영상 분류를 위한 준지도 학습 기법의 분류와 동작 원리의 이해

□ 채문주, 박재현, 조성인 / 동국대학교

요약

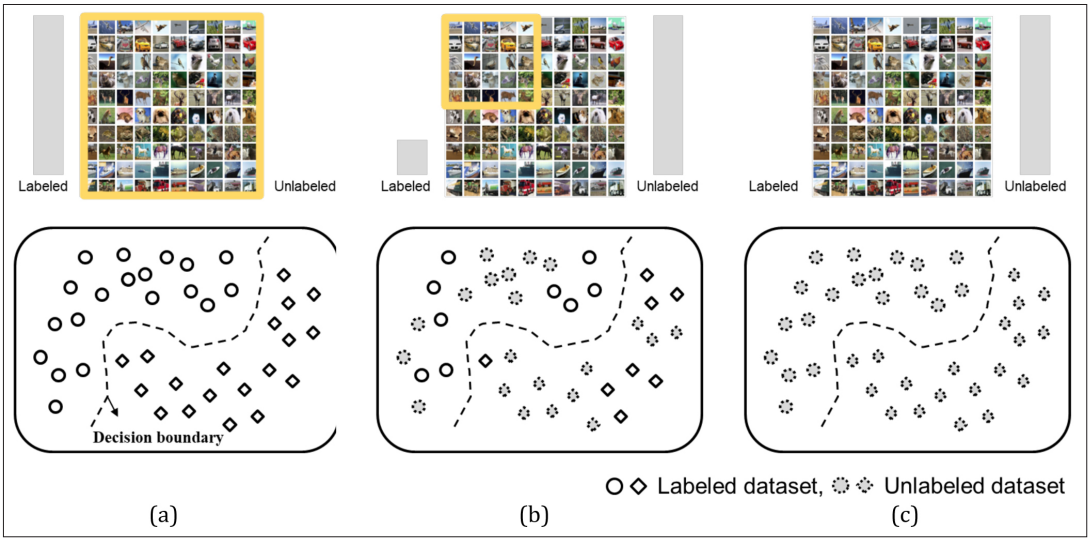
본 고에서는 준지도 학습의 개념과 목표 그리고 대표 기법들의 동작 원리에 대해서 알아본다. 구체적으로, 영상 분류를 위한 준지도 학습 기법을 크게 label propagation 기반 기법과 representation learning 기반 기법으로 나누고, 이 두 가지 기법들의 특성을 분석하고, 대표 기법들의 동작 원리에 대해서 설명한다. 또한, 영상 분류 문제에서 위 두 가지 접근법들의 대표 기법들의 성능을 평가한다.

I. 서 론

최근 딥러닝 기술의 비약적인 발전으로 인하여, 다양한 분야에서 딥러닝 기술들이 활발히 사용되고 있다. 딥러닝 기술은 레이블이 포함된 훈련 데이터가 많을수록 더 우수한 성능을 낼 수 있는 것으로 알려져 있다. 하지만, 데이터와 이에 대응되는 레이블을 도출하는 것은 엄청난 비용이 필요하므로, 적은 수의 레이블 데이터만으로도 과적합 문제없이 우수한 영상 분류 결과를 딥 뉴럴 네트워크로부터 도출하는 것에 대한 필요성이 대두되고 있다. 이에, 적은 수의 데이터만 레이블 정보의 사용이 가능할 때, 레이블이 없는 데이터를 추가로 활용하여 딥 뉴럴 네트워크의 성능을 향상시키기 위한 방법들이 개발되고 있고, 이를 준지도 학습이라고 한다. 〈그림 1〉은 영상 분류 작업에서, 준지도 학습을 포함한 3가지 대표적인 학습 기법을 나타낸다.

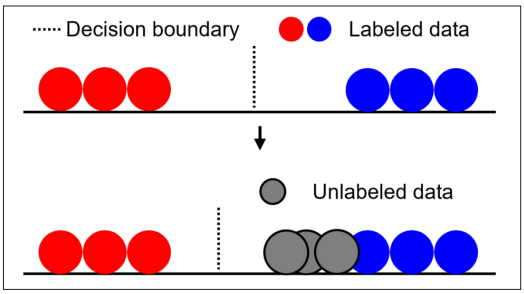
가장 일반적인 학습법은 〈그림 1〉의 (a)와 같이 레이블 정보와 입력 영상을 모두 사용하여 딥 뉴럴 네트워크를 학습하는 지도 학습 방법이다. 〈그림 1〉의 (c)는 레이블 정보 없이 입력 영상만을 이용하여 딥 뉴럴 네트워크를 학습시키는 비지도 학습 기법이고, 〈그림 1〉

※ 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1C1C1009662, NRF-2020X1A3A109).
※ This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning) (No. 2020R1C1C1009662, NRF-2020X1A3A109).



