SRP PROPOSAL

Knowledge Distillation in Encoders

STUDENTS : Wut Hmone Hnin Hlaing (1748517), Steffen Roeber (1748613)
Apurv Kumar (1748394), Abin Baby (1748437), Brhanu Atsbaha (1748398)

SUPERVISORS:
Prof. Dr. Dr. Lars Schmidt-Thieme
Diego Coello de Portugal, M.Sc

Date: 15/03/2024

Problem Setting

In the realm of machine learning, the quest for more efficient and lightweight models is a crucial endeavor. As we aim to deploy deep learning algorithms in resource-constrained environments like mobile devices and edge computing devices, we encounter significant challenges due to computational limitations. Traditional deep learning models, while powerful, often require substantial computational resources and memory. This poses a bottleneck for real-world deployment, where efficiency is paramount. Hence, the motivation arises for methods to compress and distill the knowledge within these models into smaller, more efficient versions while maintaining their performance. Enter knowledge distillation.

Knowledge distillation (KD) is a process in machine learning where a smaller, more computationally efficient model (student) is trained to mimic the behavior of a larger, more complex model (teacher). This is achieved by leveraging the rich knowledge embedded in the teacher model's parameters. In the vanilla KD setting the student model not only learns from the ground truth labels but also from the soft labels provided by the teacher model. By mimicking the teacher's predictions, the student can capture the essence of the teacher's knowledge without requiring the same computational resources. This results in a compressed model that retains much of the performance of the larger model but is more suitable for deployment in constrained environments. Thus, knowledge distillation serves as a powerful technique to bridge the gap between model performance and computational efficiency, making deep learning models more practical and scalable for real-world applications.

State of the art

In knowledge distillation, knowledge types that represent the knowledge that is being transferred to the student models, distillation strategies which are also known as training or learning schemes for both teacher and student models, and the teacher–student models architectures play a crucial role in student learning Gou et al. [2021].

Different categories of knowledge types for knowledge distillation have been explored

in the literature. The three most widely used knowledge types are response-based knowledge Hinton et al. [2015], feature-based knowledge Bengio et al. [2013], and relation-based knowledge Yim et al. [2017]. Response-based knowledge usually refers to the neural response of the last output layer of the teacher model. Response-based models often employ Kullback-Leibler divergence loss and cross-entropy loss for student models. The main idea is to directly mimic the final prediction of the teacher model which is similar to vanilla knowledge distillation Hinton et al. [2015]. Feature-based knowledge utilizes representation learning since deep neural networks excel in acquiring various levels of feature representation, progressively abstracting information Bengio et al. [2013]. Hence, both the final layer's output and the intermediate layers' output, referred to as feature maps, are viable options for supervising the training of the student model. Most of the loss functions used in feature-based learning are $l2 - norm$ distance, $l1 - norm$ distance, cross-entropy loss, and maximum mean discrepancy loss Hinton et al. [2015], Gou et al. [2021]. Both response-based and feature-based knowledge use the outputs of specific layers in the teacher model. Relation-based knowledge further explores the relationships between different layers (hint layers) or data samples Yim et al. [2017]. To train with relation-based knowledge distillation the following Earth Mover distance, Huber loss, angle-wise loss, and Frobenius norm are utilized Gou et al. [2021].

According to whether the teacher model is updated simultaneously with the student model or not, the learning schemes of knowledge distillation can be directly divided into three main categories: offline distillation, online distillation, and self-distillation. Most of the previous knowledge distillation methods work offline Hinton et al. [2015]. The offline methods mainly focus on improving different parts of the knowledge transfer, including the design of knowledge and the loss functions for matching features or distributions matching Adriana et al. [2015]. Although offline distillation methods are simple and effective, they have some limitations. Hence, researchers have proposed online distillation to address these shortcomings Mirzadeh et al. [2020]. In online distillation, updates occur concurrently for both the teacher and student models, enabling the entire knowledge distillation framework to be trained end-to-end. In self-distillation, the same networks are used for the teacher and the student models Zhang and Peng [2018].

In knowledge distillation, the teacher-student architecture serves as a versatile framework for facilitating knowledge transfer. Effectively capturing and distilling knowledge in knowledge distillation requires careful selection or design of appropriate structures for both the teacher and student networks Hinton et al. [2015]. This task is crucial yet challenging. The typical selection for the student network involves either: simplifying the architecture of the teacher network by reducing the number of layers and channels per layer, or maintaining the structure of the network while quantizing it Wang et al. [2018], Zhu et al. [2018], Polino et al. [2018], Howard et al. [2017].

