Dynamic Mutual Training - A Self-Supervised Semantic Segmentation Framework for Robust Pseudo Noise Handling

Eirik Aalstad Baekkelund

University College London

Abstract. In semi-supervised learning, generating reliable pseudo labels is crucial for effective performance. However, the reliance on a single model's prediction confidence to filter low-confidence pseudo labels results in high-confidence errors and wasted low-confidence correct labels. In this paper, we experiment with the existing Dynamic Mutual Training (DMT) framework. Specifically, we look at the labeled proportion of data, and the relative proportionality of labeled subsets for pre-training given to the two models in DMT using Different Maximized Sampling (DMS). It aims to demonstrate how the use of DMS in the DMT framework affect the reliability of pseudo-labels generated in semi-supervised learning and how model pre-training impact the performance of the DMT-DMS approach.

1 Introduction

The emergence of deep learning has accelerated research for many computer vision tasks, such as image classification and semantic segmentation [7, 8]. However, the models have a need for large amounts of data. Despite data collection for imaging being easy, annotating the collected data is not. For instance, the labeling of high-resolution images pixel-wise can take a human annotator 1.5 hours [3]. Therefore, in fields where high-resolution pixel-wise annotations are critical, pseudo labeling is immensely useful (e.g. medical imaging [10]). In semi-supervised learning, current methods use the idea of generating pseudo labels for supervision where the pseudo labels come from self-training methods. They typically face a trade-off. When predictions are high-confidence, the pseudo labels generated are less noisy. This comes at the cost of discarding low-confidence correctly classified pseudo labels [5]. In [12], they demonstrate that proportions of pseudo labels exceeding 27%-36% can cause performance degradation. Existing methods focus on random noise that can be modeled by transition matrices [2, 1]. They tend to only perform well in the case where only random noise is present, such as Dual Student [6] and Co-Teaching+ [11]. However, this may not be effective when dealing with pseudo label noise, which stem directly from the model producing labels. As a countermeasure to discarding potentially valuable pseudo labels, we seek to utilize (dis)agreement between two models to locate and counter the pseudo label errors. In contrast to the Co-Teaching+ framework, which leverages low-loss examples and use only one model for determining training targets, our

method seek to explicitly re-weighs the loss function based on the disagreement between two models. The paper is inspired by the novel Dynamic Mutual Training (DMT) framework in [4]. We demonstrate the use of the DMT framework with Different Maximized Sampling (DMS) to effectively handle pseudo noise related to semantic segmentation. We seek to experiment how different configurations of the DMT-DMS approach impact its performance compared to existing methods that rely on a single model's prediction confidence or leverage low-loss examples for determining training targets.

2 Methodology

2.1 Difference Maximized Sampling

The objective of Difference Maximised Sampling (DMS) is to select two equal-sized sub-subsets, $\mathcal{D}^A_{labeled}$ and $\mathcal{D}^B_{labeled}$, from a randomly shuffled subset $\mathcal{R} \subset \mathcal{D}_{labeled}$ of size L with the intersection between them being as small as possible. The smaller the intersection between the sub-subsets, the greater the difference between them. To achieve this, we sample $\alpha \in (0.5, 1)$ and select the subsets by

$$\mathcal{D}^A_{labeled} = \mathcal{R}_{0:\alpha L},\ \mathcal{D}^B_{labeled} = \mathcal{R}_{(1-\alpha)L:L} \tag{1}$$

2.2 Dynamic Loss

The dynamic loss is key to handling and quantifying the pseudo label noise. At each iteration, we have a batch of size N where $\mathcal{X}$ and $\mathcal{U}$ denotes the labeled and unlabeled batches. For a batch let f_{θ_A} be the teacher and f_{θ_B} is the student. Then for an unlabeled image $u \in \mathcal{U}$, we assign the pixel-wise pseudo labels (y_A) and the corresponding confidences (c_A) from the teacher as followed:

$$y_A = \arg\max_y f_{\theta_A}(y|u),\ c_A = f_{\theta_A}(y|u);\ y_A, c_A \in \mathbb{R}^{(W \cdot H) \times C}$$

where W is the width, H is the height, and C is the number of classes.

The student assign predictions (y_B) and confidences (c_B) in the same way as above (using f_{θ_B} instead). It also includes a model (dis)agreement quantification through predicted probability, p_B.

$$p_B = f_{\theta_B}(y_A|u)$$

The probability p_B can be interpreted as the level of confidence that f_{θ_B} has in the pseudo labels generated by f_{θ_A} for the unlabeled image u. It reflects the likelihood that the student model assigns to the teacher's pseudo labels.

