vaisakh Shaj1[1,3], R. Venkatesh Babu[1], Anirban Chakraborty[1]

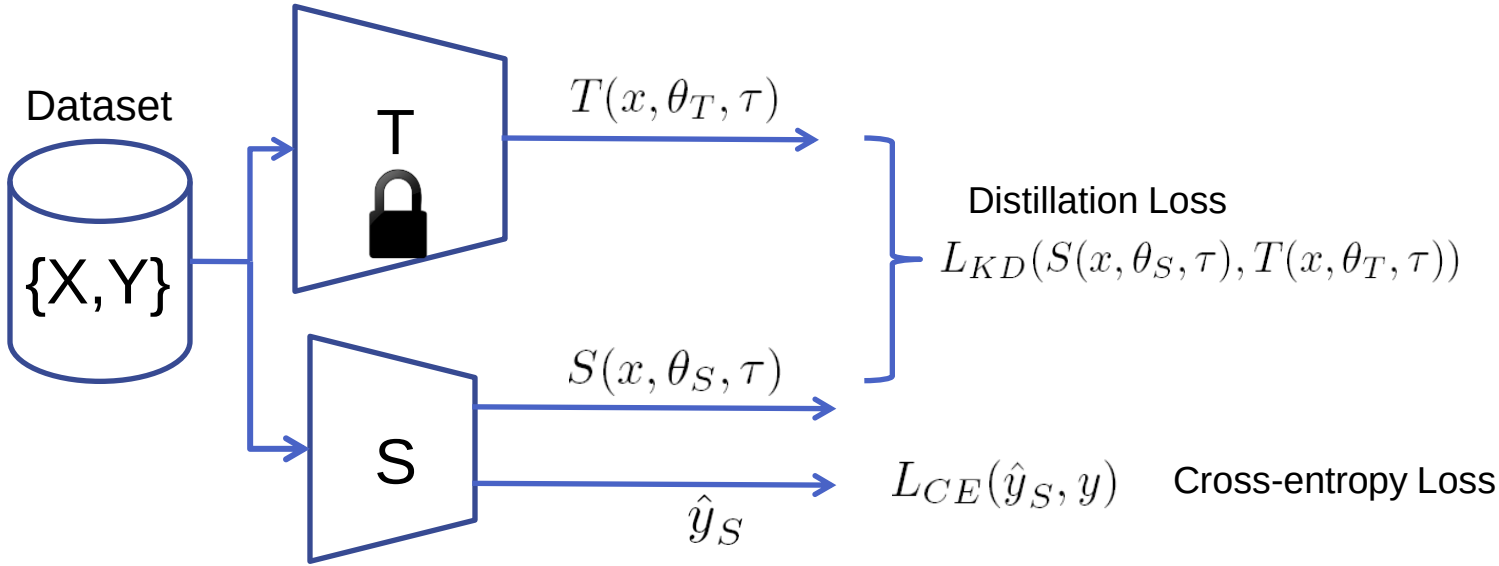
[1]Indian Institute of Science, [2]University of Edinburgh, [3]University of Lincoln

CDS
Department of Computational and Data Sciences

Objective

- **Can we do *Knowledge Distillation* without (access to) training data (Zero-Shot)?**
  - Data is precious and sensitive – won't be shared
  - E.g. : Medical records, Biometric data, Proprietary data
  - Federated learning – Only models are available, not data

