Meta Pseudo Labels

Hieu etc.(Google AI)
Cite: 3

2021.04.07.수
발표자: 임진혁

Index

1. Introduction: What is SSL(Semi-Supervised Learning)?

2. Related works

3. Proposed Method

4. Experiment

5. Conclusion

임진혁<eeplearning@units.ac.kr>

Index: 이런식으로 진행하고 싶다/해당 내용들 넣고 싶다

1. Pseudo label 설명
2. (1)이 왜 잘되는지?
3. Pseudo lable과의 차이점
4. 비교되는 기존 다른 연구들
5. 4에서 해당 문제의 challenges가 뭔지 파악
6. 해당 논문은 (5)을 어떻게 해결하였는가?

References

https://deep-learning-study.tistory.com/560?category=963091 [블로그 논문 리뷰]
[noisy student의 단점, 해당 연구의 작동 구조]

https://medium.com/@nainaakash012/meta-pseudo-labels-6480acb1b68 [영문 블로그 리뷰]
[mpl paper의 전체적인 리뷰 및 motivation 추정, sl ,kd 등과의 notation 비교]

https://yeomko.tistory.com/42 [블로그 리뷰]
[noisy student 기본 리뷰]

https://hoya012.github.io/blog/Self-training-with-Noisy-Student-improves-ImageNet-classification-Review/ [호야 블로그]
[noisy student 간단 리뷰]

https://jiwunghyun.medium.com/semi-supervised-learning-정리-a7ed58a8f023 [블로그]
[semi-supervised learning 설명]

https://blog.est.ai/2020/11/ssl/
[전반적인 semi-supervised learning 설명]

임진혁<eeplearning@units.ac.kr>

Why I choose this paper

- SOTA model (image classification)
- I was originally interested in “Semi – Supervised Learning”: Pseudo Label
- Looks similar to KD(Teacher & Student)
- Also I heard it is MAML form.

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].

Image Classification

Edit Task

Computer Vision

1430 papers with code **50** benchmarks **102** datasets

About

Edit

Image Classification is a fundamental task that attempts to comprehend an entire image as a whole. The goal is to classify the image by assigning it to a specific label. Typically, Image Classification refers to images in which only one object appears and is analyzed. In contrast, object detection involves both classification and localization tasks, and is used to analyze more realistic cases in which multiple objects may exist in an image.