Next, we define the dynamic loss weight (w_u) for the image:

$$w_u = \begin{cases} p_B^{\gamma_1}, & \text{if } y_A = y_B \\ p_B^{\gamma_2}, & \text{if } y_A \neq y_B,\ c_A \geq c_B\ , \\ 0, & \text{if } y_A \neq y_B,\ c_A < c_B \end{cases} \quad w_u \in \mathbb{R}^{H \times W} \tag{2}$$

The dynamic loss for the unlabeled samples becomes

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{N} \sum_{u, y_A \in \mathcal{U}} w_u CE(y_a, f_{\theta_B}(u) \tag{3}$$

where CE is the cross-entropy loss.

$$CE = -\sum_{c=1}^{C} y_c log(p_c) \tag{4}$$

The training with pseudo labeled data have three cases: agreement, negative disagreement, and positive disagreement. We use the current model's predicted probability on the pseudo labeled class as weight in cases 1 and 2, and set the weight to 0 in case 3. The weights are scaled by hyperparameters γ_1,γ_2. A larger γ_1 value emphasize entropy minimization and larger γ_2 emphasize mutual learning. The loss for labeled data, $\mathcal{L}_{\mathcal{X}}$, is the classical CE in the supervised setting. Our final loss becomes

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \mathcal{L}_{\mathcal{U}} \tag{5}$$

2.3 Dynamic Mutual Training (DMT)

The Dynamic Mutual Traning (DMT) framework aims to reduce pseudo label noise during training by enabling inter-model disagreement. First, we start with a set of labeled- and unlabeled data, $\mathcal{D}_{labeled}$ and $\mathcal{D}_{unlabeled}$. Then, the models, f_{θ_A} and f_{θ_B}, are initialized with separate weights. Before DMT starts, they are trained on different labeled subsets, $\mathcal{D}^A_{labeled}$ and $\mathcal{D}^B_{labeled}$, obtained by the Difference Maximized Sampling (DMS) algorithm (See 2.1). Finally, the DMT is applied where one model assumes the role as teacher and another assumes the role as student. The teacher will provide pseudo labels specified by a percentile at each epoch. These act as true labels for tuning the student. However, the comparative confidences (c_A and c_B) and predictions (y_A and y_B) determine how the models are updated. Namely, by applying a dynamically weighted cross-entropy loss function. Then, the roles are reversed such that f_{θ_B} becomes the teacher and f_{θ_A} is assigned as the student. In the traditional DMT setting, this ordering is kept fixed. The DMT paper discuss how some classes are easier to learn in semantic segmentation than others, and proposes an iterative self-training scheme to address this [4]. They are inspired by CBST [13] that use fine-tuning for faster convergence. DMT conducts two separate fine-tunings between two differently initialized models, where two models train each other equally and the inter-model disagreement is utilized. We keep the iterative framework of DMT in selecting top-confident pseudo labels by direct ranking.

Algorithm 1 Pseudo code for Dynamic Mutual Training for semantic segmentation with alternating teacher-student roles.

Require: $\mathcal{D}_{labeled}$, $\mathcal{D}_{unlabeled}$, γ_1, γ_2
Initialize f_{θ_A} and f_{θ_B} with same pre-trained weights
$\mathcal{R} \leftarrow$ Randomly shuffled $\mathcal{D}_{labeled}$
$\mathcal{D}^A_{labeled}$, $\mathcal{D}^B_{labeled}$ = DMS($\mathcal{R}$)
Train f_{θ_A} on $\mathcal{D}^A_{labeled}$ and f_{θ_B} on $\mathcal{D}^B_{labeled}$
$\alpha = \{0.2, 0.4, 0.6, 0.8, 1.0\}$
for i in $\{1, 2, 3, 4, 5\}$ **do**
    $\mathcal{D}^A_{pseudo} \leftarrow$ the top α_i pixels w.r.t. prediction confidence from $f^{i-1}_{\theta_A}$
    Train $f^{i-1}_{\theta_B}$ on $\mathcal{D}_{labeled} \cup \mathcal{D}^A_{pseudo}$ on dynamic loss
    $\mathcal{D}^B_{pseudo} \leftarrow$ the top α_i pixels w.r.t. prediction confidence from $f^{i}_{\theta_B}$
    Train f_{θ_A} on $\mathcal{D}_{labeled} \cup \mathcal{D}^B_{pseudo}$ on dynamic loss
end for
return Model with highest IoU from $f^5_{\theta_A}$ and $f^5_{\theta_B}$