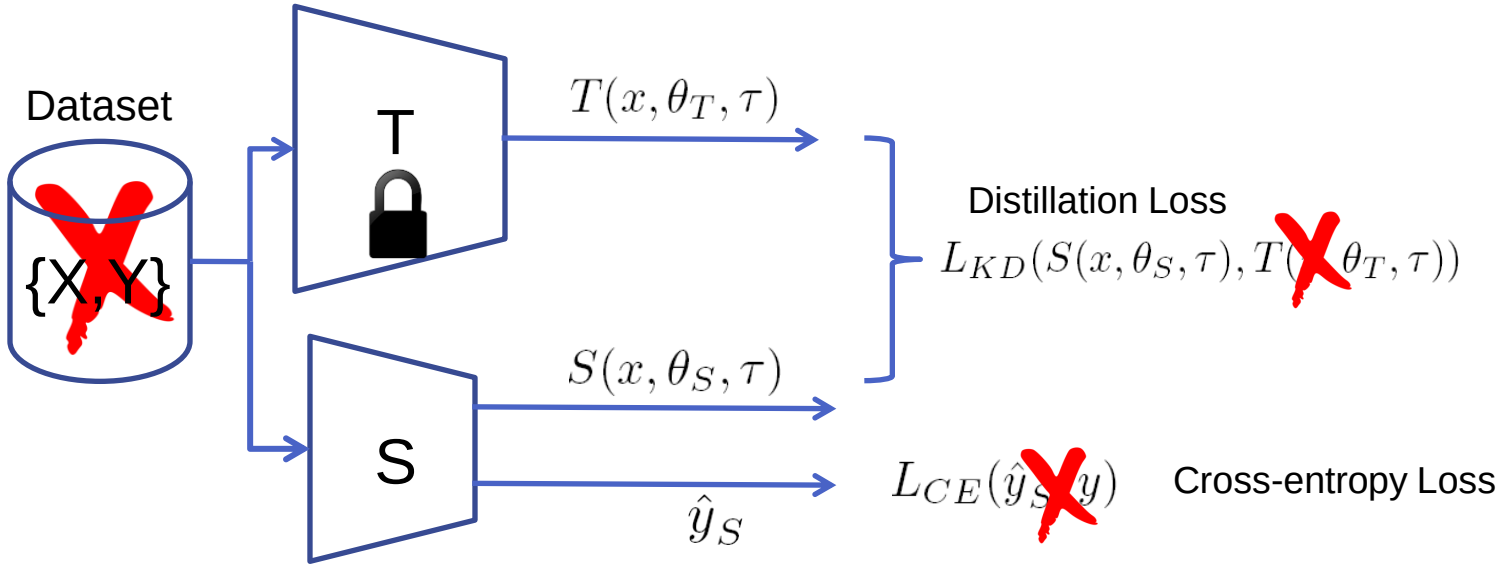
Knowledge Distillation



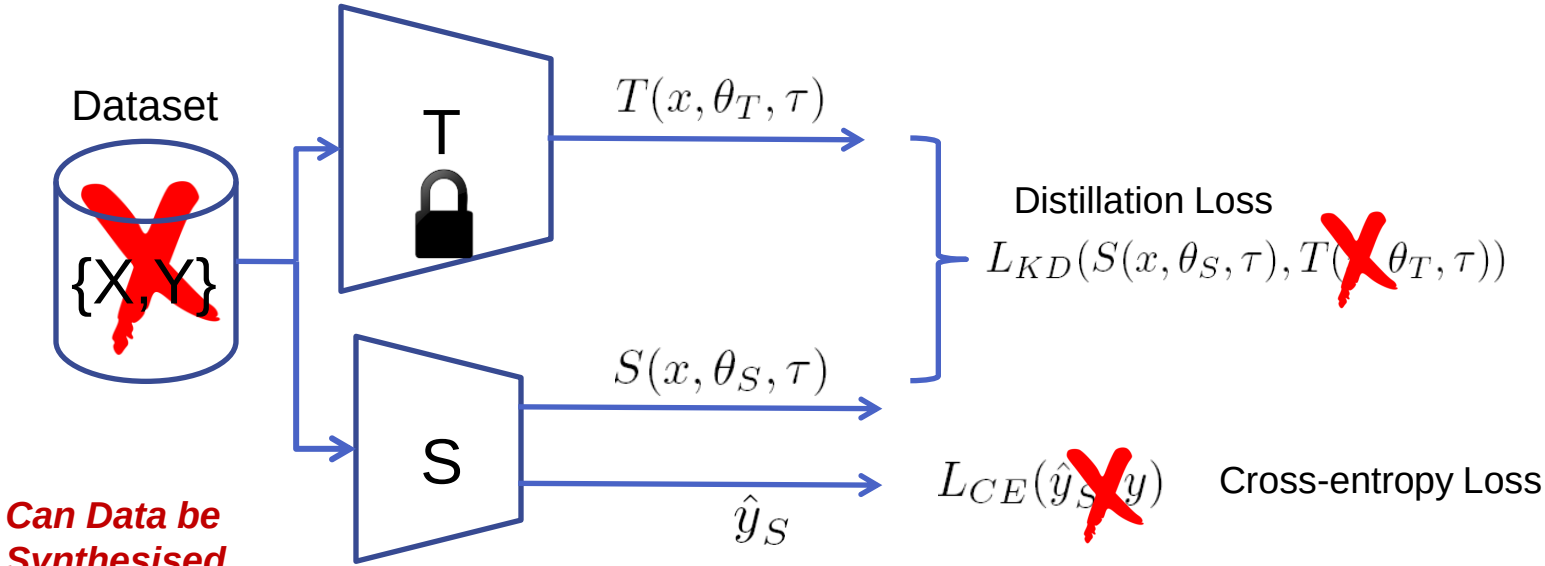
$$L = \sum_{(x,y)\in\mathbb{D}} L_{KD}(S(x, \theta_S, \tau), T(x, \theta_T, \tau)) + \lambda L_{CE}(\hat{y}_S, y)$$