Source: Metamorphic Testing for Object Detection Systems

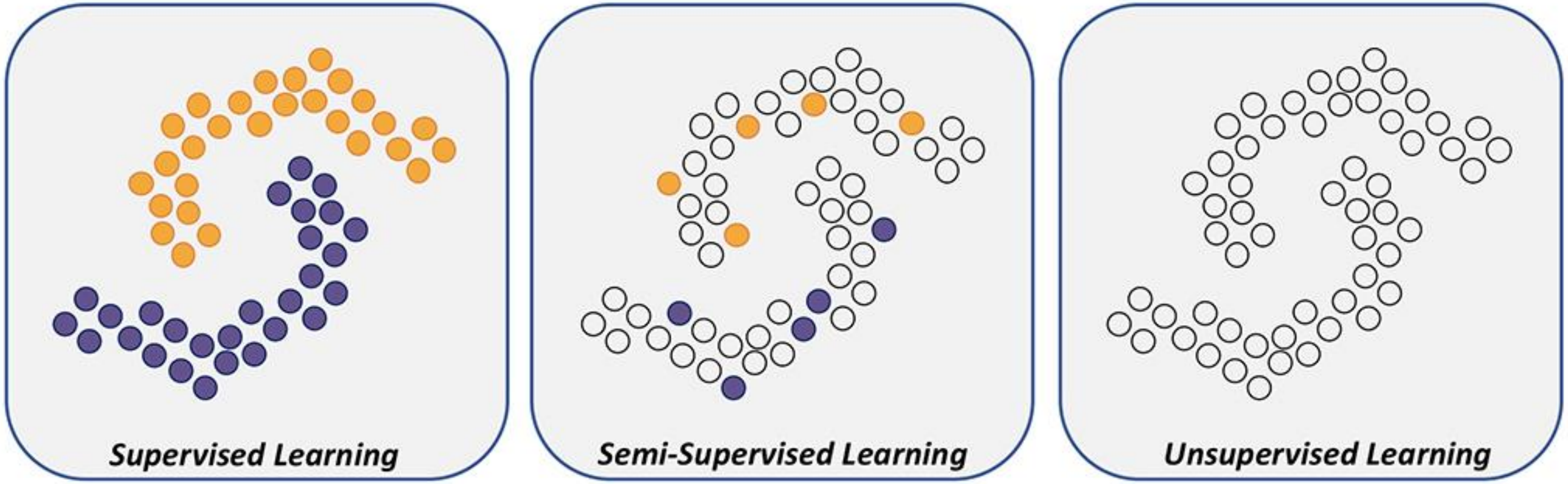
Benchmarks

Add a Result

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	ImageNet	Meta Pseudo Labels (EfficientNet-L2)	Meta Pseudo Labels			See all
	CIFAR-10	EffNet-L2 (SAM)	Sharpness-Aware Minimization for Efficiently Improving Generalization			See all
	CIFAR-100	EffNet-L2 (SAM)	Sharpness-Aware Minimization for Efficiently Improving Generalization			See all
	STL-10	Wide-ResNet-101 (Spinal FC)	SpinalNet: Deep Neural Network with Gradual Input			See all

Introduction: what is SSL(Semi-Supervised Learning)?

- **Supervised Learning (label O)**
- **Unsupervised Learning (label X)**
- **Semi - Supervised Learning**

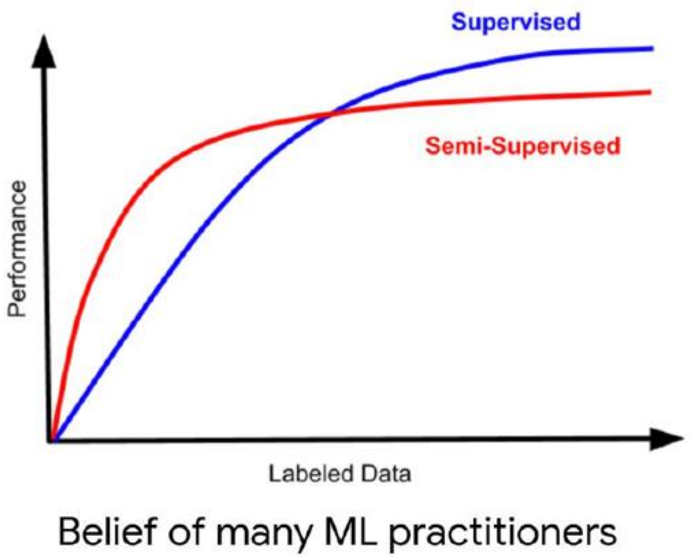


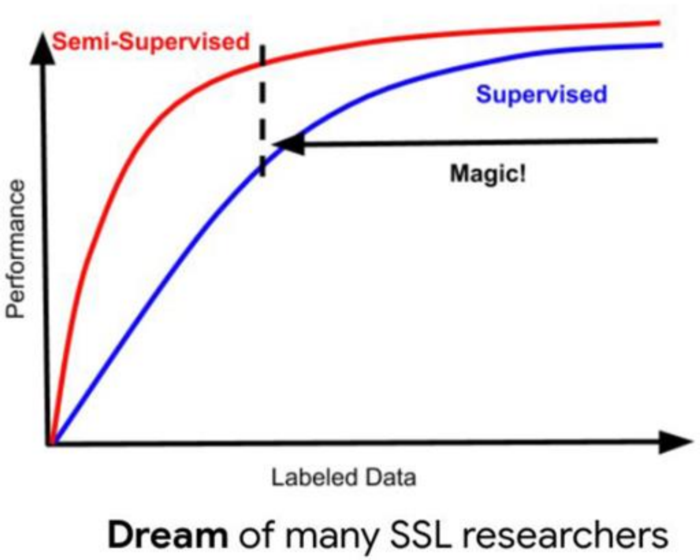
https://blog.est.ai/wp-content/uploads/2020/11/fig_1.jpg

Introduction: what is SSL?

Semi - Supervised Learning

Most motivation of SSL is Cost of Labelling.
=> We want improve the performance of SL through unlabeled data





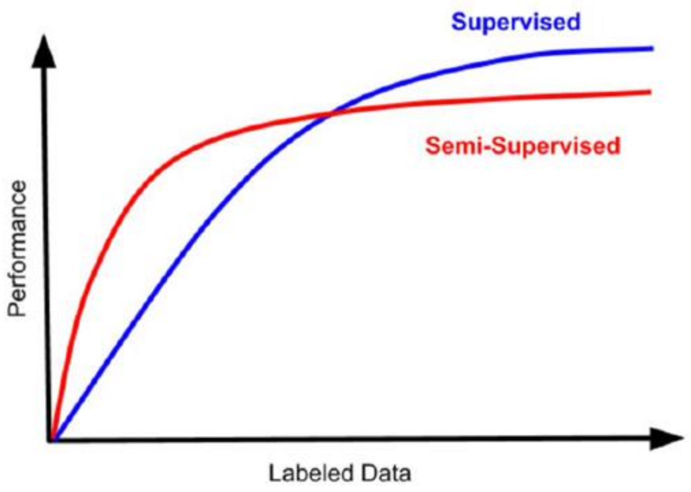
https://miro.medium.com/max/2400/1*uablqfc2X8y5vSoEOcLzAw.png

