**Pseudo labeling with Coreset selection**

GSDS 석사재학 20224314

강현구

1. **Proposal**

최근Deep learning(이하 DL)모델은 많은 양의 Labeled Data를 기반으로 다양한 분야에서 높은 성능을 이끌어 냈다. 하지만 모델이 필요로 하는 데이터가 증가함에 따라, Labeling에 필요한 Human resource 또한 비례하여 Labeling Cost가 많이 든다는 문제점이 있다. 이와 관련하여 모델 학습에 효과적인 Unlabeled Data를 우선적으로 선정하여 Labeling 하는 Deep Active learning(이하 DAL)의 중요성이 부각된다. 동일한 맥락에서 주어진 Label Data를 최대한 활용하는 Semi-supervised learning(이하 SSL) 또한 중요성이 강조된다.

DAL의 acquisition Strategy와 SSL의 Data expansion은 상호적으로 긍정적인 영향을 줄 수 있다. DAL의 acquisition Strategy는 학습 모델의 성능에 의존한다. 따라서 Data expansion을 통한 모델 성능 확보는 DAL의 성능 또한 향상시킨다. 한편으로 SSL 학습 과정에서 DAL의 acquisition Strategy은 대표성 또는 정보 가치가 높은 데이터를 우선적으로 Labeling한다. 이는 정확도 높은 Data expansion을 가능하게 한다. 하지만 각각의 방법에 대한 독립적인 연구는 많이 되었으나, 각 방법 간의 연관성에 대한 연구는 부족하다. 위 접근법을 취한 연구로는 Simeoni의 연구(Siméoni, Budnik, Avrithis, & Gravier, 2021)에서 다양한 Batch Sampling 방식과 pseudo labeling을 적용하며 관련성을 확인했다. 하지만 Simeoni의 연구는 각 방법을 Orthogonal 하게 사용하여 DAL과 SSL의 연관성을 활용하지 못했다는 한계점이 있다.

이 연구에서는 DAL의 한 방법인 coreset Selection을 활용한 새로운 Pseudo labelling 방법을 제안한다. Coreset Selection은 주어진 Batch size b개로 전체 데이터를 덮을 수 있는 subgraph들을 구성할 때, 최소한의 반지름을 가지게 만드는 데이터로 Sampling하는 방법이다 클래스가 동일한 데이터 간의 거리가 가까울 것이라는 가정 하에, Coreset selection을 통해 형성한 subgraph에 속한 데이터들은 Sampling되어 Labelling된 중점과 유사한 클래스일 확률이 높을 것이다. 이를 기반으로 기하적 및 확률 관점에서 신뢰성을 보장할 수 있는 unlabeled data에 한해 Pseudo labeling을 부여할 것이다. 본 방법론은 클래스의 수가 적고 데이터 분포가 밀접해 있는 데이터셋에 대해선 높은 효율 및 정확도를 갖춘 Pseudo labeling을 할 수 있을 것이라 기대한다. 또한 Coreset selection의 결과물을 pseudo labeling간 활용한다는 점에서, 별개로 각 방법을 활용하는 것보다 계산량이 적을 것이다.

**1. Task Description**

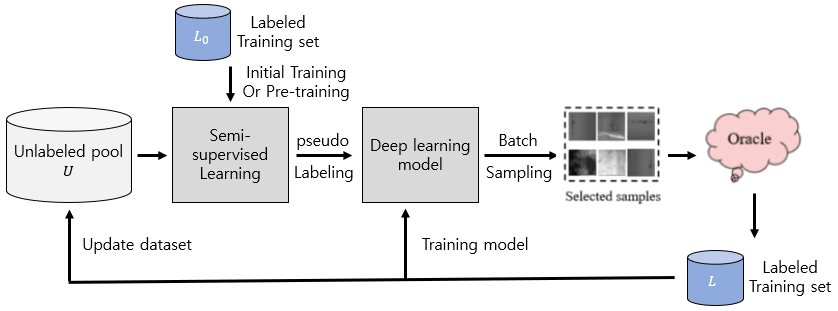
**1-1) Data set : MNIST, SVHN, CIFAR-10, CIFAR-10**

* Simeoni 연구와 비교하기 위해 동일한 Data set으로 선정한다.