Hinton et al. Distilling the Knowledge in a Neural Network, arXiv:1503.02531, 2015

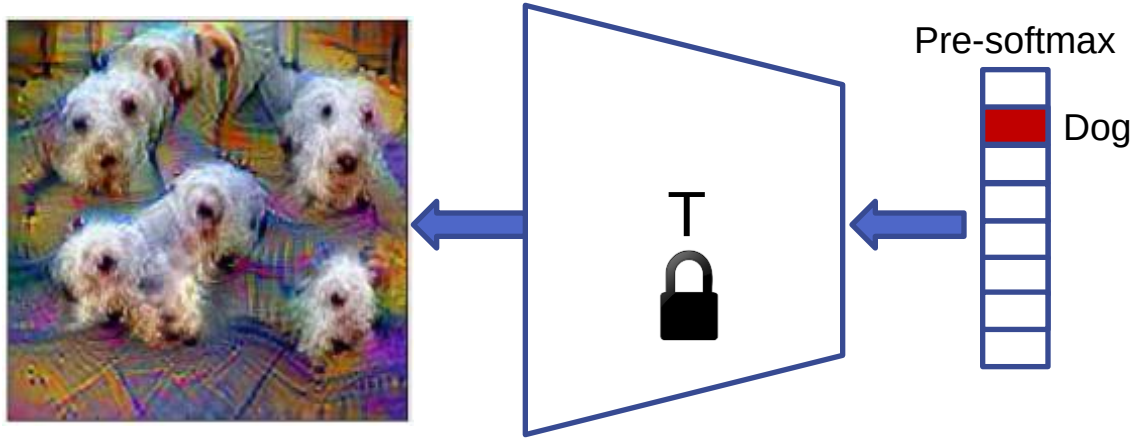
Data Free Knowledge Distillation



Data Free Knowledge Distillation



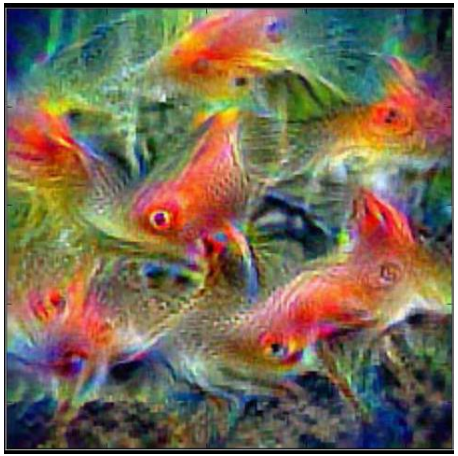
Pseudo Data Synthesis: *Class Impressions (CI)*



$$CI_c = \underset{x}{\text{argmax}} \quad f_c^{ps/m}(x)$$

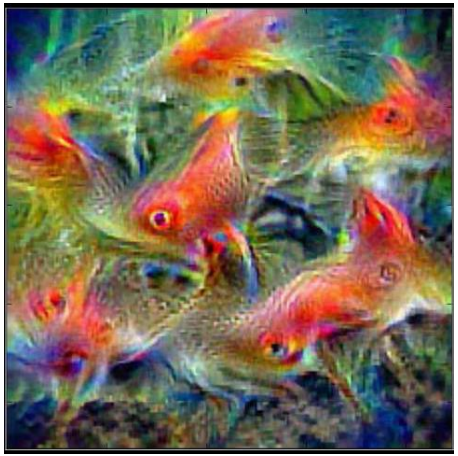
Mopuri et al., Ask, Acquire and Attack: Data-free UAP generation using Class impressions, ECCV'18