<그림 1> 학습 기법의 분류 [1] (a) 지도학습 (b) 준지도 학습 (c) 비지도 학습

의 (b)는 적은 수의 레이블 정보와 다수의 레이블 정보가 없는 영상을 이용하여 딥 뉴럴 네트워크를 학습하는 준지도 학습 기법이다. 궁극적으로 준지도 학습의 목적은 레이블 정보를 활용하여 훈련된 딥 뉴럴 네트워크의 성질을 레이블이 없는 데이터를 활용하여 변화시키는 것이다. 이와 관련된 1차원 공간에서의 가장 기초적인 예시를 아래 〈그림 2〉에서 나타내고 있다. 〈그림 2〉에서 붉은색과 푸른색의 원은 레이블이 있는 데이터이고, 이 데이터를 활용하여 decision boundary를 도출하면 그림에서 위쪽 점선과 같다. 만약 회색으로 표현된 레이블이 없는 데이터를 활용할 경우 decision boundary를 전체적인 데이터 분포를 고려하여 조정이 아래쪽 그림처럼 이루어질 수 있다. 준지도 학습은 이러한 조정이 가능하게 하며, 추후 설명될 딥 뉴럴 네트워크에서는 비단 decision boundary의 조정 뿐 아니라 representation space에서의 분포 역시 조정하는 역할을 수행한다. 하지만, 레이블이 없는 데이터가 입력 도메인에서 균등한 분포를 가질 경우는 위와 같은 역할을 수행할 수 없기 때문에, 목표로 하는 작업과 관련된 레이블이 없는 데이터의 분포 특성은 준지도 학습 기법 성능에 큰 영향을 줄 수 있고, 이를 고려한 기법 설계가 필요하다.



<그림 2> 레이블이 없는 데이터를 활용한 decision boundary의 조정 예제

준지도 학습 기법들은 다양한 작업에 적용될 수 있는데, 그 중에서도 영상 분류를 목표로 많은 연구가 수행되었다. 이어질 본론에서는 딥러닝 기반 영상 분류를 위한 준지도 학습 기법을 크게 label propagation 기반 기법과 representation learning 기반 기법으로 분류하고 각 접근법의 특성을 설명한다. 이후 각 접근법의 대표 기법들의 동작 원리를 소개한다.

II. 본 론

딥러닝 기반 영상 분류를 위한 준지도 학습은 다양한 기준을 기반으로 다수의 접근법으로 분류해 볼 수 있다. 이러한 분류법들 중에 우리는 전술한 것과 같이 크게 label propagation 기반 기법[2-5, 9-13]과 representation learning[6-8, 14]기반 기법으로 준지도 학습 기법을 분류한다. Label propagation 기법은 훈련된 모델로부터 도출되는 레이블이 없는 데이터의 레이블 추정 결과를 직·간접적으로 모델에 반영하는 기법이다. Representation learning 기반 기법으로 분류되는 기법들은 딥 뉴럴 네트워크에서의 representation space 상에서의 샘플별 분포를 조정하여 적은 수의 레이블 정보만으로도 우수한 영상 분류 성능을 달성한다.