임진혁<eeplearning@units.ac.kr>

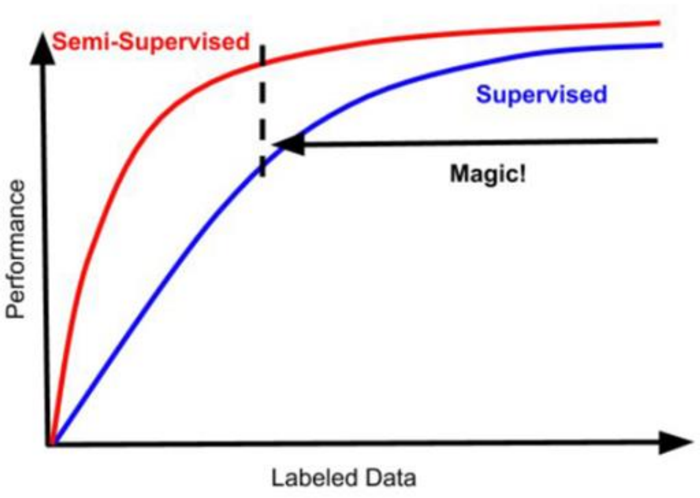
Introduction: SSL overview

Semi - Supervised Learning

Most motivation of SSL is Cost of Labelling.
=> We want improve the performance of SL through unlabeled data



Belief of many ML practitioners



Dream of many SSL researchers

https://miro.medium.com/max/2400/1*uablqfc2X8y5vSoEOcLzAw.png

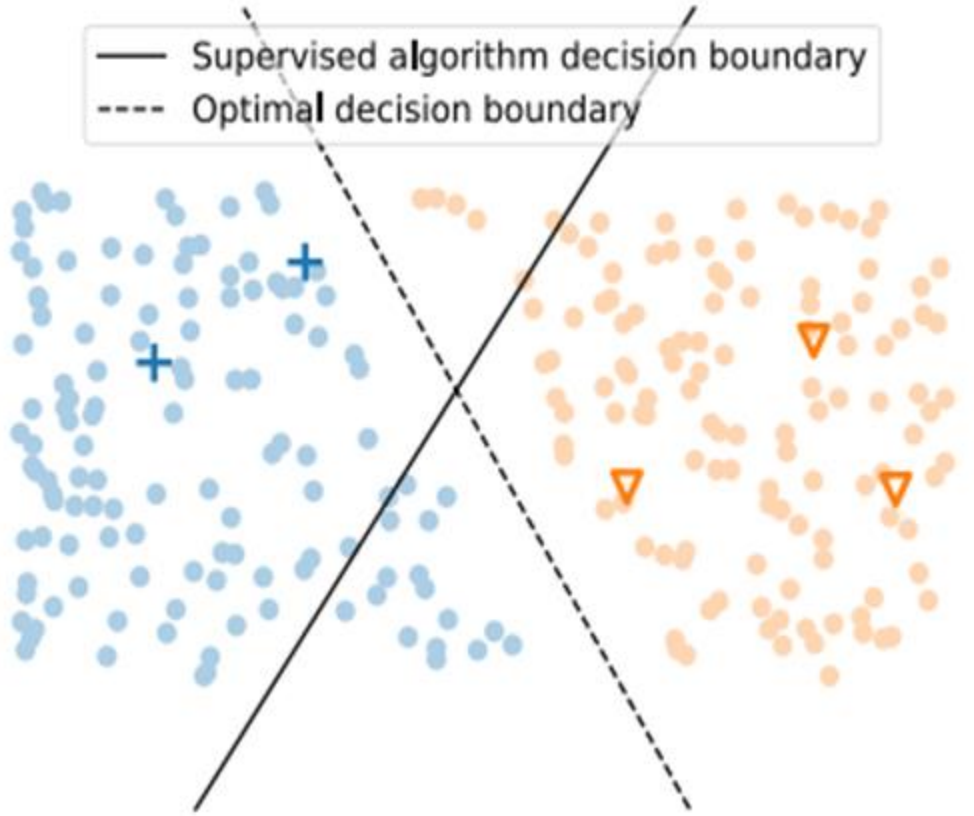
임진혁<eeplearning@units.ac.kr>

Introduction: SSL overview

SSL의 기본 가정

클러스터 가정: <u>데이터들이 같은 클러스터에 속하면 해당 데이터들은 같은 클래스에 속한다</u>

1. **smoothness 가정**
2. **low-density 가정**
3. **manifold 가정**



임진혁<eeplearning@units.ac.kr>

1 Introduction: SSL overview

SSL의 기본 가정에서 파생되는 다양한 SSL 기법들 (미완)

