Zero-Shot Knowledge Distillation in Deep Networks

Gaurav Kumar Nayak*, Konda Reddy Mopuri*, Vaisakh Shaj*, R. Venkatesh Babu, Anirban Chakraborty

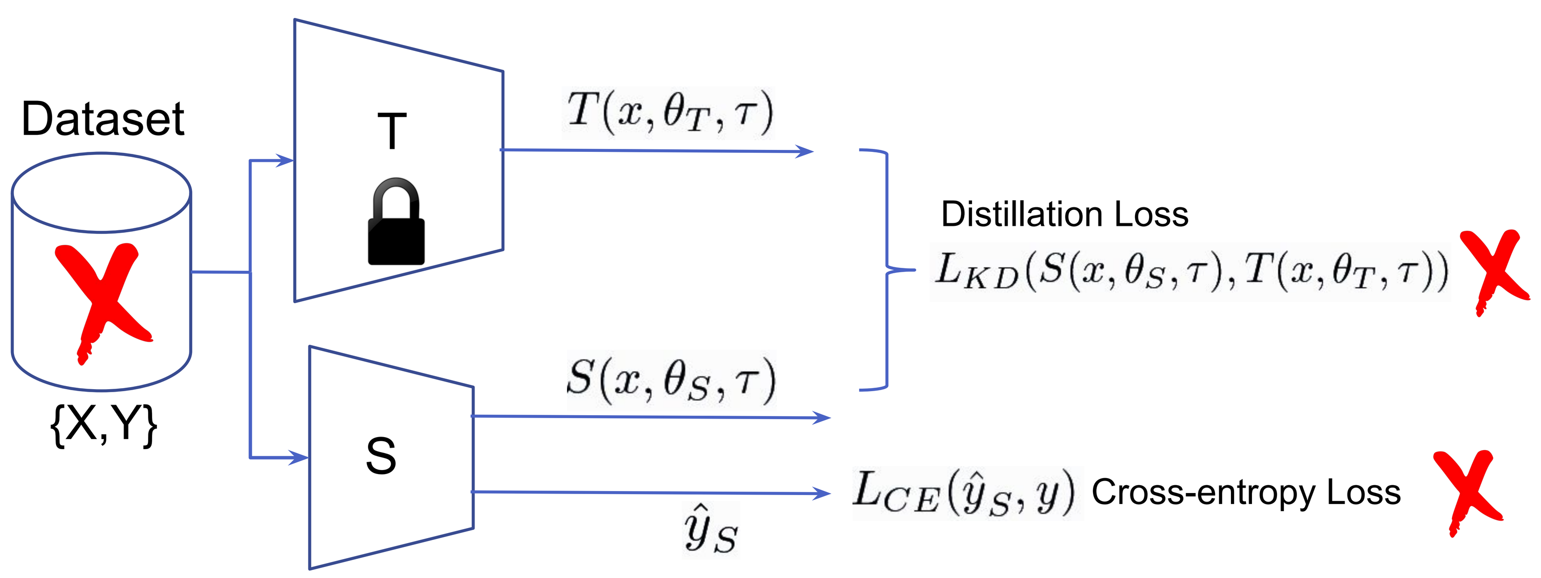
Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

ICML International Conference On Machine Learning 2019

Overview

Objective

- To do *Knowledge Distillation* without training data.
  - 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.



