**Pseudo labeling with Coreset selection**

GSDS 석사과정 20224314

강현구

1. **Abstract**

Label Data를 최대한 활용하는 Semi-supervised learning(이하 SSL)은 Labeling Cost 문제를 해결할 수 있는 주요 방안 중 하나이다. 단SSL 모델 또한 모델을 학습하거나, Dataset의 Diversity를 반영하지 못할 정도로 Label Data가 부족할 경우 적용에 제한 사한이 생긴다. SSL의 주요 방안인 Consistency Regularization은 부분적으로 해결하며, Pseudo labeling은 클래스별 Prior 확률을 활용하는 등 비현실적인 방안을 활용한다. 따라서 Label 데이터를 선정할 때부터 Dataset의 Diversity를 고려하며, 사전 지식 및 SSL 모델의 성능에 의존하지 않고 Label Data 문제를 해결할 필요가 있다.

이 연구에서는 Active learning의 coreset Selection을 활용한 새로운 Pseudo labelling 방법을 제안한다. Coreset Selection을 통해 Label Data의 Diversity를 보장한다. 동시에 Coreset Selection 간 형성한 subgraph들은 기하적 관점에서 모델의 성능에 의존하지 않고 높은 신뢰도의 Pseudo labeling을 해줘 Label Data 부족 문제를 해결해줄 것이다. 본 방법론은 클래스의 수가 적고 데이터 분포가 밀접해 있는 데이터셋에 대해선 높은 효율 및 정확도를 갖출 것으로 기대한다. 또한 Consistency 방법과 Represent learning과 병행할 수 있어 잠재성을 갖췄다 할 수 있다.

**1. Introduction**

**1. 연구주제의 배경 : 어떤 기술이며, 어떤 장점이 있는지, 어떠한 분야에 이용될 수 있는지**

- > 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 Label Data를 최대한 활용하는 Semi-supervised learning(이하 SSL)의 중요성이 강조된다. SSL은 주어진 Label Data를 활용하여 모델을 학습한다. 이후 학습된 모델을 통해 Unlabel Data을 활용한다. 그 결과 SSL은 적은 비중의 Label Data로 전체 Dataset의 성능과 유사하게 낼 수 있다. SSL은 Labeling Cost 로 인해 모든 Data에 대해 Labeling이 제한되는 분야에 폭 넓게 사용되고 있다.

[어느 분야에서 사용되고 있는지 추가할 것].

**2. 연구주제/기술에서 현재 관심사가 되고 있는 연구 분야가 무엇인지? 핵심 이슈/문제점을 기술한다.**

- SSL 은 모든 Dataset이 Labeling 된 SL 방법과 달리 오류에 취약하다. Labeled 데이터의 규모가 커 Dataset을 잘 반영하고 온전히 학습할 수 있다면 비교적 괜찮다. 하지만 Label 데이터가 적어 Dataset의 Diversity을 반영하지 못하며, 또한 모델을 학습하기에 충분하지 않다면 Confirmation bias는 확대된다. SSL이 Label Data를 통해 학습한 모델을 Leverage한다는 점을 고려할 때, Confirmation bias은 모델 전체의 성능을 크게 악화시킬 수 있다.

**3. 이 기술적 이슈를 해결하기 위해 기존의 연구에서는 어떤 연구를 하고, 어떤 결과를 얻었는지? 문헌조사를 통해 체계적으로 분석하여 소개한다.**

- SSL은 크게 Consistency 방법과 Pseudo Labeling 방법 2가지로 나눠진다. Consistency 방법은 모델의 Input에 변형이 가해져도 모델의 결과는 일정하게 유지돼야 한다는 가정 하에 이뤄진다. 이에 Label Data에 변형을 가한 데이터를 생산(augmentation) 함으로써 동일한 Class의 Label Data를 확장한다. 이로써 Confirmation bias를 피할 수 있다.

[문헌조사로 추가 필요 ]

한편 Pseudo labeling은 학습된 모델을 통해 신뢰도가 높게 나온 Unlabel Data에 대해 Pseudo labeling을 부여하는 방법이다. Pseudo labeling 한 Unlabel data를 모델 학습에 활용하여 보다 높은 모델 성능을 확보한다. 단, Confirmation bias가 주요 문제점이다. 이를 해소하기 위해 모델 학습에만 한정적으로 활용하거나, Threshold를 설정, 또는 규제항을 도입한다.