Wrapper methods => Self training => pseudo label

임진혁<eeplearning@units.ac.kr>

Introduction: SSL overview

Ssl의 기본 가정에서 파생되는 다양한 ssl 기법들: self trainng 방식 / kd와의 차이점

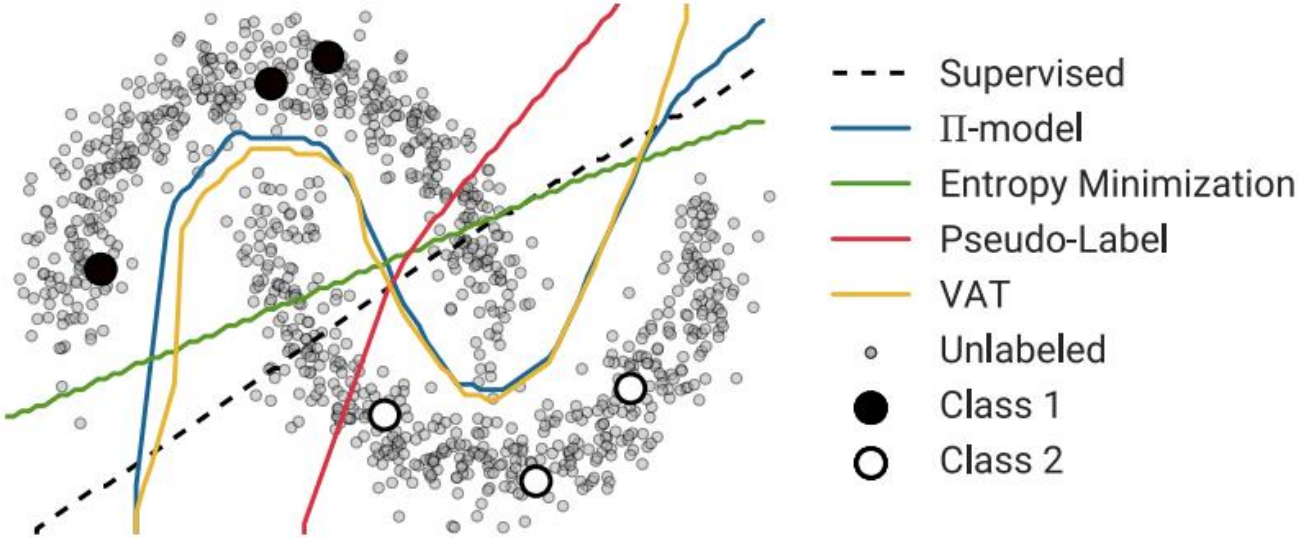
임진혁<eeplearning@units.ac.kr>

Introduction: SSL overview

self trainng

- Pseudo label
- Noisy

설명하고 이것들의 문제점을 설명하고 최종적으로 meta pseudo label 동작 구조 설명



임진혁<eeplearning@units.ac.kr>

Pseudo Label?

<Self – training>

Require: Labeled images $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ and unlabeled images $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m\}$.

1: Learn teacher model θ_* which minimizes the cross entropy loss on labeled images

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f^{noised}(x_i, \theta))$$

2: Use an unnoised teacher model to generate soft or hard pseudo labels for unlabeled images

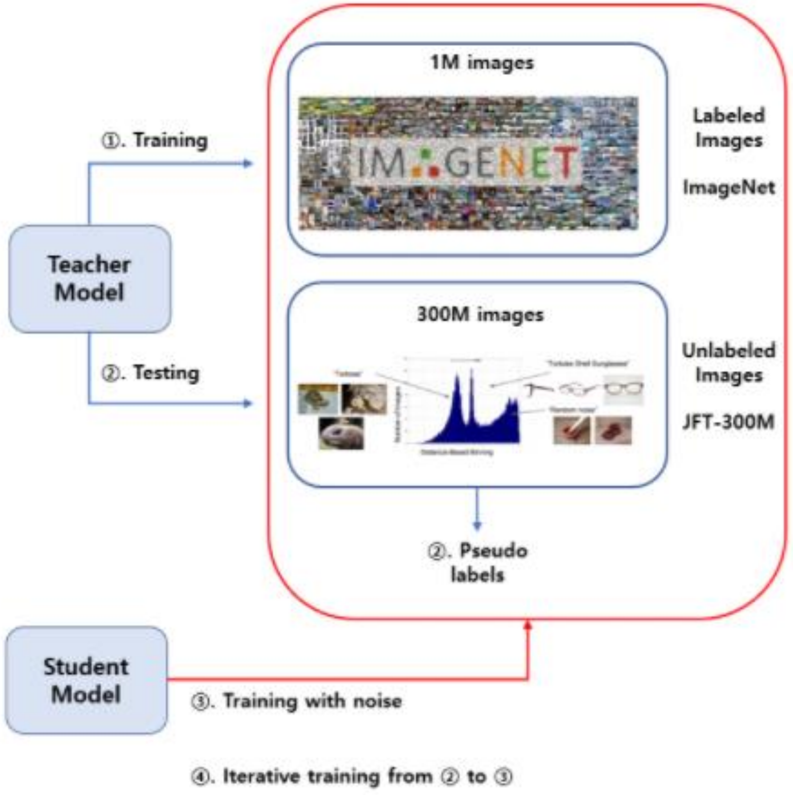
$$\tilde{y}_i = f(\tilde{x}_i, \theta_*), \forall i = 1, \cdots, m$$