Prior Works

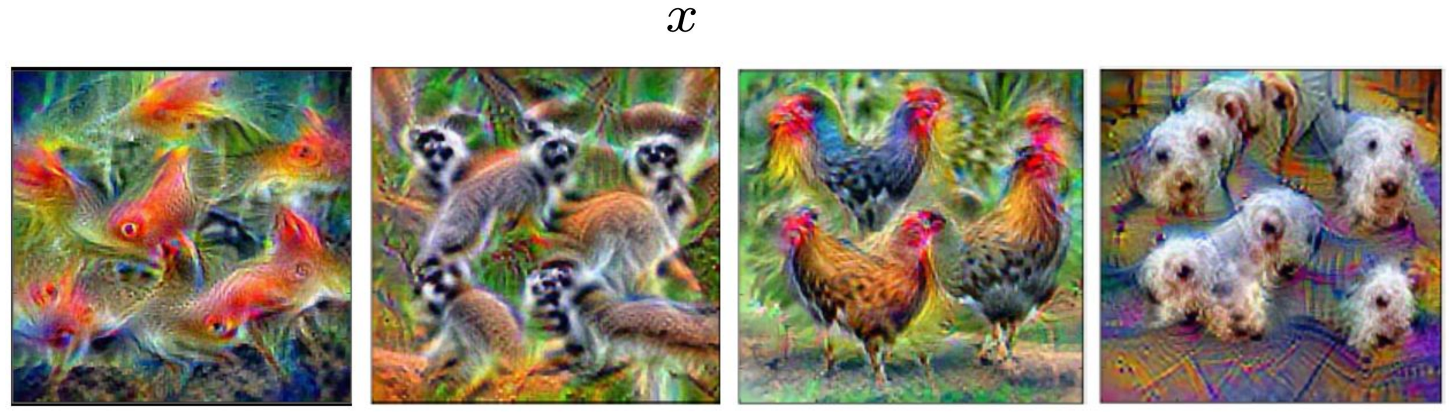
Based on the **number of training samples**:

- **Using full training data**: Matching of softmax values that are raised to a high temperature [Hinton et al., 2015].
- **Using few training samples**: Pseudo samples are augmented with few samples of training data, used to train the Student network. [Kimura et al., 2018].
- **Using Meta-data**: Uses Precomputed activation records as meta data to construct training samples [Lopes et al., 2017].

Pseudo Data Synthesis: (Class Impressions)

- The pretrained models have memory in terms of learned parameters and can be used to extract class representative samples [Mopuri et al., 2018].

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

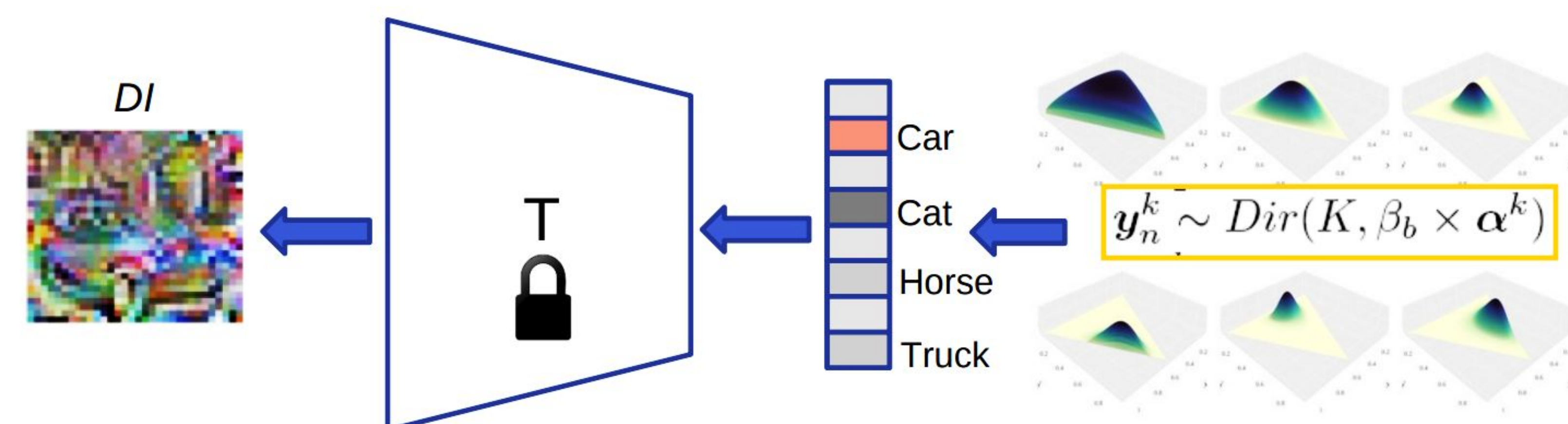


Limitations:

- Less Diverse.
- *Relative probabilities* of incorrect classes are not considered.
- Student does not *generalize* well when trained on Cis.

