Multistage Collaborative Knowledge Distillation on Semantic Segmentation

Sophia Balestrucci [1] **Chiara Ballanti** [2] **Paolo Pio Bevilacqua** [3] **Ilaria Gagliardi** [4]

Abstract

Semantic segmentation is a widely studied task in the field of computer vision. However, in many applications, a frequent obstacle is the lack of labeled images, because acquiring dense annotations of images is labor-intensive and time-consuming. In this project, we will investigate a novel distillation approach recently proposed, *Multistage Collaborative Knowledge Distillation* (MCKD) (Zhao et al., 2023), for semi-supervised sequence prediction. Inspired by the results obtained in the original paper, we propose to study the effectiveness of this technique on image semantic segmentation. Code available on github.

1. Introduction

In this project, we study a challenging semi-supervised semantic image segmentation scenario where labeled data are too few to finetune a model. We first collect pseudo-labels for a large amount of unlabeled data from a teacher. Then we perform multistage knowledge distillation. It consists of splitting the pseudo-labeled data into two partitions and performing cross-partition labeling (Chen et al., 2021). At each stage, a pair of students are trained on distinct partitions of pseudo-labeled data and produce new pseudo-labels for the data that they have not been trained on. In the final distillation stage, a single student is trained on all the latest pseudo-labeled data.

2. Related work

Knowledge distillation (Hinton et al., 2015) has proven to be a powerful technique for transferring knowledge from a complex model (teacher) to a simpler one (student) (Stanton et al., 2021; Cho & Hariharan, 2019; Beyer et al., 2022). In recent years, numerous researchers explored the application of KD in the domain of SSL semantic image segmentation. The earliest proposed technique investigates a *consistency loss* comparing regional differences in student and teacher output (Xie et al., 2018). Instead of a direct matching of student and teacher output, in *Knowledge Adaptation* (KA) (He et al., 2019) teacher output is compressed to a more dense latent space by an autoencoder before being compared to student logits. *Structured Knowledge Distillation* (SKD) (Liu et al., 2019) is the most cited publication in the field and uses a combination of three loss terms to focus more on contextual information in both intermediate and output layers. Another suggested approach, *CSCACE* (Park & Heo, 2020), introduces a *Channel and Spatial Correlation* (CSC) and an *Adaptive Cross Entropy* (ACE) term.

3. Proposed method

The MCKD method consists of performing multiple stages of vanilla KD.

Architectures. We selected **OneFormer** (Jain et al., 2023) (219M parameters) as the teacher model because it stands out as one of the best state-of-the-art (SOTA) models for semantic segmentation. In our setup, the model's backbone is *Swin Transformer Large* (Liu et al., 2021).

We designed a custom encoder-decoder model based on **DeepLabV3+** (Chen et al., 2018) as the student. The model's backbone is *Swin Transformer Tiny*. Note that the student model comprises a total of 34M parameters, but only 6,7M are trainable due to limited computational resources.

Knowledge distillation. We used an *offline* and *response-based* distillation approach, which means that the teacher network is pre-trained on the specific dataset and that the student model learns to mimic the predictions of the teacher model by minimizing the difference between the logits. Note that since both backbones are pre-trained and the architectures of the decoders are very different, it was not possible to adopt a feature-based approach.

Losses. The total loss is a weighted sum of 3 different losses: *Cross-Entropy*, *Dice*, and *Kullback–Leibler divergence*. We compute the Cross-Entropy loss to ensure the alignment of the student's prediction and the teacher's pseudo-label distributions. Additionally, we incorporated the Dice loss to address spatial overlap accuracy and the

[1]1713638 [2]1844613 [3]2002288 [4]1796812

Advanced Machine Learning, Sapienza University of Rome, a.y. 2023/2024.

Kullback-Leibler (KL) divergence loss for measuring the dissimilarity between probability distributions. Note that we conducted some experiments introducing the *Focal* loss to address the class imbalance and mitigate the impact of well-classified pixels on overall performance, but with no performance improvements.