3: Learn student model θ'_* which minimizes the cross entropy loss on labeled images and unlabeled images with noise added to the student model

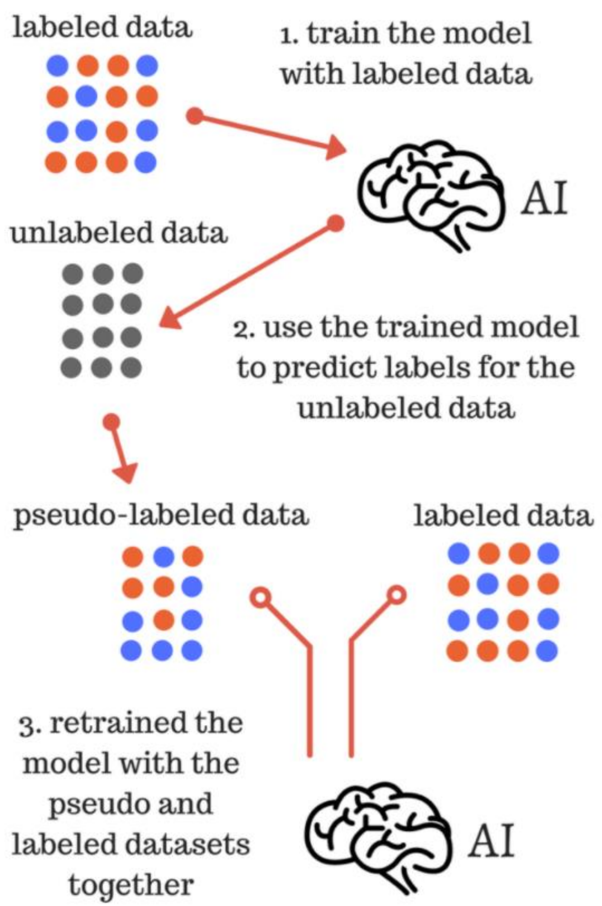
$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f^{noised}(x_i, \theta')) + \frac{1}{m}\sum_{i=1}^{m} \ell(\tilde{y}_i, f^{noised}(\tilde{x}_i, \theta'))$$

4: Iterative training: Use the student as a teacher and go back to step 2.

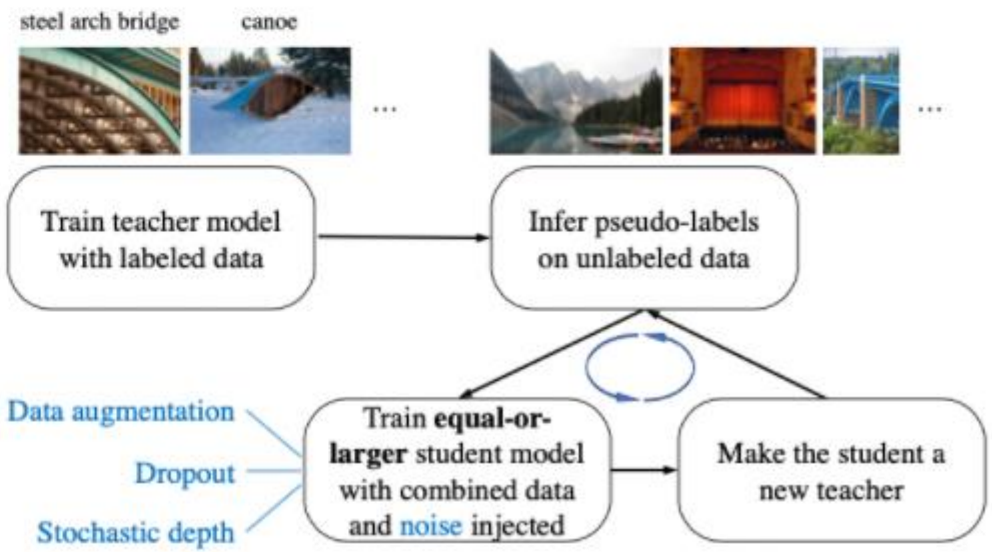
Algorithm 1: Noisy Student method



Pseudo Label



Noisy Student Network



Related: Prior Domain Generalization

Learning domain-invariant feature representation(Generalizable features)

⇒ Prior Domain Generalization :

expose Model with variety of source domains as possible as many

➔ Reduce the burden for designing ALGORITHMS for DG

⇒ Collecting data of large variety domains : high cost & impssible

Propose *MixStyle* : mix style across source domians

Why?