1. Label Propagation 기반 기법

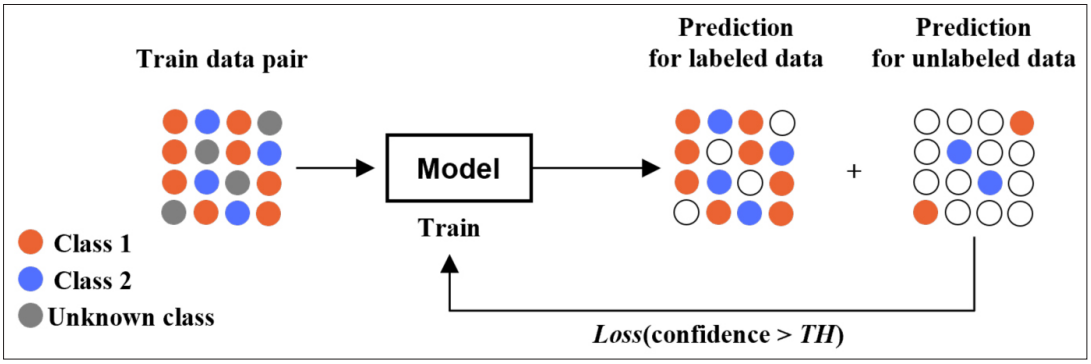
Label propagation 기반 기법은 다양한 방법이 존재하지만, 본 고에서는 아래와 같이 pseudo labeling기법과 consistency regularization 기법을 소개한다.

1) Pseudo Labeling[2]

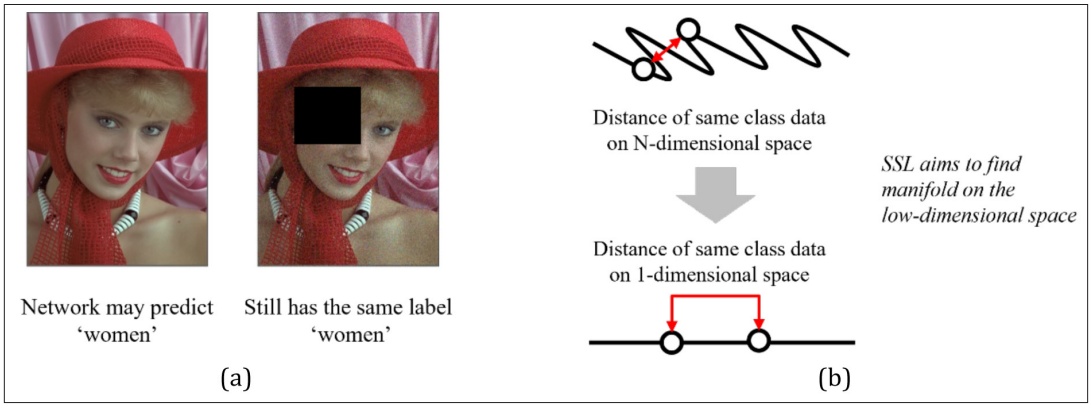
Pseudo labeling 기법은 〈그림 3〉과 같이 적은 수의 레이블 정보를 활용하여 먼저 지정된 딥 뉴럴 네트워크를 지도학습을 통해 훈련한다. 훈련된 모델에 레이블이 없는 샘플을 입력으로 제공하고 예측 결과들 중 confidence가 높은 샘플을 선별하여 훈련 데이터로 귀속시켜 훈련을 반복한다. 이를 통하여 부족한 수의 레이블 정보를 보완하고, 부족한 레이블 정보로 인한 과적합 문제를 해결하고자 한다. 다만, 잘못된 pseudo label은 모델의 성능을 크게 저하시킬 수 있다. 또한 잘못된 pseudo label이 훈련에 반영되지 않도록 pseudo label 도출을 위한 confidence 임계값을 높일수록 pseudo label을 활용한 레이블이 없는 데이터의 활용 가치가 크게 떨어질 수 있다.

2) Consistency Regularization[4-5, 9-13]

Consistency regularization 기법은 〈그림 4〉의 (a)처럼 훈련 샘플에 섭동이 존재하더라도 최종 추정 결과는 동일해야 한다는 간단한 원리를 이용한다. 특히 레이블이 없는 영상 샘플에는 레이블 정보를 사용하지 않고 위와 같은 섭동을 다양한 형태로 적용하여 섭동 적용 전



<그림 3> Pseudo labeling의 과정



<그림 4> Consistency regularization의 (a) motivation과 (b) 최종 목표

후의 영상에 대한 추정 결과에 대한 차이를 훈련 loss로 사용할 수 있다. Consistency regularization을 활용하는 준 지도 학습 기법 중 가장 널리 알려진 pi 모델[4]은 아래와 같이 정의되는 훈련 loss를 활용한다.

$$loss = -\frac{1}{|B|}\sum_{i\in(B\cap C)}[y_i]\log(z_i) + w(t)\frac{1}{C|B|}\sum_{i\in B}\|z_i - \tilde{z}_i\|^2 \quad (1)$$

여기서 B는 batch 내 sample의 수를, z_i, $\tilde{z}_i$는 i번째 입력 샘플과 해당 샘플에 섭동이 적용된 결과를 나타낸다. $w(t)$는 영상 분류를 위하여 일반적으로 사용되는 cross-entropy loss(수식 (1)의 왼쪽)와 consistency regularization loss(수식 (1)의 오른쪽)의 balancing을 위한 weight 값으로 훈련이 진행될수록(t 값이 커질수록) 값이 커지는 ramp-up 함수로 설정된다.