4. Datasets and Benchmark

We evaluate our approach on two public semantic segmentation benchmark datasets. **COCO** (Lin et al., 2014) has 133 (53 "stuff" and 80 "thing") classes. We used the *Test2017* version, which doesn't provide annotations. The dataset contains 41k images, but we were forced to work with only 10% of them due to limited computational resources. Train, validation, and test sets contain 70%, 20%, and 10% of the images, respectively. **Cityscapes** (Cordts et al., 2016) consists of a total of 19 (11 "stuff" and 8 "thing") classes with 2,975 training, 500 validation, and 1,525 test images.

Mean Intersection over Union (mIoU) and **F1-score** are the metrics used in all our experiments. OneFormer reaches 67,4% mIoU on Coco and 84,4% mIoU on Cityscapes.

5. Experimental results

We performed 6 different trainings, three for each dataset, to obtain the two students for the first stage and the final model for the last stage. Our final models, in 20 epochs with early stopping, were able to reach 30,8% mIoU on COCO and 49,3% on Cityscapes. The students' final performances increased with the second dataset, but they are still low compared to those of our teacher.

The hyper-parameters tuning was a crucial step to reach acceptable performance. We conducted experiments without neglecting any hyper-parameter, with a major focus on adjusting the losses' weights and the temperature of the KD loss. During this process, we realized two interesting phenomena. First of all, in contrast to the original paper introducing KD, where the authors suggested assigning more importance to the KD loss (KL divergence) compared to other losses, we found that such an approach did not yield favorable results for our specific case. Instead, we set the weight of the KD loss very low and we treated it as a regularization term to achieve better performance in our experiments. Second, it is commonly recommended by experts to set the temperature of KD loss to a low value (e.g., 2) when the student model is significantly smaller than the teacher. However, in our case, contrary to this convention, we found that the opposite approach yielded better results.

Looking at the qualitative results. We can see that our final models face challenges in accurately individuating small objects. We hypothesize that this difficulty may stem from the resizing of the teacher's input before it reaches the student model. The resizing process could potentially lead to a loss of fine-grained details crucial for discerning smaller entities. The necessity to resize input images originated from the intrinsic architectural limitations of our model, which is not designed to process large inputs like the teacher model. Larger images (e.g. 800x800), would make our model struggle to accurately identify the internal part of very large objects, probably due to a loss of contextual information.
While our model demonstrated overall proficiency in object detection, it struggles to accurately outline object contours, which are rarely precise and frequent occurrences of flickering inconsistencies.

Suspicious of the low performance on the COCO dataset, we observed that some objects are always identified (e.g. people), while others, regardless of their size, are rarely identified. To further investigate the issue, we examined the class distribution in the training dataset and discovered a high degree of imbalance among classes. An initial idea was not to randomly select the images to compose our training dataset and to make all classes equally represented. The challenge, however, lies in the lack of annotations to filter the images before creating the dataloader. Moreover, we could leverage OneFormer to generate labels, but once the dataloader is created, we no longer have access to image IDs to perform selective filtering.
Moreover, while fewer than 1500 images may be sufficient to train the model effectively on the Cityscapes dataset with 19 classes, this quantity proves insufficient for training the model on the COCO dataset with its extensive 133 classes. In the case of COCO, we face the challenge of having too few images to adequately cover the diverse range of classes, worsening the difficulty of training a robust and accurate model across all categories.

6. Conclusions and Future work

We saw how changing the dataset was a crucial choice to obtain better results on the MCKD. Class imbalance and reduced dataset prompted us to experiment with Cityscapes since it has 19 classes instead of 133 (COCO).
The obtained results are still not even close to the performance of OneFormer, but we consider them a good starting point for further investigation. Even if KD is an agnostic method, which means that is not architecture-dependent, we strongly think that using a smaller version of the OneFormer decoder can lead to significant improvements. In this way, it could be even possible to perform different kinds of KD, such as features-based or relation-based.