임진혁<eeplearning@unist.ac.kr>

Proposed Method: Meta – Pseudo Label

What's different from [Pseudo Label]

➔ **Teacher model is not fixed but adopted by the feedback of Student**

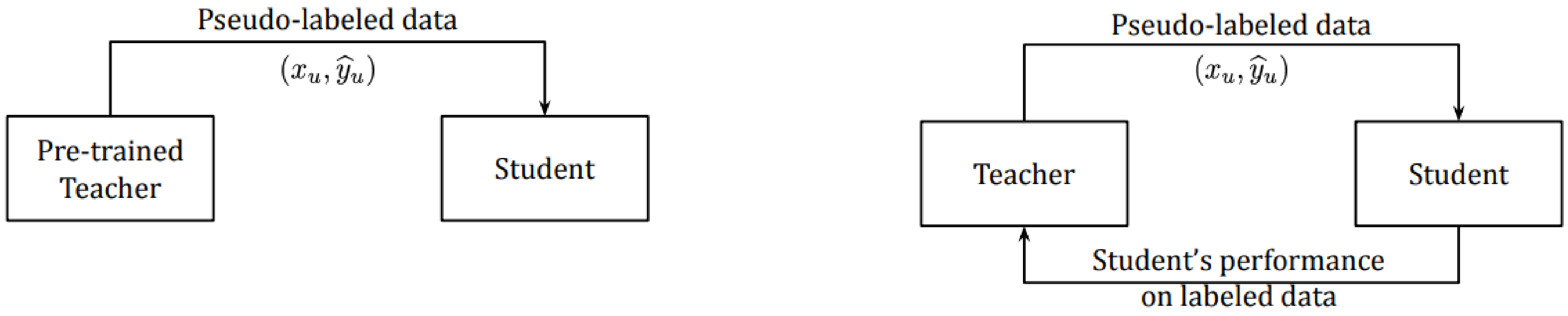


Figure 1: The difference between Pseudo Labels and Meta Pseudo Labels. **Left:** Pseudo Labels, where a fixed pre-trained teacher generates pseudo labels for the student to learn from. **Right:** Meta Pseudo Labels, where the teacher is trained along with the student. The student is trained based on 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).

Proposed Method: Meta – Pseudo Label

Why this is necessary?

➔ Teacher model is not fixed but adopted by the feedback of Student

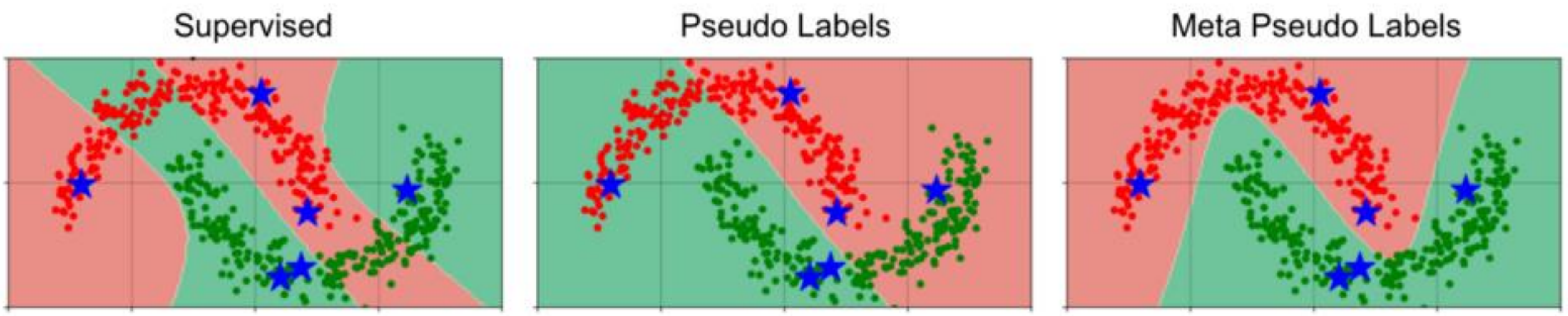


Figure 2: An illustration of the importance of feedback in Meta Pseudo Labels (right). In this example, Meta Pseudo Labels works better than Supervised Learning (left) and Pseudo Labels (middle) on the simple TwoMoon dataset. More details are in Section 3.1.