**4. 기존 연구의 한계 즉, 아직 해결하지 못한 이슈/문제는 무엇인지? "연구의 동기"를 제시한다.**

- Consistency 방법은 Label data 부족 문제는 해소하나, 주어진 Label Data가 Dataset의 Diversity를 반영하지 않을 경우 그 외 Class에 대해선 오차를 가질 수 있다. 이를 해소하기 위해선 Label Data을 선별하는 과정부터 Diversity를 고려해줄 필요가 있다. 한편 Pseudo labeling은 Unlabel에 대한 판단을 모델에 의존한다는 한계점을 가진다. 즉, Label Data이 부족하여 모델 자체의 성능을 보장할 수 없을 때 Confirmation bias는 필연적이다. 초기 모델 오류를 방지하고자 각 클래스별 비율에 대한 Prior을 규제항으로 도입하지만, Class에 대한 사전 확률을 안다는 것은 비현실적이다.

**5. 이 논문에서 해결하고자 하는 연구 목표를 제시한다.**

- 본 연구에서는 Label Data의 Diversity를 미반영 및 SSL 모델 학습 간 Label Data부족 문제를 해결할 것이다. 이를 위해 Active learing – Coreset selection 방법을 도입한다. Active learning은 최소한의 Data를 Sampling하여 전체 데이터셋을 활용했을 때와의 성능을 유사하게 만드는 방법이다. 이때 Acquisition method에 따라 Sampling Data가 Dataset의 어떤 정보를 반영하도록 할 수 있다. 본 연구에서는 Sampling 된 Label Data가 데이터 셋의 Diversity를 반영하면서 동시에 Coreset의 subgraph에 속한 Unlabel Data의 대표성을 띌 수 있도록 한다. 이후 기하적 및 확률 관점에서 Unlabel 데이터에 대해 모델의 성능에 의존하지 않고 신뢰성을 보장한 Pseudo labeling을 부여한다.

**6. 예측되는 주요한 연구 결과를 기술한다.**

본 연구는 Label Data의 diversity 문제와 Pseudo labeling의 Label Data 부족으로 인한 Confirmation bias 문제를 해소할 수 있다. 또한 Subgraph를 활용한 기하적 및 확률 관점에서의 Classification은 클래스의 수가 적고 데이터 분포가 밀접해 있는 데이터셋에 대해선 높은 효율 및 정확도를 갖출 것으로 기대한다. 이는 본 방법이 Representation learning과 Consistency 방법과 병행할 수 있다는 점에서 적은 학습양으로 높은 신뢰도의 Classification을 가능하게 만들 것이다.

**7. 과학적인 창의성과 독창성 있는 주요 결론을 제시하고, 이전 연구와의 차별성 및 중요성을 제시한다.**

Active learning과 SSL의 연계에 대한 연구는 드물다. 일부 있는 연구 또한 Active learing과 SSL을 병렬적으로 활용한다. AL은 모델 학습에 가장 도움될 것으로 여겨지는 Data를 Sampling 하는 방법이며, SSL은 주어진 Label Data를 Leverage 하는 방법이다. 즉, 두 방법을 잘 연계한다면 추가적인 시너지를 낼 것이라 판단된다. 본 측면에서 AL이 제공하는 Data의 특성을 적극적으로 활용한 본 SSL 방법론은 좋은 시작점이 될 것으로 생각한다.