References

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., and Kolesnikov, A. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10925–10934, June 2022.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Chen, X., Yuan, Y., Zeng, G., and Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

He, T., Shen, C., Tian, Z., Gong, D., Sun, C., and Yan, Y. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 578–587, 2019.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jain, J., Li, J., Chiu, M. T., Hassani, A., Orlov, N., and Shi, H. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2989–2998, June 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., and Wang, J. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2604–2613, 2019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Park, S. and Heo, Y. S. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors*, 20(16):4616, 2020.

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., and Wilson, A. G. Does knowledge distillation really work? *Advances in Neural Information Processing Systems*, 34: 6906–6919, 2021.

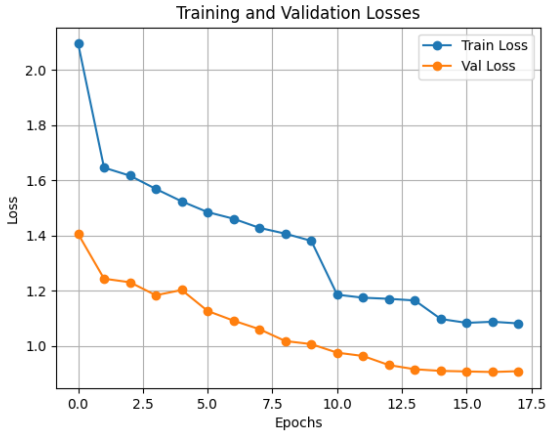
Xie, J., Shuai, B., Hu, J.-F., Lin, J., and Zheng, W.-S. Improving fast segmentation with teacher-student learning. *arXiv preprint arXiv:1810.08476*, 2018.

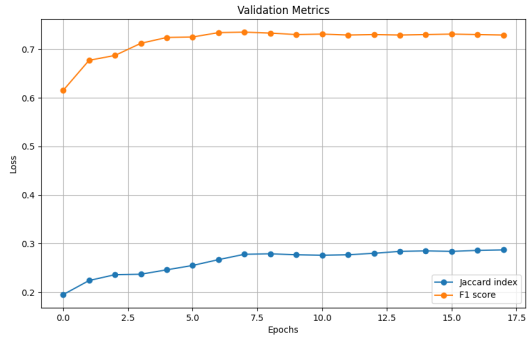
Zhao, J., Zhao, W., Drozdov, A., Rozonoyer, B., Sultan, M. A., Lee, J.-Y., Iyyer, M., and McCallum, A. Multistage collaborative knowledge distillation from large language models. *arXiv preprint arXiv:2311.08640*, 2023.

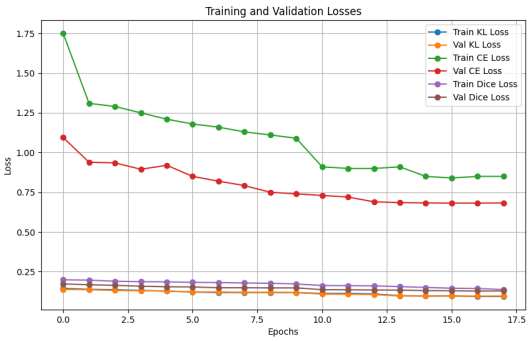
A. Quantitative results

A.1. COCO

In this section, we report the quantitative results computed on the COCO dataset. The following three figures represent the training and validation losses (1), the Jaccard (mIoU) and dice (f1-score) metrics (2), and all the single losses (3) respectively.