임진혁<eeplearning@unist.ac.kr>

Proposed Method: Meta – Pseudo Label

Why this is necessary?

➔ Teacher model is not fixed but adopted by the feedback of Student

Why this is necessary?*2

➔ Pseudo Label's limitation : Confirmation bias problem

[Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning](2020,Paul etc.) say
"It is natural to think that reducing the confidence of the network by artificially changing the labels might alleviate this problem and improve generalization"

Proposed Method: Meta – Pseudo Label

Why this is necessary?

➔ **Teacher model is not fixed but adopted by the feedback of Student**

Why this is necessary?*2

➔ **Pseudo Label's limitation : Confirmation bias problem**

Meta Pseudo Label :

Teacher correct that bias by observing how its pseudo labels would affect the student

임진혁<eeplearning@unist.ac.kr>

Proposed Method: Meta – Pseudo Label

Meta Pseudo Label :

Teacher correct that bias by observing how its pseudo labels would affect the student for the teacher to generate better pseudo labels .

Feedback from student = Performance of the student on labeled dataset
Used as reward to train teacher

임진혁<eeplearning@unist.ac.kr>

Proposed Method: Meta – Pseudo Label

Feedback from student = Performance of the student on labeled dataset

Used as reward to train teacher

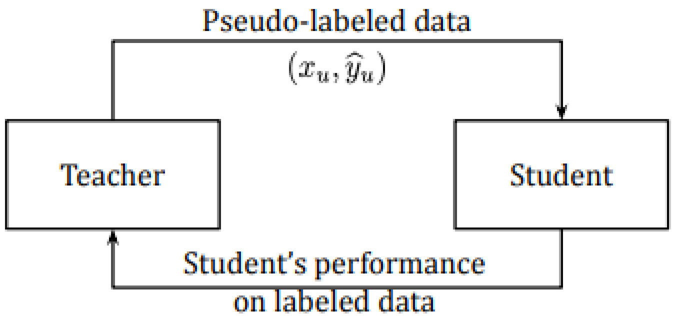
θ_T θ_S

T/S : Teacher Model / Student Model

$$Model(x_{label?}; \theta_{model})$$

=> Soft predictions of the model(T or S) on the batch x(label or unlabel)

CE(q, p) : cross-entropy loss on average of all instances in the batch



Proposed Method: Meta – Pseudo Label

Feedback from student = Performance of the student on labeled dataset

Used as reward to train teacher

$\mathrm{CE}\big(y_l, S(x_l;\theta_S)\big)$: ?

Proposed Method: Meta – Pseudo Label

Feedback from student = Performance of the student on labeled dataset

Used as reward to train teacher for

θ_S^{PL} achieve a low loss on labeled data

$$\boxed{\mathbb{E}_{x_l, y_l}\Big[\text{CE}\big(y_l, S(x_l; \theta_S^{\text{PL}})\big)\Big] := \mathcal{L}_l\big(\theta_S^{\text{PL}}\big).}$$

Pseudo Label Loss(PL) : Student model's loss on unlabeled data

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

임진혁<eeplearning@unist.ac.kr>

Proposed Method: Meta – Pseudo Label

Feedback from student = Performance of the student on labeled dataset

Used as reward to train teacher for

θ_S^{PL} **achieve a low loss on labeled data**

Always depend on the Teacher Model Parameter
(Via the pseudo targets)

$$\boxed{\mathbb{E}_{x_l,y_l}\left[\mathrm{CE}\left(y_l, S(x_l;\theta_S^{\mathrm{PL}})\right)\right] := \mathcal{L}_l\left(\theta_S^{\mathrm{PL}}\right).} \Longrightarrow \mathcal{L}_l\left(\theta_S^{\mathrm{PL}}(\theta_T)\right)$$

Pseudo Label Loss(PL) : Student model's loss on unlabeled data

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

Pseudo targets
(1): well pre-trained teacher model with fixed parameter

Proposed Method: Meta – Pseudo Label

In short, we optimize

$$\min_{\theta_T} \quad \mathcal{L}_l \left(\theta_S^{\mathrm{PL}}(\theta_T)\right),$$

$$\boxed{\text{where} \quad \theta_S^{\mathrm{PL}}(\theta_T) = \operatorname*{argmin}_{\theta_S} \mathcal{L}_u\left(\theta_T, \theta_S\right).}$$

This result pseudo labels can be adjusted to improve student's performance

임진혁<eeplearning@unist.ac.kr>

Proposed Method: Meta – Pseudo Label

$$\min_{\theta_T} \mathcal{L}_l\left(\theta_S^{\mathrm{PL}}(\theta_T)\right),$$

$$\text{where} \quad \theta_S^{\mathrm{PL}}(\theta_T) = \underset{\theta_S}{\operatorname{argmin}} \, \mathcal{L}_u(\theta_T, \theta_S).$$

Practical approximation. via Meta Learning

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

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

(treat θ(S) as fixed parameters and ignore its dependency on θ(T))

REINFORCE ?