**2. Related works**

2-1). Suvervised learning [점검 필요]

SSL은 Labeling Cost 문제로 일부의 Label Data만을 구할 수 있는 현장에서 널리 활용되는 방법이다. SSL은 크게 3가지 가정을 기반으로 한다. 첫째, Smoothness 가정으로, 확률 밀도가 높은 지역의 입력값 x1, x2가 가깝다면 각각의 연관된 Label y1, y2도 가까울 것이다. 둘, Low Density 가정으로, 모델의 결정 경계가 데이터의 확률 밀도가 높은 곳을 지나지 않아야 한다. 마지막으로 고차원의 입력 데이터가 저차원의 Manifold를 따라 놓여 있다. 즉, SSL은 Data 간의 거리가 Data의 유사성을 반영하며, 각 Class 별로 밀도를 통해 경계 지점을 파악할 수 있을 때 활용할 수 있다. 그 결과 Distance-based 방법들이 손 쉽게 적용될 수 있어 다양한 영역에 폭 넓게 활용되고 있다.

SSL의 주요 방법 중 하나는 Consistency regularization 방법이다. 모델의 Input에 변형이 가해져도 모델의 결과는 일정하게 유지돼야 한다는 가정 하에 적용된다. Confirmation bias를 효과적으로 제거하여 Pseudo labeling 보다 적극적으로 활용되고 있다. Label Data을Permutation하거나 Noise를 적용한 경우에도 Class는 변하지 않는 다는 점을 활용하여 Label Data를 Augment 한다. 한편 Unlabeled Data에 대해서도 정확한 Class를 몰라도 Data augmentation한 것과의 예측은 동일해야 함을 활용한다.

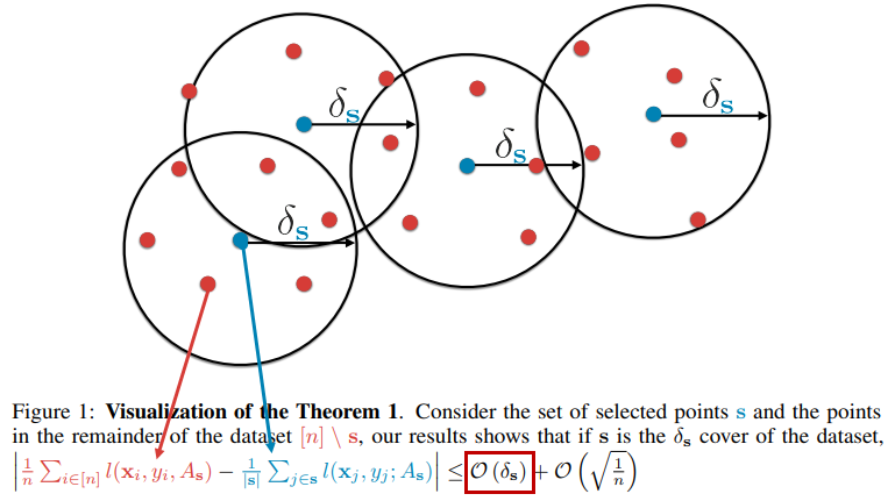
또 다른 SSL의 방식은 Pseudo labeling이며, 본 연구에서 주로 다룬다. Pseudo labeling은 신뢰도가 높은 값에 대해 Pseudo labeling 한다면 전체 데이터셋이 Labeling 되었을 때와 결과가 크게 다르지 않을 것을 가정한다. 단, 초기의 오류가 확대 재 생산되는 Confirmation bias가 발생할 수 있다. 이 문제를 해결하기 위해 Pseudo labeling을 모델 Fine Tuning에만 한정적으로 사용한 후 제거하거나, Threshold를 설정하여 일정 이상의 신뢰도를 가진 값에만 Pseudo labeling을 부여하는 방안을 쓰고 있다. Pseudo labeling의 문제는 Label Data가 모델을 학습하기에도 부족할 때 추가된다. 바로 초기 학습 모델의 오류로 인해 Pseudo labeling 된 Data들의 클래스 비율이 실제 DataSet 에서의 클래스의 비율에 수렴하지 않고 발산할 수 있다는 점이다. 이를 막기 위해 Dataset의 클래스의 사전 확률을 규제항으로 도입한다.

일부 몇 논문에서는 Pseudo labeling에서 Confirmation bias를 방지하였을 경우 Consistency regularization 만큼의 성능을 확보한 연구가 있다. 따라서 기존의 연구에서 confirmation bias를 피할 수 있도록 방안을 탐색할 것이다. 또한 기존의 Pseudo labeling 연구 간 Confirmation bias을 방지하고자 사용한 Mixup data augmentation, soft labeling, reinitiate model parameter을 활용하여 confirmation bias를 해결할 수 있는 지 확인하겠다.