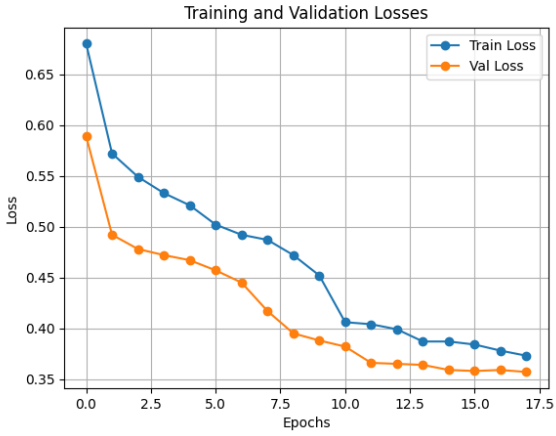


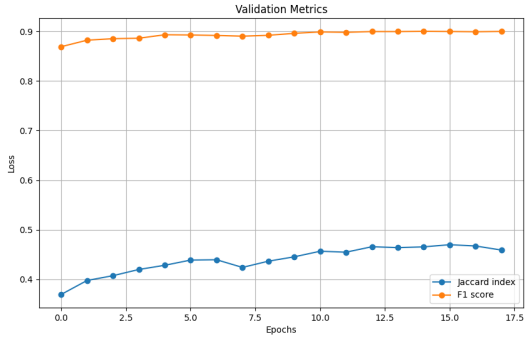


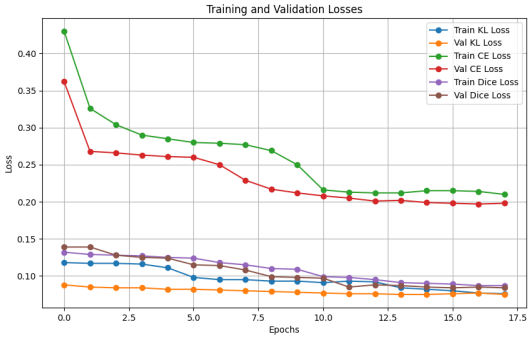


A.2. CityScapes

In this section, we report the quantitative results computed on the Cityscapes dataset. The following three figures represent the training and validation losses (1), the Jaccard (mIoU) and dice (f1-score) metrics (2), and all the single losses (3) respectively.



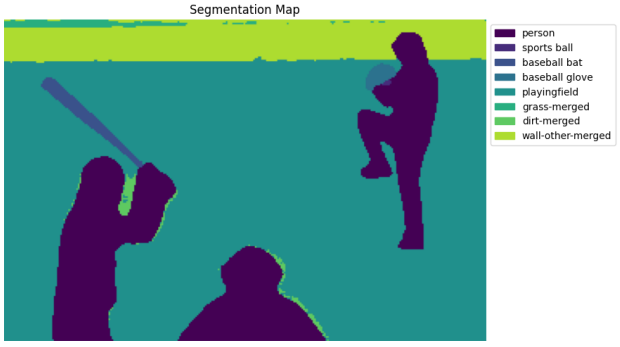


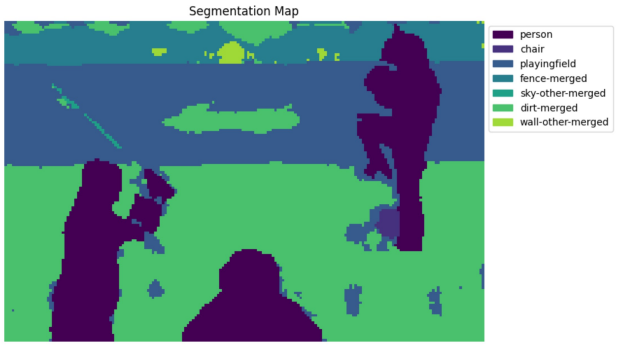


B. Qualitative results

B.1. COCO

In this section, we report the quantitative results computed on the COCO dataset. The following three figures represent the original image (1), the pseudo-label generated by the teacher (2), and the final model prediction (3) of a sample of the COCO dataset.





B.2. CityScapes

In this section, we report the quantitative results computed on the Cityscapes dataset. The following three figures represent the original image (1), the pseudo-label generated by the teacher (2), and the final model prediction (3) of a sample of the Cityscapes dataset.