임진혁<eeplearning@unist.ac.kr>

Proposed Method: Meta – Pseudo Label

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

실제적으로는 어떻게 적용되는가?

- After Meta-Train Phase , Finetune Student Model on labeld data to improve accuracy
- Both the teacher and the student share same architecture but have independent weights

?

임진혁<eeplearning@unist.ac.kr>

4 Experiment

	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

Method	Unlabeled Images	Accuracy (top-1/top-5)
Supervised [24]	None	76.9/93.3
AutoAugment [12]	None	77.6/93.8
DropBlock [18]	None	78.4/94.2
FixRes [68]	None	79.1/94.6
FixRes+CutMix [83]	None	79.8/94.9
NoisyStudent [77]	JFT	78.9/94.3
UDA [76]	JFT	79.0/94.5
Billion-scale SSL [68, 79]	YFCC	82.5/**96.6**
Meta Pseudo Labels	JFT	**83.2**/96.5

Table 3: Top-1 and Top-5 accuracy of Meta Pseudo Labels and other representative supervised and semi-supervised methods on ImageNet with ResNet-50.

임진혁<eeplearning@unist.ac.kr>

Conclusion

Method	# Params	Extra Data	ImageNet Top-1	ImageNet Top-5	ImageNet-ReaL [6] Precision@1
ResNet-50 [24]	26M	–	76.0	93.0	82.94
ResNet-152 [24]	60M	–	77.8	93.8	84.79
DenseNet-264 [28]	34M	–	77.9	93.9	–
Inception-v3 [62]	24M	–	78.8	94.4	83.58
Xception [11]	23M	–	79.0	94.5	–
Inception-v4 [61]	48M	–	80.0	95.0	–
Inception-resnet-v2 [61]	56M	–	80.1	95.1	–
ResNeXt-101 [78]	84M	–	80.9	95.6	85.18
PolyNet [87]	92M	–	81.3	95.8	–
SENet [27]	146M	–	82.7	96.2	–
NASNet-A [90]	89M	–	82.7	96.2	82.56
AmoebaNet-A [52]	87M	–	82.8	96.1	–
PNASNet [39]	86M	–	82.9	96.2	–
AmoebaNet-C + AutoAugment [12]	155M	–	83.5	96.5	–
GPipe [29]	557M	–	84.3	97.0	–
EfficientNet-B7 [63]	66M	–	85.0	97.2	–
EfficientNet-B7 + FixRes [70]	66M	–	85.3	97.4	–
EfficientNet-L2 [63]	480M	–	85.5	97.5	–
ResNet-50 Billion-scale SSL [79]	26M	3.5B labeled Instagram	81.2	96.0	–
ResNeXt-101 Billion-scale SSL [79]	193M	3.5B labeled Instagram	84.8	–	–
ResNeXt-101 WSL [42]	829M	3.5B labeled Instagram	85.4	97.6	88.19
FixRes ResNeXt-101 WSL [69]	829M	3.5B labeled Instagram	86.4	98.0	89.73
Big Transfer (BiT-L) [33]	928M	300M labeled JFT	87.5	98.5	90.54
Noisy Student (EfficientNet-L2) [77]	480M	300M unlabeled JFT	88.4	98.7	90.55
Noisy Student + FixRes [70]	480M	300M unlabeled JFT	88.5	98.7	–
Vision Transformer (ViT-H) [14]	632M	300M labeled JFT	88.55	–	90.72
EfficientNet-L2-NoisyStudent + SAM [16]	480M	300M unlabeled JFT	88.6	98.6	–
Meta Pseudo Labels (EfficientNet-B6-Wide)	390M	300M unlabeled JFT	90.0	98.7	**91.12**
Meta Pseudo Labels (EfficientNet-L2)	480M	300M unlabeled JFT	**90.2**	**98.8**	91.02

Table 4: Top-1 and Top-5 accuracy of Meta Pseudo Labels and previous state-of-the-art methods on ImageNet. With EfficientNet-L2 and EfficientNet-B6-Wide, Meta Pseudo Labels achieves an improvement of 1.6% on top of the state-of-the-art [16], despite the fact that the latter uses 300 million *labeled* training examples from JFT.

임진혁<eeplearning@unist.ac.kr>

감사합니다