📖

[논문 리뷰] Meta Pseudo Label [2020]

🌟 **본 논문은 Unlabel data에서, Pseudo label을 생성하여, Student Network를 학습하고,**
    Meta Pseudo Label의 Teacher Network는 label 대한 Student의 학습 결과를 Feedback합니다.

Introduction

이전의 Pseudo Label을 활용할 경우, 가장 **큰 문제점은 Teacher Network가 unlabeled data를 정확하지 않게 Pseudo Label하는 경우, Student Network가 inaccurate training을 하는 문제**가 있습니다.

이러한 **문제점을 해결하기 위해**, **Meta Pseudo Label은 Pseudo Label이 Student에 어떻게 영향을 미치는지 관찰함으로서, Teacher Network는 systematic mechansim하게 Design되었습니다.**

실험을 하는 과정에서, feedback signal은 label dataset에서 performance를 보였습니다.

Teacher & Student Network는 동시에 train을 진행합니다.

1. **Student Network는 teacher Network로부터, 생성된 Pseudo Label을 학습**합니다.
2. **Teacker Network는 Student Network의 학습의 reward signal로부터 학습**합니다.

Datasets	ImageNet Top-1 Accuracy	ImageNet-ReaL Precision@1
Previous SOTA [16, 14]	88.6	90.72
Ours	**90.2**	**91.02**

Table 1: Summary of our key results on ImageNet ILSVRC 2012 validation set [56] and the ImageNet-ReaL test set [6].

Meta Pseudo Labels

Meta Pseudo Label과 Pseudo Label의 차이점으로는, **Student Network의 Feedback에 관하여,**
학습유무에 따른 것이다.

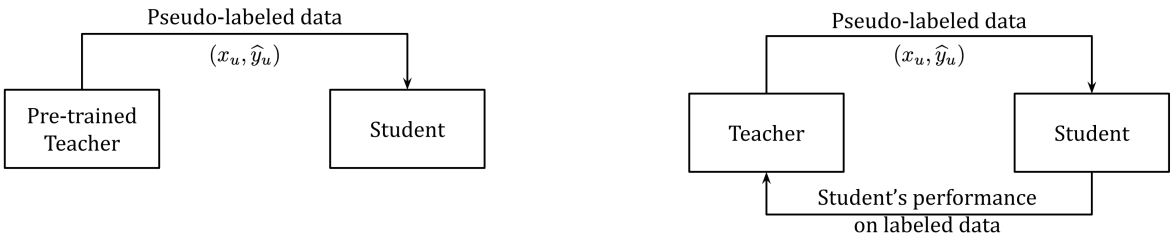


Figure 1: The difference between Pseudo Labels and Meta Pseudo Labels. **Left:** Pseudo Labels, where a fixed pre-trained teacher generates pseudo labels for the student to learn from. **Right:** Meta Pseudo Labels, where the teacher is trained along with the student. The student is trained based on the pseudo labels generated by the teacher (top arrow). The teacher is trained based on the performance of the student on labeled data (bottom arrow).

Pseudo Label Optimization

- Pseudo Label(PL)을

$$\theta_S^{\mathrm{PL}} = \underset{\theta_S}{\mathrm{argmin}} \underbrace{\mathbb{E}_{x_u}\Big[\mathrm{CE}\big(T(x_u;\theta_T), S(x_u;\theta_S)\big)\Big]}_{:=\mathcal{L}_u\left(\theta_T,\theta_S\right)}$$

와 같이 정의!

- T : 선생님 모델
- S : 학생 모델
- θ_T : 궁극적인 목적은 θ_S^{PL} 이 Labeled Data에서 낮은 Loss를 달성하는 것

즉,

$$\theta_S^{\mathrm{PL}} = \underset{\theta_S}{\mathrm{argmin}} \underbrace{\mathbb{E}_{x_u}\Big[\mathrm{CE}\big(T(x_u;\theta_T), S(x_u;\theta_S)\big)\Big]}_{:=\mathcal{L}_u\left(\theta_T,\theta_S\right)} \rightarrow \mathbb{E}_{x_l,y_l}\Big[\mathrm{CE}\big(y_l, S(x_l;\theta_S^{\mathrm{PL}})\big)\Big] := \mathcal{L}_l\big(\theta_S^{\mathrm{PL}}\big).$$

- Student 모델 (θ_S^{PL})의 Loss 기댓값을 낮게 만들어야 함
- θ_S^{PL} 은 항상 θ_T 의존적일 수밖에 없음 (다른 말로 θ_T 의 Function이다!)
- 최종 목적 함수 정의

$\mathcal{L}_l$ with respect to θ_T:

$$\begin{aligned} \min_{\theta_T} \quad & \mathcal{L}_l\left(\theta_S^{\mathrm{PL}}(\theta_T)\right), \\ \text{where} \quad & \theta_S^{\mathrm{PL}}(\theta_T) = \underset{\theta_S}{\mathrm{argmin}}\, \mathcal{L}_u\big(\theta_T, \theta_S\big). \end{aligned} \tag{2}$$

가 Meta Pseudo Label 방법의 최종적인 (이중) 목적함수

Meta Pseudo Labels의 최적화 어려움

- Teacher, Student Network간의 종속성 때문에, 계산이 어렵다.
- 예를들면, $argmin_{0_s}$ 을 모두 구해야 합니다.