Pi 모델 이외에도 adversarial attack을 활용하는 virtual adversarial training(VAT)[5], mean-teacher (MT) 구조를 활용한 기법[4] 다양한 augmentation을 활용하는 기법[10-12]들이 있다. 이렇게 다양한 consistency regularization 기법들은 기존의 훈련에 추가된 consistency regularization loss는 레이블이 없는 샘플을 활용하여 low-dimensional manifold space를 찾을 수 있도록 하고, 최종적으로 영상 분류의 정확도를 향상을 유도한다.

2. Representation Learning 기반 기법

Representation learning 기반의 준지도 학습은 영상 분류에 효과적인 representation을 도출해내는 것에 집중한다. 첫 번째로 소개할 predicting view assignments with support samples(PAWS)[6]의 경우 representation space 상에서 pseudo label과 유사한 정보를 기반으로 모델을 학습하게 되고, simple framework for contrastive learning of visual representations(SimCLR)[7]와 bootstrap your own latent(BYOL)[8]은 자기 지도 학습(self-supervised training) 기법을 기반으로 representation space를 구성한다.

1) Predicting View Assignments with Support Samples(PAWS)[6]

먼저 PAWS는 약한 augmentation과 강한 augmentation을 입력 영상에 적용하여 두 가지 출력(positive

view, anchor view)을 생성한다. 이후 모델의 feature extractor를 통과시켜 두 출력의 representations(z^+, z)을 도출한다. 이후 레이블이 있는 샘플(support sample)에 동일한 feature extractor에 통과시켜 z_s를 도출하고, representation space 상에서 (z^+, z) 와 (z_s) 유사도를 기반으로 positive view와 anchor view의 predictions(p^+, p)을 도출한다. 이 과정을 통하여 레이블이 있는 데이터의 정보가 통계적으로 p^+, p 도출에 활용될 수 있도록 한다. 이후 아래 수식을 이용하여 모델을 훈련한다.

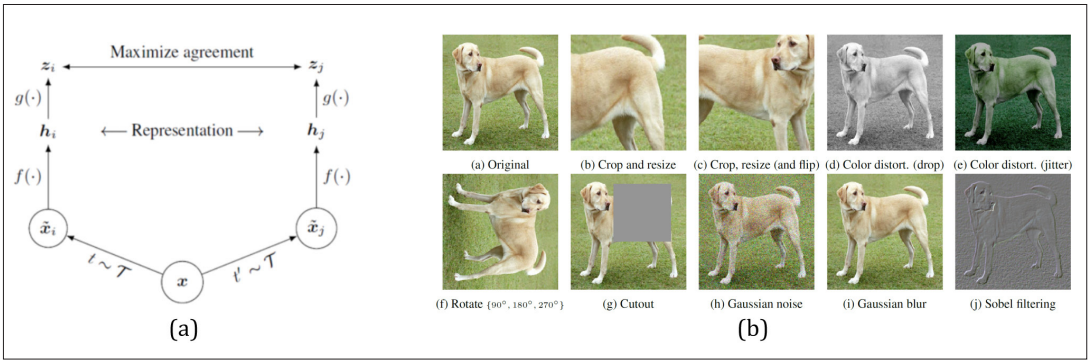
$$L = \frac{1}{2n}\sum_{i=1}^{n}\left(H\left(\rho\left(p_i^+\right), p_i\right) + H\left(\rho\left(p_i\right), p_i^+\right)\right) - H\left(\bar{p}\right)$$
$$\bar{p} = \frac{1}{2n}\sum_{i=1}^{n}\left(\rho\left(p_i\right) + \rho\left(p_i^+\right)\right) \quad (2)$$

여기서 i는 샘플 인덱스를 나타내고, H는 cross entropy 함수를, ρ 는 sharpening 함수를 나타낸다. n은 batch 내 샘플의 수를 나타낸다. 수식에서 볼 수 있듯이, positive view와 anchor view의 추정결과의 차이를 줄이면서 정규화 요소를 반영한다. 따라서 PAWS의 경우 consistency regularization 기반 기법으로도 분류될 수 있으나, 본 논문에서는 representation space 상에서의 유사도 분석을 기반으로 모델이 훈련되는 것을 고려하여 representation learning 기반 기법으로 분류하였다.