3 Data

The Oxford Pets dataset is a popular benchmark for semantic segmentation tasks, consisting of 7,400 images of pets with pixel-wise annotations for 37 classes of cats and dogs. To ensure balanced and effective learning in semantic segmentation, the data loading process is crucial. This includes the separation of data into disjoint splits for training and testing, to prevent information leakage from the test set and avoid overly optimistic evaluations. Class balance promotes effective and diverse training of the models. Finally, consistency of data across experiments is important for fair comparisons of the impact of the experiments on model performance, without potential confounding effects from changes in the data itself. All these measures are accounted for within the data loading of the project to enforce efficient training and direct comparison between experiments. Additionally, different sizes of labeled and unlabeled data can occur. Then, the batch sizes of the labeled and unlabeled datasets are adjusted to have similar numbers of mini-batches, ensuring the whole unlabeled dataset is used at least once if it has more data. The data is also split in a train- and test set with 5913 and 1480 images, respectively. The training set is further split to a 5% validation set used for determining who is assigned the teacher first role in DMT, which is the model with the highest IoU after pre-training.

4 Experiments

We conducted the following set of experiments. First, training the models on varying proportions of labeled data; $\{1\%, 2\%, 5\%, 10\%, 50\%, 80\%, 95\%\}$. Second, varying the number of DMT epochs before teacher-student reversal: $N = \{5, 10, 20, 30\}$.

4.1 Training Settings

Apart from the experiment specification parameters, they follow the default settings of the original DMT paper. We used SGD with a momentum of 0.9, the poly rate learning scheduler with a base learning rate of 0.004, and a batch size of 32. We followed the suggestion with $\gamma_1 = \gamma_2 = 3$. Additionally, we used other default settings: 10,000 epochs for (pre-)training, 10 epochs in DMT fine-tuning, α (percentiles) as in the pseudo-code, $\alpha = 0.7$ for DMS with $L = |\mathcal{D}_{labeled}|$, and a labeled proportion and a validation proportion of 10% and 5%, respectively. We used U-Net[1] models with random initialization for the baselines and for the models in the DMT as it has shown state-of-the-art performance on semantic segmentation tasks [9]. It also provides a direct comparison for the effectiveness of DMT.

4.2 Test Results

When comparing results, we compare against a lower bound Baseline model trained on default label proportion (10%) and an upper bound model, Oracle, which sees all the labeled data. We also compare against a Pseudo-Label self-training model that first converge on labeled data, then generates pseudo labels on the unlabelled training set and incrementally increases the weight of the pseudo labels during training to minimize entropy over unseen predictions. We select each model by the highest Intersection over Union (IoU) on the validation set during training. The models are then evaluated on the test set.

The default setting experiment show that DMT outperforms the baseline U-Net model and the Pseudo-Label model. increasing the labeled data proportion improves the performance 1. Still, improvements are only marginal compared to the baseline. We hypothesize that the identical models and insufficient dataset diversity may limit the effectiveness of the DMT framework with inter-model disagreement. Also, we did not conduct consistency regularization, which can limit the consistency when unlabeled data is perturbed.

Model	**IoU**
Oracle	0.888 ± 0.001
DMT	0.823 ± 0.002
Pseudo-Label	0.814 ± 0.002
Baseline	0.798 ± 0.002

Table 1: Test IoU for all models with label fraction of 10%. The standard error on DMT is calculated over 3 runs, whereas the other models have 5 runs.

[1] GitHub link to U-Net

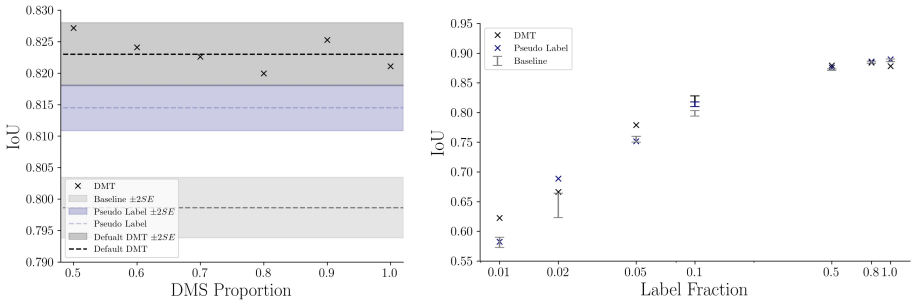

Fig. 1: Experiments plotted against IoU. (Left) Varying DMT Epoch teacher-student reversal. (Right) Varying label proportion.