Pseudo Data Synthesis: *Class Impressions (CI)*



Mopuri et al., Ask, Acquire and Attack: Data-free UAP generation using Class impressions, ECCV'18

Pseudo Data Synthesis: *Class Impressions (CI)*



Goldfish

Squirrel Monkey

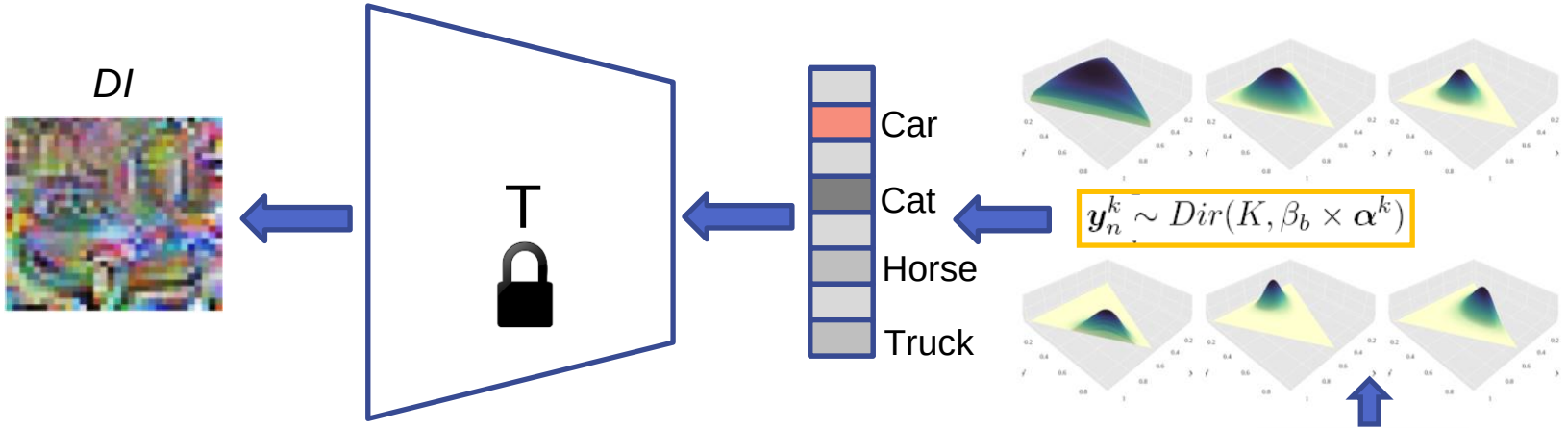
Cock

Mopuri et al., Ask, Acquire and Attack: Data-free UAP generation using Class impressions, ECCV'18

Class Impressions (CI): Limitations

- Generated samples are *less diverse*
- *Relative probabilities* of incorrect classes are not considered
- Student does not *generalize* well when trained on CIs

Data Impressions (DI)

