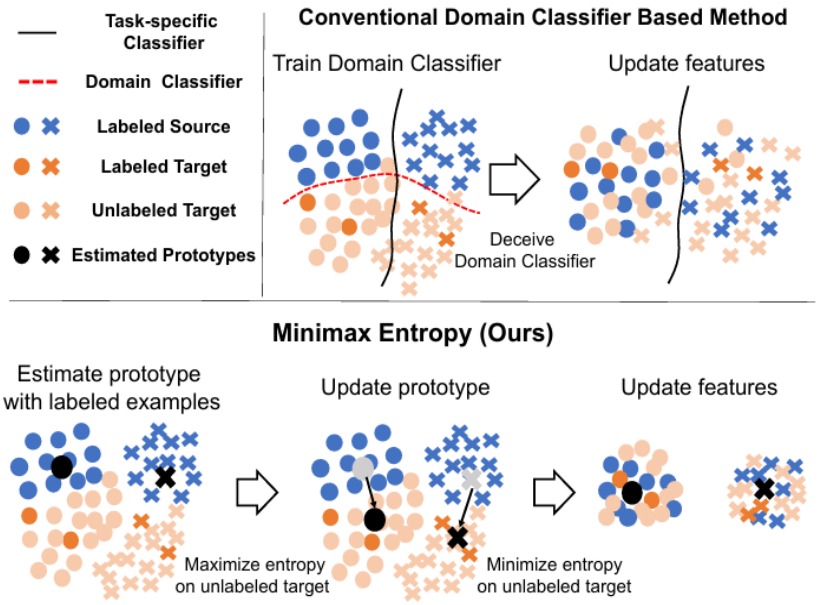
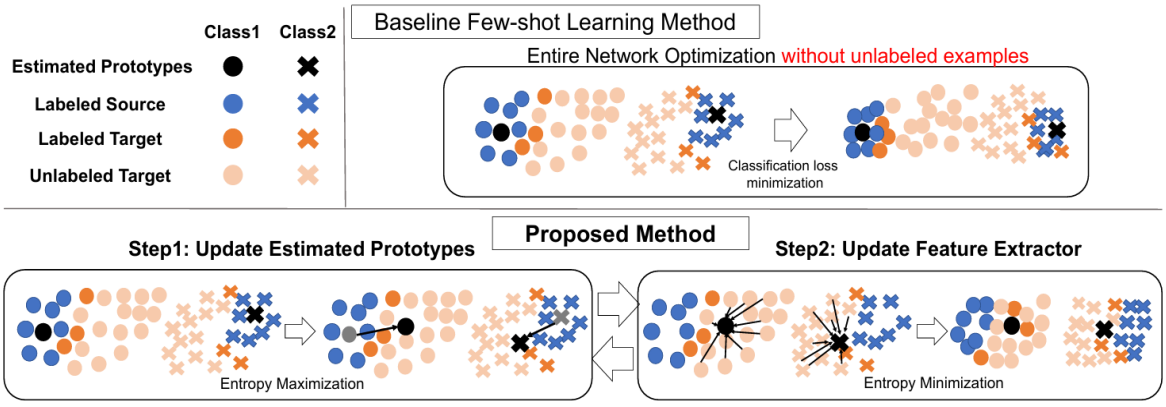
Semi-supervised Domain Adaptation via Minimax Entropy

Background

인공지능 신경망이 발전함에 따라 학습 도메인에 대한 정확도가 매우 높아지고 있다. 하지만, 해당 모델을 새로운 도메인으로 일반화하는데 아직 어려움을 겪고 있다. 따라서 Domain Shift 를 통해 모델의 일반화 성능을 높이는 연구가 진행되어 왔고, UDA(Unsupervised Domain Adaptation)와 같은 방법들은 Source domain 과 Target domain 의 distribution 을 맞춤으로써 문제를 해결하려고 했다. 하지만 이와 같은 방법론은 아래 그림과 같이 class boundaries 를 구별하는데 한계가 존재했다.





이 논문은 Few-shot Learning Method 의 Baseline 인 cosine similarity-based classifier architecture 를 사용한다.

- cosine similarity-based classifier architecture
    - Feature Extractor 를 통해 Image 로부터 feature vector 를 추출한다
    - 각 class 를 대표하는 weight vector 가 존재한다. 학습 초기엔 random 또는 다양한 방식으로 초기화된다.
    - feature vector 와 weight vector 의 cosine similarity 를 비교하여 data 가 각 class 에 속할 확률을 예측한다.
    - 예측된 class 와 실제 label 간 loss 를 계산하고 Back propagation 을 통해 weight vector 가 올바른 방향으로 이동한다.

하지만 이 방식은 그림에서 나타나듯이 Labeled 된 domain 에서는 잘 align 된 것을 볼 수 있는데, Unlabeled 된 data 에 대해서는 성능을 내지 못한다는 단점이 있었다.

