Self-training 기술 동향 분석

황유진(Y.J Hwang, yujine92@gmail.com)

목차

I. 머리말

딥러닝은 지난 기간 동안 수많은 발전을 이루어왔다. ImageNet과 같은 대용량 데이터 셋은 딥러닝의 성능 향상에 큰 기여를 하였다. 데이터 셋의 구축 과정은 일반적으로 레이블링 생성 과정에서 사람이 개입하기 때문에 비용이 크다. 따라서 ImageNet 이후 대용량 데이터 셋의 공개가 적었으며, 모델의 표현력 증가와 연산능력의 증가 속도에 비해 데이터 셋의 발전속도가 느려 최근에는 데이터 셋의 부족이 모델 성능 발전 병목현상의 원인이 되고 있다.

데이터 셋 부족으로 인한 딥러닝 기술 발전병목을 해결하기 위하여 효율적으로 레이블링을 진행하기 위해 모델이 직접 학습할 데이터를 선정하여 레이블링을 요청하는 Active Learning, 레이블링이 없는 데이터로 학습을 진행하는 비지도학습(Unsupervised Learning)이 있으며 지도학습(Supervised Learning) 과정과 유사하나 레이블링이 없는 데이터까지 이용하여 효율성을 최대화 한 준지도학습(Semi-supervised Learning) 과 같은 연구가 진행되고 있다.

본고에서 다루고자 하는 Self-training은 준지도학습 방법에 속하며 Pseudo-labeling으로도 불린다. 해당 방법론을 통해서 준지도학습 분야에서 state-of-the-art 성능을 보이는 연구들[1,2,3]을 소개한다. 본고의 II장에서는 준지도학습 기술 개괄을 소개를 하고 III장에서는 Self-training에 속하는 최신 연구를 소개한다. IV장에서는 데이터 셋 구축 비용을 줄이기 위한 다양한 연구로 Active Learning을 소개하며 마지막으로 V장에서는 결론을 맺는다.

II. 준지도학습 기술 개괄

딥러닝은 크게 지도학습과 비지도학습으로 분류할 수 있다. 딥러닝의 가장 대표적인 학습 방식인 지도학습은 모델의 입력이 되는 x데이터와 예측 값인 y 데이터를 쌍으로 갖고 있는 Labeled Data를 통해 분류기를 학습하는 방법으로, x를 입력으로 할 때 y를 출력하도록 학습한다. 그러나 비지도학습의 경우는 입력이 되는 x 데이터로만 구성된 Unlabeled Data에서 x에 내재되어 있는 분포를 학습하는 학습방식이다. 비지도 학습의 경우 지정된 y값이 없어 지도학습 보다 학습난이도가 어려운 것이 일반적이다.

준지도학습 방법론은 위의 지도학습 방법론과 비지도학습 방법론을 결합한 방법론으로 Labeled Data를 이용할 뿐만 아니라 Unlabeled Data까지 이용하여 모델의 성능을 최대한 향상시키는 것이 목적이다. 최근 [4]에서 정의한 준지도학습의 대표적인 학습 방법은 다음과 같다. (1) Inductive methods (2) Graph-based method. (1) 의 방법론은 input space에 속하는 어떠한 데이터든 예측할 수 있는 분류기를 학습하는 것을 목적으로 한다. 이와 다르게 (2)은 학습 중에 사용한 데이터만을 예측할 수 있다. 따라서 3)은 기존의 train data로 학습하고 test 데이터로 성능 평가를 진행하는 딥러닝의 성능 평가 방식으로 모델의 성능 평가를 진행할 수 없다는 단점이 있다. (1)의 Inductive method는 다음의 두가지로 다시 나누어진다. Wrapper methods, Intrinsically semi-supervised method.