2) Framework for Contrastive Learning of Visual Representations(SimCLR)[7]

SimCLR 기법은 contrastive learning을 활용하여 일반화가 성공적으로 이루어졌으면서도, 영상 분류에 효과적인 visual representation을 도출한다. Contrastive learning은 같은 입력에 대해서는 유사한 representation을 제공하고, 서로 다른 입력에 대해서는 다른 representation을 제공하는 방향으로 feature extractor를 훈련하는 방식이다. 따라서 레이블이 없는 입력에 대해서도 적용이 가능하고, 자기 지도 학습 기법에서 pretrained model의 생성과정으로 이해할 수 있다. 자기 주도 학습 기법은 비지도 학습으로 분류될 수 있지만, pretrained model을 생성하고, 추후 적은 수의 레이블 정보를 이용한 fine tuning(downstream task)을 적용한다면 준지도 학습 기법으로 이해될 수 있다.

〈그림 5〉의 (a)는 SimCLR의 구조를 보여준다. 주



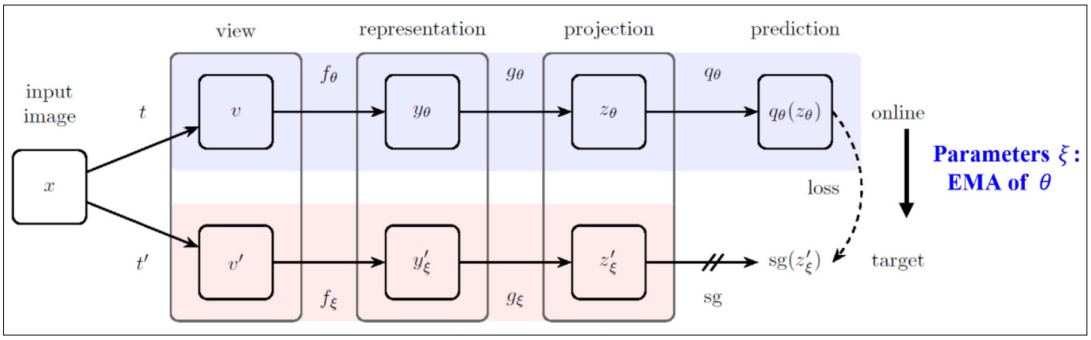
<그림 5> SimCLR의 (a) 구조 (b) augmentation의 종류[7]

어진 입력 영상 x에 대해서 서로 다른 augmentation을 적용 (〈그림 5〉 (b))하여 두 가지 형태의 입력을 생성하고 base encoder인 f를 통과시켜 h_i와 h_j를 생성한다. 논문에서는 ResNet 50을 f로 사용하였다. 이후 128개의 latent space를 활용하는, 두 개의 레이어로 이루어진 multi-layer perceptron(MLP)으로 이루어진 projection head인 g를 통과시켜 최종 출력 z_i와 z_j를 생성한다. 이후 동일한 입력의 z_i와 z_j의 (positive pair) 유사도는 향상시키면서, 나머지 샘플 쌍(negative pair)에 대해서는 유사도를 저하시는 방향으로 모델이 학습된다. 각 batch 에서는 sample별로 1쌍의 positive pair와 2N-2쌍의 negative pair를 구성할 수 있다. Batch의 크기는 커질수록 우수한데, 이는 다양한 negative pair를 도출할 수 있기 때문이다. SimCLR에서는 다양한 batch 크기(256~8192)를 사용하여 실험을 진행하여, 영상 분류 작업에 최적화된 representation을 도출하였다. 이후, 최적화된 representation은 간단한 classifier에 대해서만 훈련하여 결과를 도출하여도 매우 우수한 영상 분류 결과를 제공하게 된다. 앞서 설명한 과정을 통하여 훈련된 모델에 작은 수의 레이블 샘플을 활용하여 fine-tuning 하여 모델을 평가하게 되면 준지도 학습으로의 성능 평가가 가능하다.