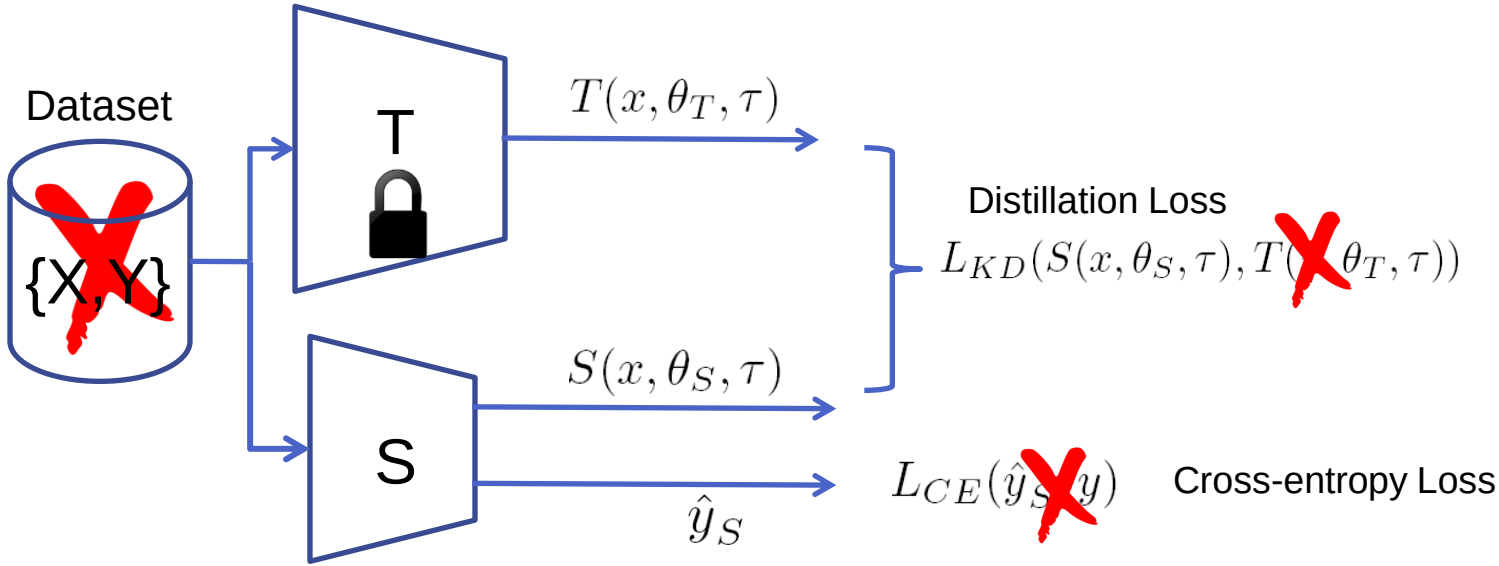


$$\bar{x_i}^k = \underset{x}{\operatorname{argmin}}\, L_{CE}(\boldsymbol{y}_i^k, T(x, \theta_T, \tau))$$

$$C(i,j) = \frac{\boldsymbol{w}_i^T \boldsymbol{w}_j}{\|\boldsymbol{w}_i\| \|\boldsymbol{w}_j\|}$$