Wrapper method는 지도학습방법론과 가장 유사한 학습 프로세스를 가졌다. 학습

프로세스는 다음과 같다. 먼저 Labeled data에 대하여 모델 학습을 진행한 후 Unlabeled data에 대한 예측을 진행한다. 이후 Unlabeled data에 대한 예측을 pseudo-label로 하여 (x, y) 쌍을 구성한 후 해당 데이터와 기존 Labeled data를 합하여 모델의 재학습을 진행한다. 즉, x값으로만 구성된 Unlabeled data에 대해 y값을 생성하여 지도학습방식과 유사하게 학습을 진행하는 방식이다. III장에서 다룰 Self-training도 해당 방법론에 속한다.

Intrinsically semi-supervised method는 쌍으로 구성된 (x,y) 데이터를 이용하여 지도학습기반 방법론과 유사하게 학습하는 Wrapper method와 다르게 레이블이 없는 x 데이터로도 학습 가능한 목적함수를 구성한다. 해당 방법론은 보통 지도학습방식을 Unlabeled data에 그대로 확장하여 적용한다. 예를 들어 기존 support vector machine이 마진을 최대화하는 방식으로 학습한다면 이를 Unlabeled data에도 확장하여 결정경계가 데이터와의 거리 마진을 최대화하도록 학습한다.

III. 대표적인 Self-training 연구

1. Pseudo-label[5]

Pseudo-label[5]는 2013년에 소개되었으며, DNNs을 준지도학습방식으로 훈련시키기 위해 연구되었다. 해당 연구는 MNIST와 같은 작은 데이터셋을 대상으로 실험을 진행하였으며 Classification task를 위하여 연구되었다. 본 연구에서는 Unlabeled data에 대한 Pseudo-label을 설정하기 위하여 가장 높게 예측한 클래스를 이용한다. Pseudo-label을 생성하기 위한 수식은 (1)과 같다.

$$y_i' = \begin{cases} 1 \; if \; i = argmax_{i'} f_{i'}(x) \\ 0 \; otherwise \end{cases} \quad (1)$$

$$L = L_L + a(t)L_U \quad (2)$$

$$a(t) = \begin{cases} 0 & t < T_1 \\ \dfrac{t - T_1}{T_2 - T_1} a_f & T_1 \le t < T_2 \\ a_f & T_2 \le t \end{cases} \quad (3)$$

y_i'는 Pseudo-label을 의미하며 $f_{i'}(x)$는 x가 모델의 입력일때 i' 클래스에 속할 확률이다. 모델의 목적함수는 Unlabeled data에 대한 학습오차를 L_U, Labeled data에 대한 학습오차를 L_L로 할 때 이 둘을 가중치 합하여 구성하며 수식(2)와 같다. 이때 a(t)값이 너무 크면 Unlabeled data의 영향으로 학습이 방해를 받게 되고, a(t)값이 너무 적으면 지도학습 방법론과 다름이 없게 된다. 따라서 해당 논문에서는 진행 속도 t에 따라 a(t)가 증가하도록 하였으며 수식(3)과 같다. 이때 a_f, T_1, T_2는 상수이다.

2. MixMatch[1]

MixMatch[1]는 2019년에 소개된 연구로 CIFAR-10, STL-10 데이터셋에서 당시 state-of-the-art 성능을 보였다. 본 연구에서는 Unlabeled data에 대한 Pseudo-label을 단순히 예측 확률의 최대값으로 사용했던 [5]와 다르게 추가적인 과정을 거쳐 더욱 정확한 Pseudo-label을 생성하였으며 그 과정은 그림1과 같다. MixMatch에서 하나의 Unlabeled data에 대하여 k번 weak augmentation을 진행하여 augmentations을 생성한 후 augmentations에 대한 모델 예측의 평균에 SoftMax Temperature 방식의

$$Sharpen(p,T) = P_i^{\frac{1}{T}} \Big/ \sum_{j=1}^{L} P_i^{\frac{1}{T}} \qquad (4)$$

그림 1. MixMatch의 Pseudo-label 생성 과정

Entropy Minimization을 적용하여 보정한 분포 값을 Pseudo-label로 사용한다. Entropy Minimization 과정은 Sharpen이라는 함수로 정의되며 수식 (4)와 같다. 입력된 확률 분포 p를 Sharpening하는 과정이며 temperature변수 T가 0에 가까워질수록 Entropy가 작아져, one-hot 벡터의 형태로 sharpening 된다. MixMatch에서는 위와 같은 방식으로 생성된 Pseudo-labeled data와 Labeled data를 mixup[6] 방식으로 결합하여 학습에 사용하였다.