- **Meta Pseudo Labels 최적화 해결방안**

  - Approximate Multi-step $argmin_{\theta_S}$ 을 사용

$$\theta_S^{\mathrm{PL}}(\theta_T) \approx \theta_S - \eta_S \cdot \nabla_{\theta_S} \mathcal{L}_u(\theta_T, \theta_S),$$

$\mathcal{L}_l$ with respect to θ_T:

$$\begin{aligned} \min_{\theta_T} \quad & \mathcal{L}_l\left(\boxed{\theta_S^{\mathrm{PL}}(\theta_T)}\right), \\ \text{where} \quad & \theta_S^{\mathrm{PL}}(\theta_T) = \operatorname*{argmin}_{\theta_S} \mathcal{L}_u(\theta_T, \theta_S). \end{aligned} \qquad (2)$$

$$\min_{\theta_T} \quad \mathcal{L}_l\Big(\theta_S - \eta_S \cdot \nabla_{\theta_S} \mathcal{L}_u(\theta_T, \theta_S)\Big). \qquad (3)$$

  - 방정식 (3)을 최적화 :

    1. θ_S 를 고정 매개변수로 취급
    2. θ_T 에 대한 고차 종속성을 무시합니다.

✨ **자세한 내용은 하단 링크를 참고하시기 바랍니다.**

Experiment

실험을 하였을 때, **Meta Pseudo Label이 Pseudo Label, Supervised Learning**에 비해,
좋은 성능을 보여주었습니다.

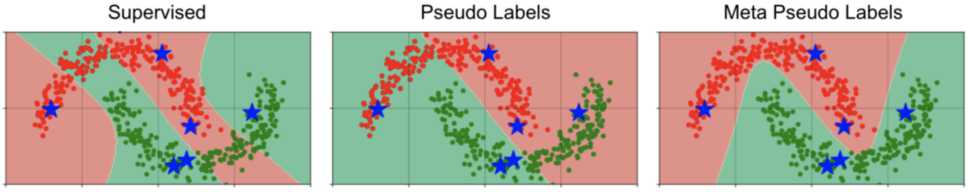


Figure 2: An illustration of the importance of feedback in Meta Pseudo Labels (right). In this example, Meta Pseudo Labels works better than Supervised Learning (left) and Pseudo Labels (middle) on the simple TwoMoon dataset. More details are in Section 3.1.

	Method	CIFAR-10-4K (mean ± std)	SVHN-1K (mean ± std)	ImageNet-10% Top-1	ImageNet-10% Top-5
Label Propagation Methods	Temporal Ensemble [35]	83.63 ± 0.63	92.81 ± 0.27	–	
	Mean Teacher [64]	84.13 ± 0.28	94.35 ± 0.47	–	
	VAT + EntMin [44]	86.87 ± 0.39	94.65 ± 0.19	–	83.39
	LGA + VAT [30]	87.94 ± 0.19	93.42 ± 0.36	–	
	ICT [71]	92.71 ± 0.02	96.11 ± 0.04	–	
	MixMatch [5]	93.76 ± 0.06	96.73 ± 0.31	–	
	ReMixMatch [4]	94.86 ± 0.04	97.17 ± 0.30	–	
	EnAET [72]	94.65	97.08	–	
	FixMatch [58]	95.74 ± 0.05	97.72 ± 0.38	71.5	89.1
	UDA* [76]	94.53 ± 0.18	97.11 ± 0.17	68.07	88.19
Self-Supervised Methods	SimCLR [8, 9]	–	–	71.7	90.4
	MOCOv2 [10]	–	–	71.1	–
	PCL [38]	–	–	–	85.6
	PIRL [43]	–	–	–	84.9
	BYOL [21]	–	–	68.8	89.0
	Meta Pseudo Labels	**96.11 ± 0.07**	**98.01 ± 0.07**	**73.89**	**91.38**
	Supervised Learning with full dataset*	94.92 ± 0.17	97.41 ± 0.16	76.89	93.27

Table 2: Image classification accuracy on CIFAR-10-4K, SVHN-1K, and ImageNet-10%. Higher is better. For CIFAR-10-4K and SVHN-1K, we report mean ± std over 10 runs, while for ImageNet-10%, we report Top-1/Top-5 accuracy of a single run. For fair comparison, we only include results that share the same model architecture: WideResNet-28-2 for CIFAR-10-4K and SVHN-1K, and ResNet-50 for ImageNet-10%. * indicates our implementation which uses the same experimental protocols. Except for UDA, results in the first two blocks are from representative important papers, and hence do not share the same controlled environment with ours.

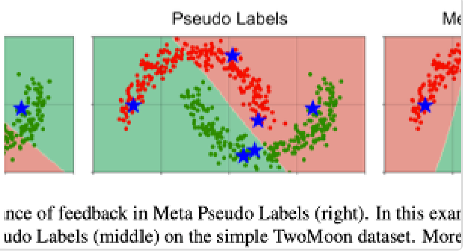
Conclusion

- Semi-Supervised Learning의 Meta Pseudo Label 방법을 제안합니다.
- **Meta Pseudo Labels의 Keypoints**:
    - Student Network의 training에 도움이 되는 방식으로, Pseudo Label을 생성하기 위해,
      Teacher Network는 Student Network의 피드백을 계속해서 반영함
    - Student Network의 성과를 기반으로, Teacher Network는 Update

Have A Nice AI

https://arxiv.org/pdf/2003.10580v4.pdf 설명 : Semi-supervised Learning으로도 엄청난 성능을 기록한 훌륭한 방법론 - Semi-Supervised Learning의 가능성 을 여김 없이 보여준 논문 - 엄청난 모

https://kmhana.tistory.com/33



Meta Pseudo Labels

We present Meta Pseudo Labels, a semi-supervised learning method that achieves a new state-of-the-art top-1 accuracy of 90.2% on ImageNet, which is 1.6% better than the existing

https://arxiv.org/abs/2003.10580