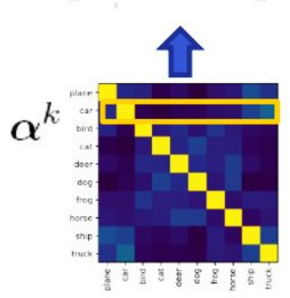
Approach

- We tap the memory (**learned parameters**) of the *Teacher* and synthesize pseudo samples, naming as *Data Impressions* (DI).
- Let $\boldsymbol{s} \sim p(\boldsymbol{s})$ be the random vector that represents the neural softmax outputs of the *Teacher*, $T(x,\theta_T)$. We model $p(\boldsymbol{s}^k)$ for each class *k*, using a **Dirichlet distribution.**



$$\bar{x}_i^k = \underset{x}{\text{argmin}}\ L_{CE}(\mathbf{y}_i^k, T(x,\theta_T,\tau)) \qquad C(i,j) = \frac{\mathbf{w}_i^T\mathbf{w}_j}{\|\mathbf{w}_i\|\|\mathbf{w}_j\|}$$

$$k \in \{1\ldots K\} \qquad \alpha^k = [\alpha_1^k,\alpha_2^k,\ldots,\alpha_K^k], \text{ and } \alpha_i^k > 0, \forall i$$

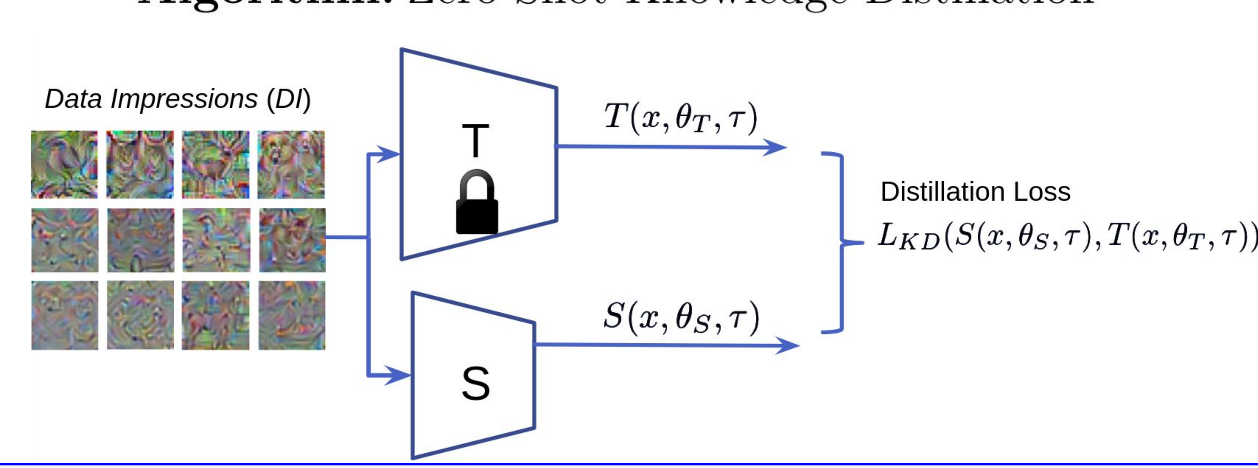


Class similarity matrix

- K is the count of total classes, $\boldsymbol{\alpha}^k$ is the concentration parameter of the distribution modelling class *k* and β scales the concentration parameter to model the spread of the Dirichlet distribution.