**1-2) Notation**

* n examples where := [n]:= {1, …, n}
* Given a set X :=
* labels for training
* U: Unlabeled pool Dataset
* L: Labeled pool Dataset
* of labels , a set if indices with
* Batch

**1-3) Naive Model**

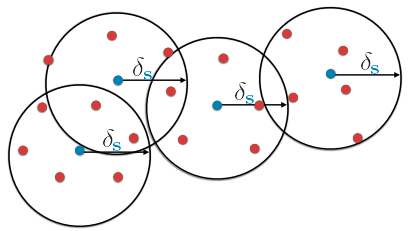


* Phase 1: Random sampling from Unlabeled pool U to make
* Phase 2: Initial Training or Pre-training
* Phase 3: Batch sampling with acquisition strategy of Deep Active learning
* Phase 4: Labelling Sampled data
* Phase 5: Generate Pseudo labeling by SSL

-> Repeat phase 3~ 5.

**2. Idea Sketch**

2-1). Coreset Selection



Coreset Selection은 주어진 Batch size b개로 전체 데이터를 덮을 수 있는 subgraph들을 구성할 때, 최소한의 반지름(threshold)를 가지게 만드는 데이터로 Sampling하는 방법이다{Sener, 2017 #4}. 는 Sampling 된 데이터로 각 i번째 subgraph의 중점이 된다. 는 j번째 data가 i번째 subgraph에 속했는가를 의미한다. 는 subgraph들의 반지름을 의미한다. Coreset은 전체 데이터 셋을 활용했을 때와 Sampling Data만을 활용했을 때의 성능 차이를 최소화하는 방식이다. 또한 각각 Sampling된 subgraph의 중점들은 최소 이상의 간격을 두고 있다. 즉, 각 중점들은 데이터의 밀집 정도에 비교적 무관하게 넓게 퍼져 있다. 이로 인해 Sampling point 는 특정 클래스의 값 만을 포함하는 것이 아닌 전반적인 데이터 셋을 반영한다. 한편으로 의 Labelling 분포는 각 클래스별 데이터의 밀도가 아닌 넓이에 기반한다고 할 수 있다.

**2-2) Pseudo labeling**

 동일한 클래스(=Label)를 가진 데이터들은 거리가 가까울 것이란 가정 하에 2가지 방식의 Pseudo labelling을 고려한다.

첫번째는 i번째 subgraph가 접하는 subgraph들의 중점 Label과 의 Label이 모두 동일할 때, i번째 subgraph에 속하는 점들에 대해서 와 동일한 pseudo label을 부여한다. 거리 가정으로 인해 동일한 클래스를 가진 데이터들은 서로 밀집해 있을 것이다. 따라서 Subgraph 속해있는 데이터들은 중점과 동일한 Label을 가질 확률이 높다. 하지만 주변 subgraph의 중점 Label과 다를 경우, 해당 subgraph는 서로 다른 클래스의 경계에 속해 있을 확률이 크다. 반대로 접하고 있는 subgraph의 중점 Label이 동일할 경우, 두 subgraph의 사이에 있는 데이터는 높은 확률로 동일한 Label을 가질 것이다. 이에 접하고 있는 모든 subgraph 중점의 Label과 동일한 중점의 Label를 가진 subgraph의 데이터에 한해 pseudo labeling을 부여한다. 이 방식의 장점은 라벨 클래스의 수가 적거나 Unlabeled data가 많아 특정 클래스의 분포가 명확히 구분할 때 부각된다. 기하적 관점에서 Subgraph의 영역을 각 클래스별로 높은 가능성의 영역으로 봄으로써 각 클래스의 주변부와 중심부를 구분하며, 중심부로 판단되는 subgraph에 대해서 pseudo labeling을 적용한다. 즉, 적절한 , 클래스별 데이터 밀집이 보장된다면 각 클래스별로 중심부에 해당하는 모든 데이터를 한번에 Pseudo labeling을 할 수 있다. 단, Confirmation bias 문제를 해소하지 않는다면 본 방식은 효율적이나 부정확한 결과를 낼 것이다. 따라서 이후 기존의 SSL 연구에서 Confirmation bias를 해소할 방법들을 적용할 것이다. 그 외에도 기존에 부여한 pseudo labelled data가 이후 coreset selection 간 다른 클래스를 가진 subgraph에 한 번이라도 속할 경우 pseudo label을 제거한다.