3. ReMixMatch[2]

ReMixMatch[2]는 MixMatch의 후속 논문으로 2019년에 소개되었으며 CIFAR-10, STL-10, SVHN 데이터 셋에서 당시 state-of-the-art 성능을 보였다. ReMixMatch의 전체적인 프로세스는 [1]과 같으나 Pseudo-label 생성 방식에서 두가지 차이가 있다. Unlabeled data에 대한 예측 확률 분포(p)를 보정하는 과정에서 [1]의 경우 Entropy Minimization을 통해 진행하였으나, [2]의 경우 labeled data의 분포를 반영하여 p를 보정하였으며 이를 통해 Labeled data와 유사한 분포를 갖도록 Distribution alignment를 진행하였다.

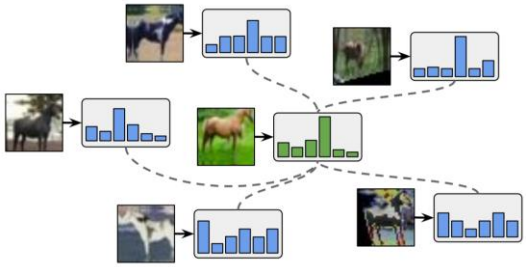

그림 2 CTAugment 예시로 가운데 말 이미지에 strong augmentation인 CTAugment를 적용한 예시는 주변의 5개의 이미지와 같다.

다음으로 data augmentation 과정에서도 차이가 있다. [1]에서는 Flip, Crop과 같은 단순한 변형을 적용하는 weak augmentation만을 사용하였다면, [2]에서는 입력 이미지의 형태를 매우 변형시키는 strong augmentation을 학습에 같이 사용한다. 논문에서 strong augmentation을 위해 제안한 CTAugment의 적용 예시는 그림1과 같다. 이때 strong augmentation가 적용된 데이터의 모델 예측 값은 비교적 노이즈가 많기 때문에 해당 데이터에 대한 Pseudo-label은 augmentation을 적용하기 전 이미지의 Pseudo-label를 사용한다.

4. FixMatch[3]

FixMatch[3]는 2020년에 소개된 연구로 CIFAR-10, CIFAR-100, SVHN, STL-10에서 state-of-the-art의 성능을 보였다. 해당 방법론은 Unlabeled data에 대한 Pseudo-label을 생성하기 위하여 Unlabeled data의 weak augmentations을 모델의 입력으로 하며, 예측 분포 값을 one-hot label로 변형한다. 이는 [5]와 유사하지만 [3]에서는 모든 예측을 Pseudo-label로 사용하지 않고 클래스에 대한 예측 분포 값 중 일정 threshold 이상인 확률이 존재하는 경우에만 Pseudo-labeled data로 Labeled data에 추가한다.

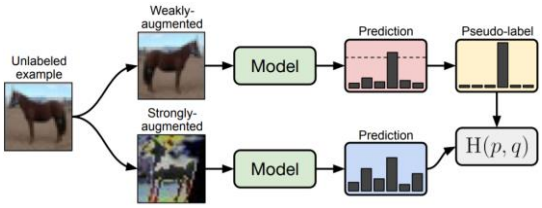

그림 3. FixMatch에서 Unlabeled data의 학습 프로세스

또한 FixMatch는 [2]에서 strong augmentations에 대한 Pseudo-label을 원본 이미지에 대한 값으로 대체한 것을 목적함수로 구현하여 consistency regularization을 진행하였으며 해당 목적함수가 Self-training을 내재하도록 설계하였다. 해당 목적함수는 그림3의 과정으로 구성되었으며 같은 이미지에 대해 weak augmentation된 데이터에 대한 모델의 예측과 strong augmentation된 데이터에 대한 예측이 같아야 한다는 가정을 갖고 있다. 모델의 학습은 Labeled data에 대한 지도학습과 앞서 설계한 consistency regularization을 위한 목적함수를 가중치 합하여 동시에 진행한다.