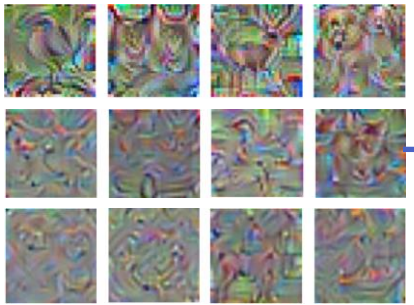
Class similarity matrix

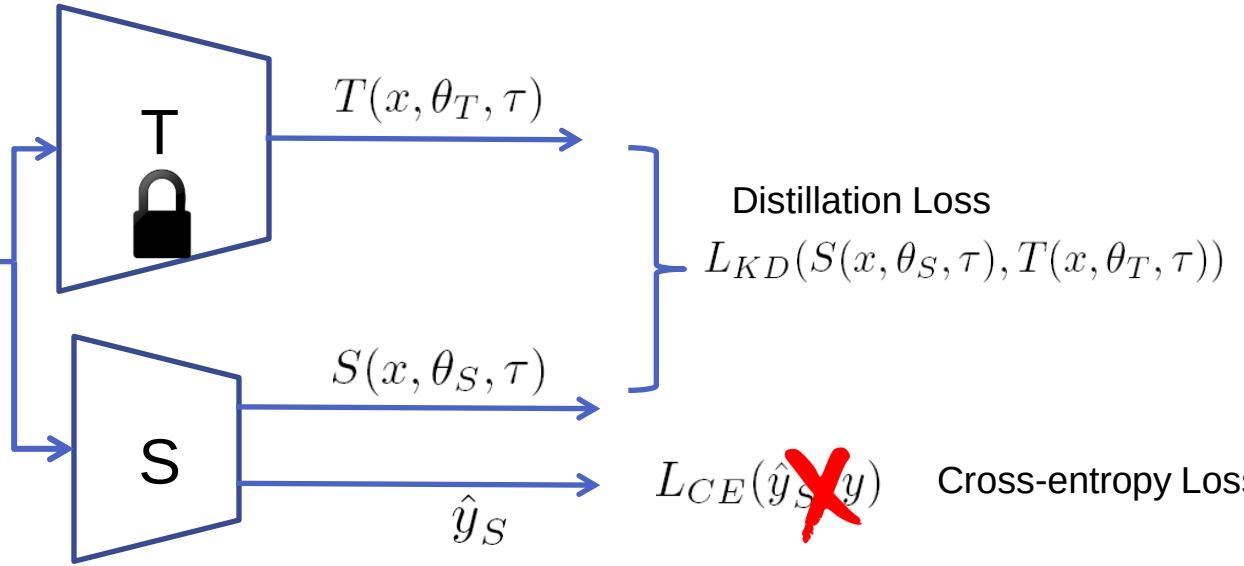
Data Free Knowledge Distillation



Data Free Knowledge Distillation

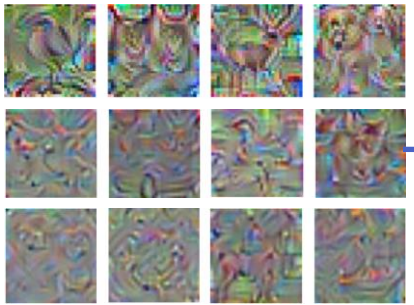
Data Impressions (*DI*)