두번째로 각 Coreset selection을 진행할 때 마다 각 데이터가 속한 subgraph들의 중점 클래스 개수와 각 Subgraph의 반지름 을 고려하여 각 클래스에 속할 확률을 계산한다. 해당 방법은 Subgraph의 중점 과 첫번째 방식으로 pseudo label을 부여한 데이터를 제외한 모든 데이터에 대해 적용한다. 각 데이터들이 중점이 특정 클래스인 subgraph에 많이 속할수록 동일한 클래스일 확률이 클 것이다. 예로 들어 한 데이터가 4번의 Coreset selection 간 중점의 클래스가 a인 subgraph에 3번 속하고, b 클래스가 중점인 subgraph에 1번 속했다고 하자. 그럼 해당 데이터는 a에 속할 확률이 가장 크며, 그 다음은 b에 속할 확률이 클 것이다. 또한 한번의 Coreset selection 간 다수의 Subgraph에 속하는 경우도 같이 고려해준다. 또한 데이터들은 각 subgraph의 반지름인 에 반비례하여 중점 와 동일한 클래스일 확률이 높을 것이다. 따라서 각 데이터들을 subgraph에 속하는 횟수와 반지름을 기반으로 softmax를 통해 확률화한다.

K는 가 지금까지 Subgraph에 속한 횟수이다. 이때 데이터의 클래스를 가장 높은 확률 값으로 고정할 경우 Confirmation bias가 발생할 수 있다. 따라서 Confirmation bias를 방지하기 위해 SSL에서 자주 사용하는 Mix up data augmentation{Zhang, 2017 #8}을 적용하여 매번 Stochastic하게 pseudo labeling을 부여한다. 한편으로 충분한 횟수의 subgraph의 중첩이 생기기 전엔 틀린 pseudo label을 부여할 확률이 높다. 따라서 허용 오차율 을 도입하여 특정 class일 확률이 1-보다 클 경우만 pseudo label을 부여한다.

**3. 예상 문제점**

3-1). 적절한 선정

Coreset subgraph의 기반 Pseudo labeling에서 가장 중요한 것은 적절한 subgraph의 반지름 을 선정하는 것이다. Coreset에서 은 Sampling size b(= ||)에 달려 있다. 즉, b의 크기가 작을 땐 Labelling Cost는 줄어들지만 이 커져 유의미한 pseudo labeling을 적용하기 어렵다. 특히 특정 Minor 클래스 데이터가 분포한 영역이 보다 작다면, 단일 subgraph 안에 모두 포함되어 아예 고려를 못하는 경우가 생긴다. 반대로 b의 크기가 크다면 이 줄어들어 전체 데이터를 덮는 subgraph의 밀도가 높아져 높은 정확도로 pseudo labeling을 부여할 수 있다. 따라서 Sampling 개수 b와 pseudo labeling 정확도 사이에 Trade off 관계가 형성된다. 기존의 Coreset 연구에선 b를 임의로 부여한다. 따라서 적정한 을 선정할 가이드라인을 추가로 제안해야 한다.

본 연구에서는 subgraph의 데이터 밀도를 통해 가이드라인을 제안하고자 한다. Pseudo labeling 간 가장 피해야할 문제는 confirmation bias이다. 따라서 Labelling cost를 많이 들여 생기는 문제보다, 가 커서 pseudo label을 잘못 부여하는 것을 피하는 데 중점을 둔다. 데이터셋에서 데이터 밀도가 가장 높은 곳을 기준으로, 1개의 subgraph에 속하는 일정 이상의 데이터의 수(ex- M)보다 적게 포함하도록 하는 을 찾는다. 모든 subgraph에 동일한 반지름 을 적용하기 때문에 모든 subgraph에 속한 데이터 수를 M개 아래로 고정할 수 있다. Subgraph의 밀도는 를 통해서 구한다. 은 j번째 data가 i번째 subgraph에 속했는가를 의미한다. 따라서 i번째 row에 대한 값을 다 더하면 i번째 subgraph에 속하는 데이터의 개수를 구할 수 있다. 정리하자면 임의로 Coreset selection을 진행하여 밀도가 가장 높은 subgraph를 찾는다. 그 다음 최고 밀도의 subgraph에 속한 데이터에 한하여 subgraph당 속한 데이터가 M개 보다 적게 하는 를 찾는다. 이후 전체 데이터 셋에 공통적으로 을 적용한다.

단, 추가적으로 해결해야할 문제가 있다. 값을 찾았다고 해서 최적의 Sampling 개수 b를 구할 수 있는 것이 아니다. 주어진 b에 대해 을 최적화하는 것은 Coreset에서 제안한 K-center-greedy를 통해 구할 수 있으나, 반대 방향의 최적화할 방안이 현재 없다. 이에 대한 추가 방안을 제안해야 한다.

3-2). Confirmation bias

SSL 방식의 Pseudo labeling의 주요 문제점 중 하나는 Confirmation bias이다. 이전의 Pseudo labeling 연구에서 Mixup data augmentation, soft labeling, reinitiate model parameter을 통해서 일부 해결함을 보였다. 본 연구에서도 나와있는 방법들을 활용하여 confirmation bias를 해결할 수 있는 지 확인하겠다.

