

Efficient Online Subclass Knowledge Distillation for Image Classification

M. Tzelepi, N. Passalis and A. Tefas

E-mails: mtzelepi@csd.auth.gr, passalis@csd.auth.gr, tefas@csd.auth.gr

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece





Knowledge Distillation

- Knowledge Distillation (KD) has been established as a highly promising approach for **training compact and faster models** by **transferring knowledge** from more complex powerful models
- However, KD in its conventional version constitutes an **enduring**, **computationally** and **memory demanding** process
- Thus, many **online KD** approaches have recently been proposed



Online Knowledge Distillation

- Online KD describes the process where the teacher and student networks are trained **simultaneously**, without requiring a separate stage for pre-training the teacher network
- Current online KD works propose to train multiple models mutually from each other, or to create ensembles of multiple identical branches of a target network in order to build a strong teacher and distill the knowledge to the target network



Motivation

- In conventional KD methods it is manifested that it is advantageous for each sample to **maintain its similarities with the other classes**, instead of merely training with the hard labels
- In this work, we considered that inside each class there is also a set of **sub-classes** that **share semantic similarities** (e.g., blue cars, inflatable boats, etc.)
- The sub-classes inside each class are unknown and thus, we proposed to **estimate** them using the **neighborhood** of each sample



Motivation

- We assume that the **nearest neighbors** of each sample inside a class belong to the **same sub-class**
- The model is trained **synchronously** both with the conventional supervised loss (hard labels) and the **soft labels** so as to **maintain these sub-class similarities**



Proposed Method

- We proposed a novel online self distillation method, namely **Online Subclass Knowledge Distillation** (OSKD), that distills additional knowledge **online** from the **model itself** throughout the network's training
- We introduced an additional **distillation objective** which encourages the data representations to come closer to the nearest representations of the same class and concurrently to move further away from the nearest representations of the other classes



Proposed Method

• Considering an input vector \mathbf{y}_{i} , a neural network $\phi(\cdot, W)$ with a set of parameters W, and the output vector of \mathbf{y}_{i} given the network $\phi \phi(\mathbf{y}_{i}, W)$ the additional distillation objective is formulated as follows:

$$\min_{\mathcal{W}} \mathcal{J}_1 = \min_{\mathcal{W}} \sum_{\boldsymbol{y}_i, \boldsymbol{y}_j \in \mathcal{R}^i} \|\phi(\boldsymbol{y}_i; \mathcal{W}) - \phi(\boldsymbol{y}_j; \mathcal{W})\|_2^2 \qquad \max_{\mathcal{W}} \mathcal{J}_2 = \max_{\mathcal{W}} \sum_{\boldsymbol{y}_i, \boldsymbol{y}_l \in \mathcal{V}^i} \|\phi(\boldsymbol{y}_i; \mathcal{W}) - \phi(\boldsymbol{y}_l; \mathcal{W})\|_2^2.$$

• The above equations can also be formulated as follows: $\min_{\mathcal{W}} \mathcal{J}_1 = \min_{\mathcal{W}} \sum_{\boldsymbol{y}_i \in \mathcal{R}^i} \|\phi(\boldsymbol{y}_i ; \mathcal{W}) - \boldsymbol{\mu}_r^i)\|_{2^1}^2 \qquad \max_{\mathcal{W}} \mathcal{J}_2 = \max_{\mathcal{W}} \sum_{\boldsymbol{y}_i \in \mathcal{V}^i} \|\phi(\boldsymbol{y}_i ; \mathcal{W}) - \boldsymbol{\mu}_v^i)\|_2^2$

where μ_r and μ_u correspond to the mean vectors of the nearest representations of the same class and the nearest representations of the other classes, respectively

Proposed Method



- Thus the overall distillation loss is formulated as: $J_{oskd} = J_1 + (1-J_2)$
- Therefore, in the proposed distillation training process we seek for the parameters W* that minimize the overall loss of cross entropy, J_{ce}, and distillation, J_{oskd}:

$$\mathcal{W}^* = \underset{\mathcal{W}}{\operatorname{arg\,min}} \sum_{i=1}^{N} [J_{ce}(c_i, \phi(y_i; \mathcal{W})) + \lambda J_{oskd}(\mu_r^i, \mu_v^i, \phi(y_i; \mathcal{W}))]$$



Experimental Evaluation



MNIST (odd vs even digits with 3 subclasses per class): LDA Visualization



Experimental Evaluation



Test accuracy for different numbers of nearest neighbors inside each class

Method	Cifar-10	SVHN-10
W/o Distillation	$64.83\% \pm 0.57\%$	$88.82\% \pm 0.21\%$
OSKD - 2NN	$66.16\% \pm 0.76\%$	$89.00\% \pm 0.14\%$
OSKD - 4NN	$66.39\% \pm 0.77\%$	$89.52\% \pm 0.23\%$
OSKD - 8NN	$66.59\% \pm 0.78\%$	$89.61\% \pm 0.29\%$
OSKD - 12NN	$67.36\% \pm 0.82\%$	$89.67\% \pm 0.28\%$

Comparisons against existing online distillation methods using WRN 16-2

Method	Test Accuracy
WRN 16-2	$93.55\% \pm 0.11\%$
ONE	$93.76\% \pm 0.16\%$
FFL-S	$93.79\% \pm 0.12\%$
ONE-E	$93.84\% \pm 0.20\%$
FFL	$93.86\% \pm 0.11\%$
OSKD	$93.96\% \pm 0.13\%$

Acknowledgements

This research is co-financed by Greece and the European Union (European Social Fund - ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020" in the context of the project "Lightweight Deep Learning Models for Signal and Information Analysis" (MIS 5047925).



Επιχειρησιακό Πρόγραμμα Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση



Ευρωπαϊκή Ενωση Ευρωπαϊκό Κοινωνικό Ταμείο Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης Thank you!