2-2) Active learning [점검 필요 ]

Active learning 또한 Sampling Cost로 인해 모든 데이터를 Labeling 할 수 없을 때 사용하는 방법이다. 적은 수의 Labeling으로도 전체 데이터셋이 Labeling 되었을 때와 유사한 성능을 내는 정보 가치가 높은 Unlabel Data를 선별한다. AL의 유형은 크게 3가지가 있다. Membership query Synthesis은 모델이 Labeling을 요청할 데이터를 생성한다. Stream-based Selective sampling은 새로운 샘플에 대해서 레이블이 필요한지를 판단한다. 마지막으로 Pool-based sampling은 주어진 데이터 셋에서 중요한 데이터를 선별한다. 대부분의 AL 연구는 Pool-based sampling 연구에 치중되어 있으며, 본 연구에서도 Pool-based sampling에 초점을 둔다.

Pool-based sampling AL에서 Acquisition strategy에 따라 Unlabel Data를 선별한다. 크게 Uncertainty-based approach, Expected-based approach, Diversity-based approach가 있다. Uncertainty-based approach는 주로 Bayesian network Model에 의거하여 각 Data별 Uncertainty를 측정한다. 이후 Uncertainty score가 높은 순으로 데이터를 Sampling 하여 Labeling을 한다. Uncertainty 가 높은 데이터들이 각 Class 별 경계 값 주변에 놓여 있는 것을 고려할 때, Uncertainty가 높은 데이터들은 서로 비슷한 상황에 놓여 있는 경우가 많다. 즉, 각 Class 별로 경계 영역을 학습하는 데에는 도움이 되나 Dataset의 Diversity를 반영하지 못한다. Expected-based approach는 데이터를 Sampling 했을 때 전후로 변화한 모델 성능 expectation의 차이를 기준으로 데이터를 선별한다. 단, Uncertainty based approach와 동일하게 모델의 성능에 의거하여 판단한다는 점에서 SSL의 모델의 성능을 신뢰할 수 없는 상황에선 적용하기 어렵다. 마지막으로 Diversity-based approach는 데이터셋의 Diversity를 보장할 수 있는 데이터를 선별하는 방법이다. 대표적으론 Coreset selection이 있다.

[참고용. 나중에는 지우거나, 대체 해야함]

Coreset Selection은 모델 성능 향상에 가장 효과적일 것으로 기대되는 Data를 Sampling 한다. 결과적으로 모델 성능 향상하는 Data를 Sampling 하는 것은 주어진 Sampling size b개로 전체 데이터를 덮을 수 있는 subgraph들을 구성할 때, 최소한의 반지름(threshold)를 가지게 만드는 데이터를 Sampling 하는 것과 동일하다{Sener, 2017 #4}. 모든 데이터를 덮으면서 반지름을 최소로 하는 subgraph를 형성하는 방법은 K-Center greedy 알고리즘과 동일하다. 그 결과 기존에 선정된 Sampling Data로 부터 거리가 가장 먼 Data를 선정한다. 각 Sampling Data들은 Dataset의 밀집 정도에 무관하게 넓게 퍼져 있다. 이로 인해 Sampling point 는 특정 클래스의 값 만을 포함하는 것이 아닌 전반적인 데이터 셋을 반영한다. 또한 Acquisition Strategy가 Data 간의 거리를 활용함으로써 SSL 모델의 성능에 의존하지 않으면서 데이터셋의 Diversity를 반영하는 Data들을 Sampling 할 수 있다. SSL 모델의 성능에 의존하지 않기에 SSL 모델의 초기 Confirmation bias 없이 활용할 수 있다.