The test results show that DMT outperforms the baseline U-Net model, and increasing the labeled data proportion improves the performance (Figure 1). Pseudo label training after seeing labeled data is more effective for low fractions of labeled data, as seen from the Pseudo- and DMT models. For the DMS, an increasing proportion of data to model A (teacher first) seem to degrade performance compared to even splits. This could lead to inter-model disagreement not being as effective. DMS aims to improve the performance of a model by encouraging it to learn patterns for the same class of data, which can prevent overfitting and improve generalization. So, DMS make the subsets as different as possible with respect to instances but they may see an equal distribution of classes. Thus, to improve disagreement efficiency, deciding who is teacher first at each iteration based on its current performance (as measured by IoU on validation) could reduce the pseudo noise as the higher accuracy can reduce the pseudo noise from the teacher. Additionally, the framework can adapt to changing conditions if the best model is used to guide the learning process for the less accurate model. Thus, alternating who is teacher first may prove effective in combination with a balanced size of pre-train sets for the two models pose an interesting ablation study.

Additionally, DMT paper suggested pre-trained weights, which we did not have. Further, we used online self-training in DMT, making the training more susceptible to pseudo noise. This pose a limitation to our implementation in addition to the lack of consistency regularization. Future research should explore the use of more diverse datasets or different model architectures to better utilize the DMT framework. Additionally, investigating the weighting factor used in the loss function, which acts as the pseudo noise "regularizer" could improve performance. The DMT paper propose a sigmoid-like function for γ but it is not used in their implementation.

Overall, our study provides insight into the potential of DMT for improving the performance of semantic segmentation models, and highlights the need for exploration and optimization of the framework to achieve optimal results.

References

[1] Dana Angluin and Philip D. Laird. "Learning From Noisy Examples". In: *Machine Learning* 2 (1988), pp. 343–370.

[2] Alan Bekker and Jacob Goldberger. "Training deep neural-networks based on unreliable labels". In: Mar. 2016, pp. 2682–2686. DOI: 10.1109/ICASSP.2016.7472164.

[3] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: June 2016. DOI: 10.1109/CVPR.2016.350.

[4] Zhengyang Feng et al. "Semi-Supervised Semantic Segmentation via Dynamic Self-Training and Class-Balanced Curriculum". In: *CoRR* abs/2004.08514 (2020). arXiv: 2004.08514. URL: https://arxiv.org/abs/2004.08514.

[5] Wei-Chih Hung et al. "Adversarial Learning for Semi-Supervised Semantic Segmentation". In: *CoRR* abs/1802.07934 (2018). arXiv: 1802.07934. URL: http://arxiv.org/abs/1802.07934.

[6] Zhanghan Ke et al. *Dual Student: Breaking the Limits of the Teacher in Semi-supervised Learning*. 2019. arXiv: 1909.01804 [cs.LG].

[7] Xiaolong Liu, Zhidong Deng, and Yuhan Yang. "Recent progress in semantic image segmentation". In: *Artificial Intelligence Review* 52.2 (2019), pp. 1089–1106. DOI: 10.1007/s10462-018-9641-3. URL: https://doi.org/10.1007/s10462-018-9641-3.

[8] Keiron O'Shea and Ryan Nash. "An Introduction to Convolutional Neural Networks". In: *CoRR* abs/1511.08458 (2015). arXiv: 1511.08458. URL: http://arxiv.org/abs/1511.08458.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[10] Constantin Marc Seibold et al. "Reference-Guided Pseudo-Label Generation for Medical Semantic Segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (June 2022), pp. 2171–2179. DOI: 10.1609/aaai.v36i2.20114. URL: https://doi.org/10.1609%5C%2Faaai.v36i2.20114.

[11] Xingrui Yu et al. *How does Disagreement Help Generalization against Label Corruption?* 2019. arXiv: 1901.04215 [cs.LG].

[12] Yang Zou et al. "Confidence Regularized Self-Training". In: *CoRR* abs/1908.09822 (2019). arXiv: 1908.09822. URL: http://arxiv.org/abs/1908.09822.

[13] Yang Zou et al. *Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training*. 2018. arXiv: 1810.07911 [cs.CV].