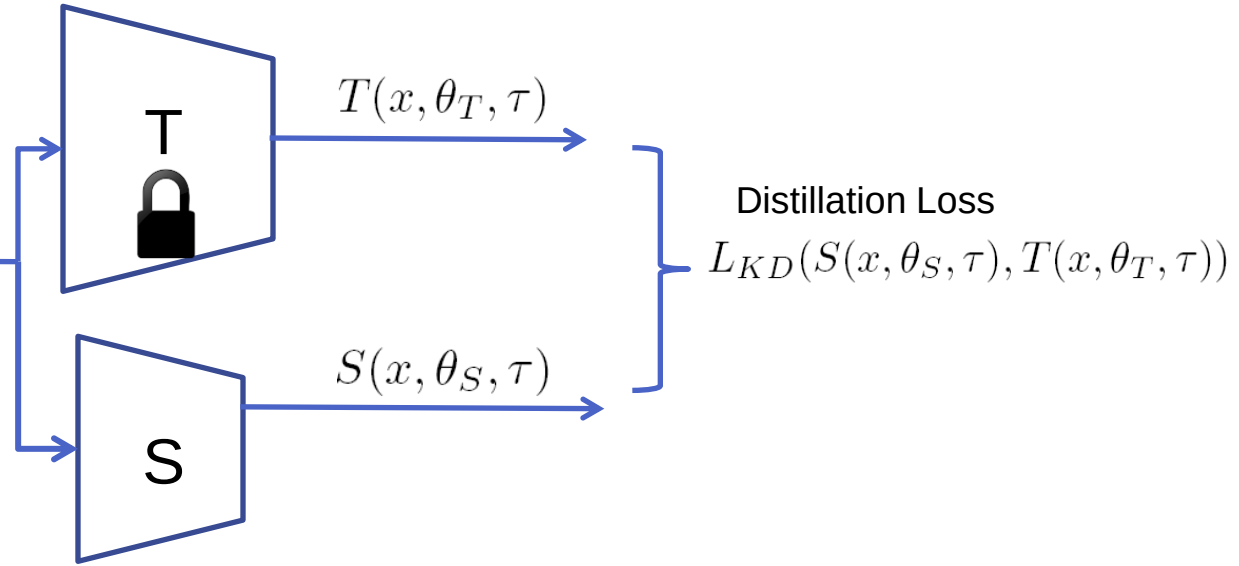




Data Free Knowledge Distillation

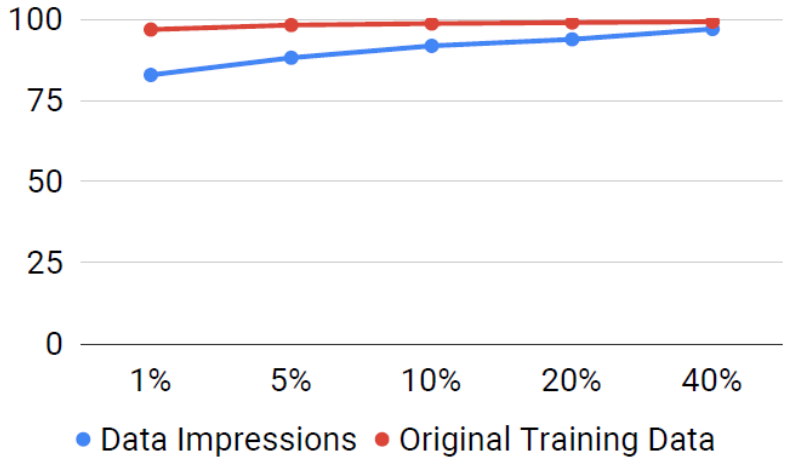
Data Impressions (*DI*)

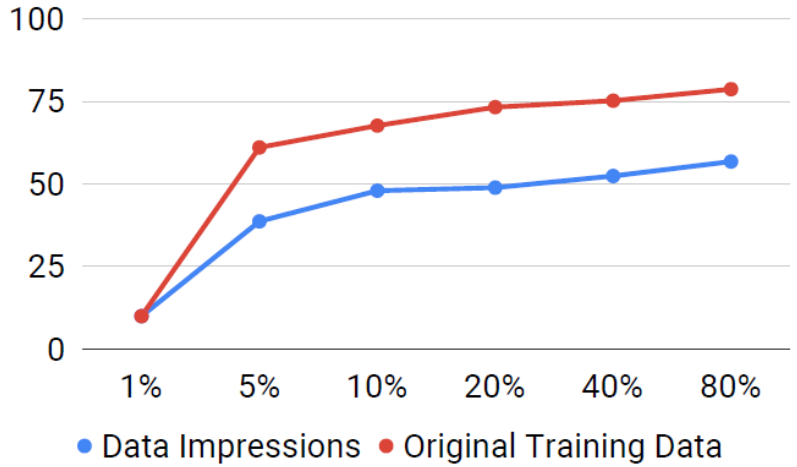




$$\theta_S = \underset{\theta_S}{\operatorname{argmin}} \sum_{\bar{x} \in \bar{X}} L_{KD}(T(\bar{x}, \theta_T, \tau), S(\bar{x}, \theta_S, \tau))$$