따라서 이 논문은 **Unlabeled data 의 Entropy** 를 이용한다.

Contribution

Unlabeled Target Domain 의 Entropy 를 최소화 및 최대화하는 과정을 통해 Alignment 를 유도한다.

Labeled 된 data 로부터 cosine similarity-based classifier architecture 를 이용하게 되면 각 class 를 대표하는 prototype(weight vector)이 나오게 된다. 하지만 Labeled data 의 양이 Source domain 가 절대적으로 많기 때문에 해당 prototypes 은 Source Domain 에 지배된다. 따라서 class prototypes 을 근처의 Unlabeled target sample 과 거리가 최소화 되게 만드는 것이다. 즉 Unlabeled data 에 맞추어서 Entropy 를 최대화하여 class prototype 을 전체 분포의 중앙에 위치 시키고, 이후 다시 Entropy 를 최소화하여 Source domain, Target domain 둘 다 prototype 기점으로 모이게 하는 방식이다.

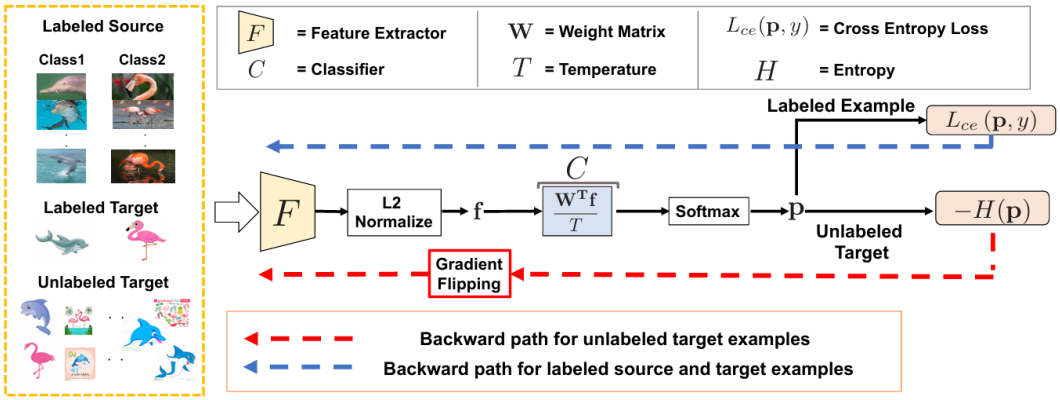


Figure 3: An overview of the model architecture and MME. The inputs to the network are labeled source examples (y=label), a few labeled target examples, and unlabeled target examples. Our model consists of the feature extractor F and the classifier C which has weight vectors ($\mathbf{W}$) and temperature T. $\mathbf{W}$ is trained to maximize entropy on unlabeled target (Step 1 in Fig. 2) whereas F is trained to minimize it (Step 2 in Fig. 2). To achieve the adversarial learning, the sign of gradients for entropy loss on unlabeled target examples is flipped by a gradient reversal layer [11, 37].

Algorithm

Input Data 는 Labeled Source Domain, Labeled Target Domain, Unlabeled Target Domain 을 받는다.

Input data 를 CNN(Feature Extractor)에 거쳐 L2 Normalization 을 적용하면 각 sample 에 해당하는 feature vector 가 나오게 된다.

output feature vector 에 정규화 작업을 거쳐 Classifier(weight matrix 와 similarity)에 넣게 되는데, 둘 사이의 유사도에 temperature, softmax 를 추가로 사용하여 최종 class prediction 이 나오게 된다.

Labeled data 의 경우 supervised learning 으로 cross entropy 를 통해 loss 를 계산한다.

Unlabeled data 의 경우 prototype 의 위치를 수정하는 것으로 Weight Matrix W 와 Unlabeled Target data 의 similarity entropy 를 증가하는 방식을 취한다. 높은 Entropy 를 가지기 위해서, 모든 class 에 대한 균일한 similarity 를 가져야 한다. 이 과정을 거치면, 모델은 Domain invariant prototypes 을 추정하게 된다.

최종적으로, 넓게 퍼져있는 Unlabeled Target data 분포를 maximum 된 prototype 주위로 clustering 해야 한다. 이는 prototype 에 대하여 unlabeled data 로부터 추출된 feature vector 의 entropy 를 감소 시키는 것에 해당하며, Feature extractor 에서 적용될 수 있다.

Conclusion

해당 논문은 Classifier 와 Feature extractor 의 적대적 학습으로 요약할 수 있다.

Classifier 는 Unlabeled target data 의 entropy 를 최대화하도록 훈련하고, Feature Extractor 는 entropy 를 최소화하도록 훈련한다.

uniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, Kate Saenko. Semi-Supervised Domain Adaptation via Minimax Entropy. In *ICCV*, 2019