Besides, with the assumption that knowledge distillation bears a resemblance to human learning, recent advancements in knowledge distillation have expanded to encompass various approaches that primarily focus on compressing deep neural networks such as teacher-student learning Hinton et al. [2015], mutual learning Zhang and Peng [2018], assistant teaching Mirzadeh et al. [2020], lifelong learning, Zhai et al. [2019] and self-learning Yuan et al. [2019].

Data Foundation

Following Chen et al. [2022] we will use the CIFAR-100 dataset (Krizhevsky [2009]) to do our experiments on. CIFAR-100 is a standard benchmark dataset consisting of 60000 32x32 colour images with a total of 100 target classes, split into train and test datasets with 50k and 10k images respectively. We will also use the CIFAR-10 (Krizhevsky [2009]) dataset which has 10 target classes for our initial simple experiments.

Research Idea

Our project will be based on a study by Chen et al. [2022]. They apply a feature-based approach, where they try to align the last representation layer of the teacher model with that of the student model, using an l_2-loss. To compare the two representations, they use a projector to upscale the feature dimension of the student model to that of the teacher model. Crucially, they use the pre-trained classification layer of the teacher model for the student model as well.

We plan to further explore the following aspects with our experiment:

Different loss functions: Since Chen et al. [2022] only used a simple l_2-loss, we want to experiment with other loss functions. Experimentation with novel functions can enhance the effectiveness of knowledge distillation; this could involve exploring alternatives to standard loss functions like mean squared error or cross-entropy loss, such as attention-based losses or adversarial losses tailored for distillation tasks.

Representation learning: Chen et al. [2022] only focus on aligning the feature representations of the last layers between the teacher and student model. We want to explore aligning the features in earlier layers as well. Aligning the features already in earlier layers could improve the performance of the student model by receiving additional guidance from the teacher model.

Ablation Studies: We also plan to do ablations on e.g. different projector architectures, different losses, and how to weigh them (earlier vs. later projections), down-scaling the feature representation of the teacher to that of the student or meeting somewhere in the middle.

Tangible Outcomes

In case our methods beat the state-of-the-art, we plan to write a research paper and submit it to a conference. Otherwise, we will write a project report about the experiments we conducted and the results we acquired.

Work Plan

We first want to implement the approach by Chen et al. [2022] and run some baseline experiments on the cluster of the university. The results of that should already be part of the first interim presentation, so it should be ready by June. All the other things we want

to implement rely on that, so it should be done first. The original code by Chen et al. [2022] is available online and written in Python using PyTorch, so we will also implement our experiments in the same framework.

Once we handled the baseline and got used to working on the cluster, we would like to implement and run experiments on our new ideas. The easiest and fastest one to implement should be experimenting with different losses. Next, we want to work on combining knowledge distillation with representation learning and using multiple projectors between the teacher and the student model. While the experiments on the losses should be ready to be presented on the 1st interim presentation, we plan to be done with experiments on representation learning by the 2nd and with the ablation studies by the 3rd and last presentation. The last task is to write a report or paper, depending on the outcome of our experiment.

To sum up, our list of tasks will look like this:

Task	**Time-frame**
Implement baseline experiment	April 2024
Implement different losses experiments	May 2024
First presentation preparation	May 2024
Implement representation learning experiments	July 2024
Second presentation preparation	October 2024
Ablation study	November 2024
Final presentation preparation	December 2024
Final report/paper writing	March 2025

Resources

We will need access to the cluster of the university in order to store the training data, train our models, and run the experiments.

Team Members

- **Steffen Roeber** (roebers@uni-hildesheim.de)
  Python, Computer Vision, CNNs, TensorFlow
- **Wut Hmone Hnin Hlaing** (hlaing@uni-hildesheim.de)
  Python, PyTorch, MLOps, Docker, Kubernetes
- **Apurv Kumar** (kumara@uni-hildesheim.de)
  Python, PyTorch, Computer Vision, CNNs
- **Abin Baby** (baby@uni-hildesheim.de)
  Python, PyTorch, Tableau
- **Brhanu Atsbaha** (atsbaha@uni-hildesheim.de)
  Python, TensorFlow

References

Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. Fitnets: Hints for thin deep nets. *Proc. ICLR*, 2(3):1, 2015.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier, 2022.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018.

Mengjiao Wang, Rujie Liu, Narishige Abe, Hidetsugu Uchida, Tomoaki Matsunami, and Shigefumi Yamada. Discover the effective strategy for face recognition model compression by improved knowledge distillation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2416–2420. IEEE, 2018.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisit knowledge distillation: a teacher-free framework. 2019.

Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2759–2768, 2019.

Chenrui Zhang and Yuxin Peng. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. *arXiv preprint arXiv:1804.10069*, 2018.

Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31, 2018.