Results: MNIST and CIFAR-10

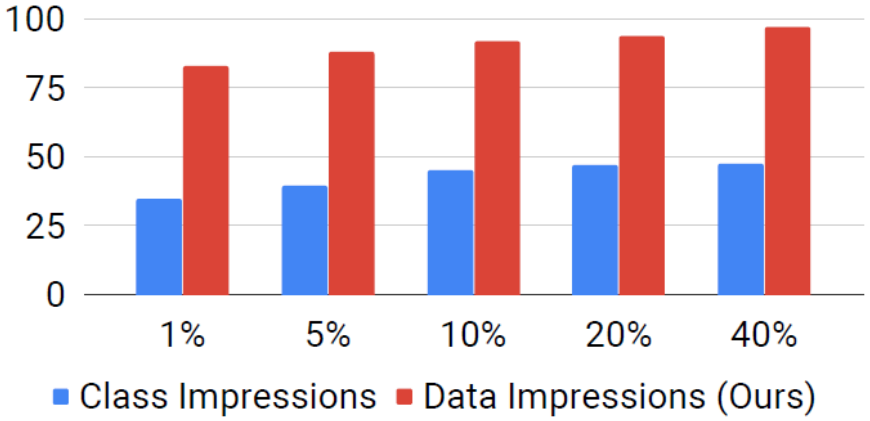




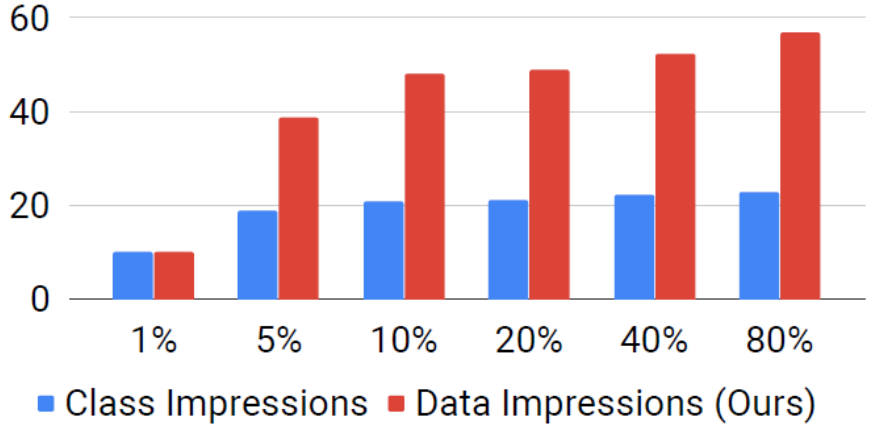
MNIST T: LeNet S: LeNet-Half

CIFAR-10 T: AlexNet S: AlexNet-Half

CI Vs *DI* : MNIST and CIFAR-10



MNIST



CIFAR-10

Results: Comparison

Model	Performance
Teacher – CE	99.34
Student – CE	98.92
Student–KD (Hinton et al., 2015) 60K original data	99.25
(Kimura et al., 2018) 200 original data	86.70
(Lopes et al., 2017) (uses meta data)	92.47
ZSKD (Ours) (24000 *DI*s, and no original data)	98.77

MNIST

Model	Performance
Teacher – CE	83.03
Student – CE	80.04
Student – KD (Hinton et al., 2015) 50K original data	80.08
ZSKD (Ours) (40000 *DI*s, and no original data)	69.56

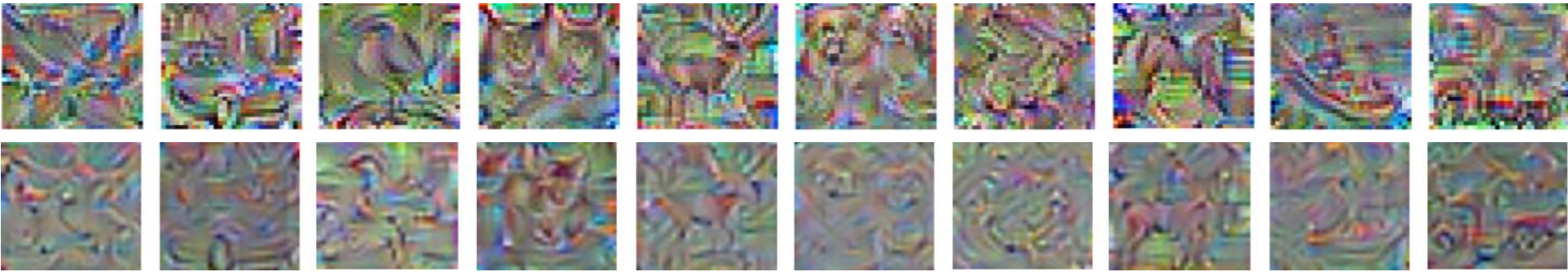
CIFAR-10

Recent works along this direction

- Micaelli P, Storkey A. **Zero-shot Knowledge Transfer via Adversarial Belief Matching**. arXiv preprint arXiv:1905.09768. 2019 May 23.
- Chen H, Wang Y, Xu C, Yang Z, Liu C, Shi B, Xu C, Xu C, Tian Q. **Data-Free Learning of Student Networks**. arXiv preprint arXiv:1904.01186. 2019 May 29.

Summary

- For the first time we have proposed a Zero-Shot KD approach
- The effectiveness of the *Data Impressions* is demonstrated by training a student network from scratch.
- Hope our ZSKD can inspire researchers to explore more interesting dimensions and applications in this area.



Thanks! Please Visit Our Poster (#74)