3-3). Distance - based approach

본 연구는 Distance-based approach로 해당 방식이 가지고 있는 단점을 물려 받는다. 일례로 Distance 계산이 많아지는 고차원의 큰 데이터셋 일 경우 계산양이 크게 늘어난다. 이에 Dimension reduction을 통한 계산량 축소를 위해 representation learning을 적용한다. 본 연구에선 기본적으로 CNN을 적용할 것이며, 또한 현장에서 많은 성과를 내고 있는 deep learning 기반의 Conditional autoencoder를 고려한다. Conditional autoencoder은 주어진 label data를 고려하면서 효과적으로 data의 representation을 학습할 수 있다. 한가지 유의해야 하는 점으로 차원 축소를 했을 때 클래스 데이터 별 거리 관계가 유지되는지 확인해야 한다. Hidden variable 들은 decode, encode에 최적화되어 있어 클래스별 거리 관계를 고려하지 않았을 수 있다. 다른 한편으론 클래스별 특징을 잘 반영했기 때문에 거리 관계 또한 재현되어 있을 수 있다.

또한 차원 축소 뿐만 아니라 각 클래스별로 데이터를 더욱 밀집시키는 특징을 가진 t-SNE를 고려한다. 하지만 계산량이 O()으로 큰 데이터셋에 적용하기 어렵다는 제한이 있다. 실제 구현에는 Barnes hut t-SNE 라는 방법을 통해 더 효율적으로 계산할 수 있어 도입 유무에 따른 효과를 확인하겠다.

**4. 추가 고려사항**

여러 Coreset selection을 통해 생성한 subgraph들은 동일한 node type을 가지면서 다수의 edge type을 가지는 Multiplex network 구조를 띈다. 즉, graph node embedding 방법의 적용 가능성이 있다.

본 방식은 데이터의 밀도를 고려하지 못한다는 선천적인 단점이 있다. 하지만 각 subgraph에 속한 데이터의 수로 밀도 측정이 가능하다. 따라서 각 subgraph의 밀도를 활용한다면 좀 더 정확한 pseudo labeling 부여가 가능할 것으로 보인다. 하나의 예시로 밀도가 일정 이상이며, 기하적인 관점에서 중심부일 때에만 pseudo labeling을 부여하는 등 제한 조건으로 활용할 수 있다.

**5. 예상 기대효과**

본 방식은 Coreset의 Subgraph를 기하적인 관점에서 활용하여 각 클래스의 주변부와 중심부를 구분한다. 그리고 클래스의 중심부에 대해서 한번에 pseudo labeling을 부여한다. 즉, unlabeled 데이터가 많으면서 클래스의 수가 적은 데이터셋과 같이 클래스 별로 밀도가 높은 데이터에 대해 높은 효율의 pseudo labeling이 가능할 것이다. 동시에 각 데이터 별로 속했던 subgraph 중점의 클래스와 반지름 을 같이 고려하며 Mix up augmentation을 통해 soft labeling을 한다. 이로써 confirmation bias를 해소할 것으로 판단된다.

본 방식은 DAL Coreset의 최소 데이터의 수로 전체 데이터셋을 대표하는 특성을 적극 활용한다. 전체 데이터를 고려하도록 subgraph을 형성하기에 특정 클래스 데이터에 한정되거나, 일부 데이터가 다수의 데이터로부터 떨어져 있는 경우도 함께 고려해줄 수 있다. 또한 representation learning 과의 연계 활용도가 높다. 어떠한 형태의 데이터라도 representation learning을 통해서 distance를 측정할 수 있는 Structured data 형식으로 만들 수 있다면 항상 적용 가능하다. 또한 representation learning을 통한 차원 축소는 Coreset selection 간 계산양을 상당수 줄이며, 클래스별로 밀집 정도를 심화한다.

**<참고문헌>**

Liu, P., Zhang, H., & Eom, K. B. (2016). Active deep learning for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(2), 712-724.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., . . . Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR), 54*(9), 1-40.

Sener, O., & Savarese, S. (2017). Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Siméoni, O., Budnik, M., Avrithis, Y., & Gravier, G. (2021). *Rethinking deep active learning: Using unlabeled data at model training.* Paper presented at the 2020 25th International Conference on Pattern Recognition (ICPR).

Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2016). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology, 27*(12), 2591-2600.

Zhu, J.-J., & Bento, J. (2017). Generative adversarial active learning. *arXiv preprint arXiv:1702.07956*.