3) Bootstrap Your Own Latent(BYOL)[8]

앞서 설명한 SimCLR의 경우 negtive pair를 풍부하게 도출하기 위해서는 batch의 크기를 매우 크게 설정해야 한다. 실제로 SimCLR에서 최대 8192의 batch 사이즈를 사용하였는데, 이는 실제 구현에 매우 큰 부담이 될 수 있다. 이에 BYOL은 두 개의 모델을 활용하여 메모리 부담을 줄이면서 augmentation에 의존성을 줄이기 위한 기법을 제안한다.

BYOL은 random한 파라미터로 초기화된 모델(target)에서 도출된 정확도가 매우 낮은 추정 값을 다른 모델(online)이 학습하게 하고, 간단한 classifier를 학습시켜 보면 target 모델의 정확도보다 향상된 결과를 도출할 수 있다는 것에 기인하여 〈그림 6〉과 같은 구조를 제안하게 된다. 그림에서 볼 수 있듯이, online과 target model은 모두 representation 도출을 위한 모듈(f)과 projection 모듈(g)을 포함한다. Online 모델은 마지막으로 target 모델의 출력을 재현해내기 위한 prediction 모듈이 추가된다. Target 모델은 gradient 기반으로 파라미터를 업데이트하지 않고, online 모델의 파라미터를 moving average한 값을 사용한다. Online 모델은 식 (2)에서 정의한 loss 함수를 활용하여 업데이트를 진행한다. 두 모델의 입력은 각각 다른



<그림 6> BYOL의 구조[8]

augmentation을 적용한 영상을 사용한다.

$$\mathcal{L} = \mathcal{L}_{\theta}^{BYOL} + \tilde{\mathcal{L}}_{\theta}^{BYOL},$$
$$\mathcal{L}_{\theta}^{BYOL} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z'_\xi} \right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\xi\|_2} \quad (3)$$

위 수식에서 볼 수 있듯이, target 모델의 결과를 online 모델이 재현해 낼 수 있도록 하고, 각 모델의 입력을 서로 교환하여서도 동일하게 loss를 계산하여 적용한다. 이를 통하여, 영상 분류에 적합한 representation을 학습하게 된다. 해당 방법은 준지도 학습으로 확장이 가능한데, 위 과정을 통하여 representation을 학습하고, 작은 수의 레이블 샘플을 활용하여 fine-tuning 하여 모델 영상 분류 성능을 평가해 볼 수 있다.

III. 실험 결과

앞서 설명한 방법들의 성능을 평가하기 위해서, 우리는 영상 분류에서 널리 사용되는 CIFAR-10[15], SVHN[16], ImageNet[17]을 사용하되, 레이블의 수를 제한하여 훈련을 진행하고, 평가를 진행한다. 〈표 1〉에서 각 실험 결과는 각 방법의 논문에서 발췌하였고, 일부는 배포된 코드를 통하여 도출하였다. CIFAR-10은 4000개, SVHN은 1000개, ImageNet은 1%, 10%의 레이블 샘플만 활용하여 훈련을 진행하고 결과를 도출하였다. 방법은 전술했던 것과 같이, label propagation 기반 기법과 representation learning 기반 기법으로 분류하여 성능을 평가하였다. 표에서 볼 수 있듯이, CIFAR-10과 SVHN 데이터 세트에서 representation learning 기반 기법인 PAWS가 가장 우수한 결과를 제공하는 것을 볼 수 있다. ImageNet에서는 SimCLR 기반 기법과 BYOL 기법이 대체로 우수한 결과를 제공하는 것을 확인할 수 있었다.

IV. 결 론

본 고에서 우리는 영상 분류를 위한 대표적인 준지도 학습 기법에 대해서 크게 두 가지 접근법으로 나

<표 1> 비교 방법들의 분류 에러 (%)