Input : *Teacher* model T; N: number of DIs crafted per category; $[\beta_1,\beta_2,...,\beta_B]$: B scaling factors; τ: Temperature for distillation
Output : Learned *Student* model $S(\theta_S)$; $\bar{X}$: *Data Impressions*
Obtain K: number of categories from T
Compute the class similarity matrix : $C = [\mathbf{c}_1^T,\mathbf{c}_2^T,\ldots,\mathbf{c}_K^T]$
$\bar{X} \leftarrow \emptyset$
for k=1:K **do**
  Set the concentration parameter $\boldsymbol{\alpha}^k = \mathbf{c}_k$
  for b=1:B **do**
    for n=1:$\lfloor N/B \rfloor$ **do**
      Sample $\boldsymbol{y}_n^k \sim Dir(K,\beta_b\times\boldsymbol{\alpha}^k)$
      Initialize $\bar{x}_n^k$ to random noise and craft
      $\bar{x}_n^k = \underset{x}{\text{argmin}}\ L_{CE}(\boldsymbol{y}_n^k, T(x,\theta_T,\tau))$
      $\bar{X} \leftarrow \bar{X} \cup \bar{x}_n^k$
    end
  end
end
Transfer the *Teacher*'s knowledge to *Student* using the *DIs* via
$\theta_S = \underset{\theta_S}{\text{argmin}} \sum_{\bar{x}\in\bar{X}} L_{KD}(T(\bar{x},\theta_T,\tau),S(\bar{x},\theta_S,\tau))$

Algorithm: Zero-Shot Knowledge Distillation



Results

- β values: a mix of 0.1 and 1.0 encourages higher diversity (variance) and at the same time does not result in highly sparse vectors.

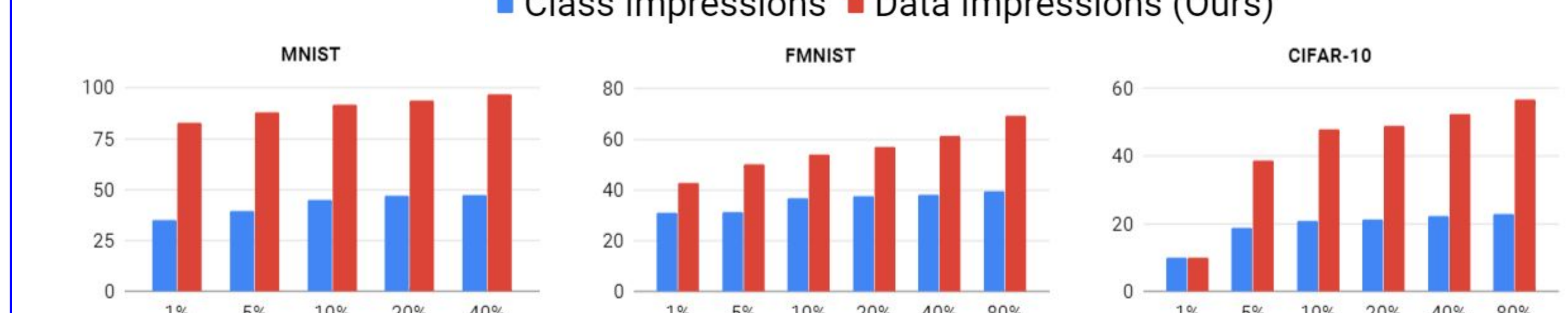
Dataset	Teacher	Student
MNIST	Lenet	Lenet-Half
Fashion MNIST (FMNIST)	Lenet	Lenet-Half
CIFAR-10	Alexnet	Alexnet-Half

Model on **CIFAR-10**	Acc.
Teacher – CE	83.03
Student – CE	80.04
Student – KD (Hinton et al., 2015) 50K original data	80.08
ZSKD (Ours) (40000 *DIs*, and no original data)	69.56

Model on **MNIST**	Acc.
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 (24000 *DIs*, no original data)	98.77

Model on **FMNIST**	Acc.
Teacher – CE	90.84
Student – CE	89.43
Student – KD (Hinton et al., 2015) 60K original data	89.66
(Kimura et al., 2018) 200 original data	72.50
ZSKD (48000 *DIs*, no original data)	79.62

Performance of ZSKD on DIs v/s Original Training data



Performance of ZSKD on CI v/s DI

References

- Hinton et al., Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
- Kimura et al., Few-shot learning of NN from scratch by pseudo example optimization. In BMVC, 2018.
- Lopes et al., Data-free knowledge distillation for deep neural networks. In NIPS Workshop
- Mopuri et al., AAA: Data-free uap generation using class impressions. In ECCV, 2018.

Acknowledgements

This work was partially supported by Tata Trust grant.

* Equal contribution