IV. 데이터 셋 구축 비용을 줄이기 위한 다양한 연구

1. Active Learning 소개

Active Learning 이란 모델이 학습이 필요한 데이터를 쿼리로 보내어 레이블링을 요청하는 방법론으로 모든 Unlabeled data를 레이블링하는 것이 아닌 요청받은 쿼리에 대해서만 레이블링을 진행함으로써 Labeled data로의 가공 비용을 줄이는 데에 목적이 있다. 어떠한 데이터에 대해 레이블링 요청 여부를 결정하기 위해 모델은 데이터의 가치판단을 진행하며 데이터의 가치판단 방식은 Labeled data들의 분포를 고려하는 분포기반 방법론과 모델의 예측 분포를 통해 학습 정도를 고려하는 불확실성(Uncertainty)기반 방법론이 있다.

분포기반 방법론으로는 Core-set 기술을 Active Learning에 적용한 [7] 연구가 있다. Unlabeled data를 subset 단위로 평가하며, subset의 데이터 분포를 모두 반영할 수 있는 클러스터를 구성하고, 클러스터의 중심값을 쿼리로 요청한다.

다음으로 대표적인 불확실성 기반 방법론으로는 [8]이 있다. [8]은 데이터에 대한 모델의 학습 정도를 학습 오차를 통해 측정하였으며 학습 오차가 높을수록 모델이 어려워하는 데이터로 모델이 쿼리로 요청할 가능성이 높아진다.

V. 결론

본고에서는 Self-training 연구동향과 준지도학습법에 대해 살펴보았다. 또한 해당 연구의 동기인 데이터 셋 구축 비용 문제를 해결하기 위한 다른 관점의 연구로 Active Learning 방법론을 IV장에서 소개하였다. 또한 Self-training에 속하는 대표적인 연구들에 대해 Pseudo-label 생성하기 위한 가장 간단한 방식인 [5]에서부터 다양한 준지도학습방식을 결합한 [3]까지 소개하였다. 최근 개인 플랫폼의 증가로 Unlabeled data에 대한 수집은 매우 쉬워지는 반면 Task의 다양화로 레이블링의 난이도는 점점 높아지고 있다. 또한 레이블링은 진행하는 사람에 따라 약간의 차이가 있어 검토과정이 필수적이기 때문에 Labeled data로 가공하는 과정에는 큰 사회

적 비용이 요구된다. 이러한 문제를 다루기 위하여 Self-training뿐만 아니라 준지도학습, 비지도학습에 관한 연구와 Active Learning과 같이 레이블링 할 데이터를 효율적으로 선별하는 기술의 개발도 필수적이다. Pseudo-label[5] 방법론에 다양한 연구를 접목한 MixMatch[1], ReMixMatch[2], FixMatch[3]와 같이 본고에서 소개한 방법론과 소개하지 못한 분야의 결합을 통해 Labeled dataset 부족으로 인한 딥러닝 모델의 발전의 병목을 해결할 수 있을 것이다.

참고문헌

[1] Berthelot, David, et al. "Mixmatch: A holistic approach to semi-supervised learning." arXiv preprint arXiv:1905.02249 (2019).

[2] Berthelot, David, et al. "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring." arXiv preprint arXiv:1911.09785 (2019).

[3] Sohn, Kihyuk, et al. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence." arXiv preprint arXiv:2001.07685 (2020).

[4] Jesper E. Van Engelen, and Holger H. Hoos. "A survey on semi-supervised learning." *Machine Learning* 109.2 (2020): 373-440.

[5] Lee, Dong-Hyun. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." *ICML. Workshop on challenges in representation learning.* Vol. 3. No. 2. 2013.

[6] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.

[7] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489, 2017

[8] In So Kweon Donggeun Yoo. Learning loss for active learning. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2019.