Methods		CIFAR-10[15]	SVHN[16]	ImageNet[17]	
Labels		4000	1000	1%	10%
Label Propagation	Π-model[4]	12.36 ± 0.31	4.82 ± 0.17	-	-
	TE[4]	12.16 ± 0.24	4.42 ± 0.16	-	-
	MT[9]	12.31 ± 0.28	3.95 ± 0.19	-	-
	FixMatch[10]	4.26 ± 0.05	2.28 ± 0.11	-	-
	MixMatch[11]	4.95 ± 0.08	3.27 ± 0.31	-	-
	UDA[12]	4.32 ± 0.08	2.23 ± 0.07	-	31.22
	MPL[13]	3.89 ± 0.07	1.99 ± 0.07	-	26.11
Representation Learning	SimCLR[7]	-	-	35.9	23.5
	SimCLRv2[14] (self-distil)	-	-	-	19.9
	BYOL[8]	-	-	28.8	22.3
	PAWS[6]	4.0 ± 0.2	*1.99±0.21	33.5	24.5
Labeled only		*18.4	*14.2	-	-

누어 그 동작 원리를 살펴보았다. 첫 번째 접근법인 label propagation 기반 기법으로 pseudo labeling과 consistency regularization 방법을 소개하였다. 이 기법들은 적은 수의 레이블 샘플로부터 모델을 학습하면서, 학습된 모델이 추정해내는 결과를 훈련에 반영할 수 있는 기법이었다. 두 번째 접근법인 representation learning 기반 방법은 영상 분류에 적합한 representation을 적은 수의 레이블 데이터를 가지고도 효과적으로 도출해는 것이 핵심이었다. 실험 결과에서 볼 수 있듯이, 준지도 학습 환경에서 일반적으로 representation learning 기반 기법들이 영상 분류에서 우수한 성능을 제공할 수 있음을 확인할 수 있었다.

참 고 문 헌

[1] J. H. Park, "Pseudo-labeling Technique for Image Classification with Limited Labeled Data," M.S. thesis, Dept. Multimedia Engineering, Dongguk University, Seoul, Republic of Korea, 2021, Available: http://lib.dongguk.edu/search/detail/CATTOT000001239284

[2] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in Proc. Workshop Challenges Represent. Learn. (ICML), vol. 3, 2013. p. 2.

[3] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2004, pp. 529536.

[4] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in Proc. Int. Conf. Learn. Represent. (ICLR), 2017, pp. 113.

[5] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," IEEE Trans. Pattern Anal. Mach. Intell., Vol. 41, No. 8, pp. 19791993, Aug 2019.

[6] M. Assran, M. Caron, I. Misra, P. Bojanowski, and A. Joulin, "Semi-Supervised Learning of Visual Features by Non-Parametrically Predicting View Assignments with Support Samples," arXiv preprint arXiv:2104.13963, 2021.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," In International conference on machine learning. PMLR, 2020. pp. 1597-1607.

[8] J.-B. Grill et al., "Bootstrap Your Own Latent A New Approach to Self-Supervised Learning," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Vancouver, Canada, 2020.

[9] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in Proc. Int. Conf. Learn. Represent. (ICLR), 2017, pp. 1-16.

[10] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semisupervised learning with consistency and confidence," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2020, pp. 1-21.

[11] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2019, pp. 5050-5060.

[12] Q. Xie, Z. Dai, E. H. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2020.

[13] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," In Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR), 2021, pp. 11557-11568.

[14] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big Self-Supervised Models are Strong Semi-Supervised Learners," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Canada, 2020.

[15] A. Krizhevsky and G. Hinton, “Learning Multiple Layers of Features from Tiny Images,” technical report, Univ. of Toronto, 2009.
[16] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” In NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.
[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li. ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision, 2015.

필 자 소 개

채 문 주

- 2022년 : 동국대학교 멀티미디어공학과 학사
- 2022년 ~ 현재 : 동국대학교 멀티미디어공학과 석사
- 주관심분야 : 딥러닝 영상 분류, 3D 데이터 처리

박 재 현

- 2019년 : 대구대학교 전자공학과 학사
- 2021년 : 동국대학교 멀티미디어공학과 석사
- 2021년 ~ 현재 : 동국대학교 멀티미디어공학과 박사
- 주관심분야 : 영상처리, 딥러닝 영상 분류, 특징 공학

조 성 인

- 2010년 : 서강대학교 전자공학과 학사
- 2015년 : 포항공과대학교 전자전기공학부 박사
- 2017년 : LG 디스플레이 선임연구원
- 2019년 : 대구대학교 전자전기공학부 조교수
- 2019년~ 현재 : 동국대학교 멀티미디어공학과 조교수
- 주관심분야 : 영상처리, 컴퓨터 비전, 딥러닝