**4. Method**

**4-1) Pseudo labeling**

일반적으로 SSL 모델의 Classification 신뢰도가 높은 Unlabel Data에 대해 Pseudo labeling을 부여한다. Label Data 부족으로 인해 SSL 모델을 활용할 수 없다. 따라서 본 연구에서는 Coreset selection 간 형성된 subgraph를 활용하여 신뢰도 높은 Classification을 진행한다. 앞서 는 Sampling 된 데이터로 각 i번째 subgraph의 중점으로 정의한다. 는 subgraph들의 반지름으로 정의하겠다. 각 값들은 Coreset selection 간 반지름 을 최소화하는 과정에서 모두 구할 수 있다. SSL의 Smoothness 가정에 의거하여, 확률밀도가 높은 i번째 subgraph G\_i에 속한 Data x\_j s.t. j \in O\_i(u\_i) 들은 중점인 와 클래스가 동일할 확률이 높을 것이다. 확률 밀도는 각 subgraph에 포함된 Data의 개수로 측정할 수 있다. 즉, Label Data이기도 한 은 확률 밀도가 높을 때 subgraph에 속한 unlabel Data를 대표할 수 있다.

더 나아가 주어진 Sampling Size를 Active selection을 통해 형성한 subgraph들이 Dataset을 촘촘하게 덮을 수 있을 정도로 충분히 작은 반지름 을 가졌다고 가정한다. G\_j 를 G\_i 와 접하고 있는 subgraph라고 하자. 이때 각 subgraph의 중점 u\_i, u\_j 은 Labeling 되어 클래스를 알고 있다. 이때 u\_i 와 u\_j 의 Class가 동일할 경우 해당 subgraph들은 특정 클래스 데이터 분포의 중심부에 위치하며, 반대로 다를 경우 주변부에 위치할 것이라고 기하적 관점에서 유추해볼 수 있다. 즉, subgraph 간의 기하적 관계를 통해 각 subgraph들이 클래스별 데이터 분포간 어디에 위치해 있는지 유추할 수 있다. 정리하면 반지름 가 충분히 작아 Dataset을 잘게 나누고 있을 때, 확률 밀도가 높고, 접하고 있는 subgraph의 중점들의 Class와 u\_i의 Class가 동일하다면 G\_i가 특정 클래스별 데이터 분포간 중심부에 속할 것이라 유추할 수 있다. 위의 조건을 모두 충족하는 subgraph에 속하는 x\_j에 대해서 일괄적으로 와 동일한 pseudo label을 부여한다. 이 방식의 장점은 라벨 클래스의 수가 적거나 Unlabeled data가 많아 특정 클래스의 분포가 명확히 구분할 때 부각된다. 충분히 작은 반지름 , 클래스별 데이터 밀집이 보장된다면 각 클래스별로 중심부에 해당하는 모든 데이터를 한번에 Pseudo labeling을 할 수 있다. Confirmation bias를 피하기 위해 Pseudo labeling Data들은 모델의 Fine-tuning에만 활용한다. 이로써 Dataset의 Diversity를 반영하는 Label Data를 선정하고, subgraph 간의 기하적 관계를 활용하여 Label Data 부족 문제를 해소할 수 있다.

Coreset selection을 다회차 진행하여 sugraph가 중첩되었을 때 또 다른 방식으로 Pseudo labeling이 가능하다. 계산상 효율을 높이기 위해 Active selection의 대상은 앞서 Pseudo labeling되지 않은 Unlabel Data에 한해 진행한다. 매번 Coreset selection 간 생성된 subgraph들은 모든 Dataset를 cover 해야 하기 때문에 모든 Unlabeled Data들은 다수의 subgraph안에 속하게 된다. 각 Unlabel Data들이 특정 Class 중점을 가진 subgraph에 많이 속할수록 동일한 클래스일 확률이 높아질 것이다. 또한 각 subgraph들의 반지름 \_k 에 반비례하여 확률이 높아질 것이다. 즉, 따라서 각 데이터들을 subgraph에 속하는 횟수와 반지름을 기반으로 softmax를 통해 확률화할 수 있다.

[표기 손 볼 필요 있음]

K는 가 지금까지 Subgraph에 속한 횟수이다. 본 Pseudo labeling 방법은 Label data 및 첫번째 방식으로 pseudo label을 부여한 데이터를 제외한 모든 데이터에 대해 적용한다. 한번의 Coreset selection 간 다수의 Subgraph에 속하는 경우도 같이 고려해준다. 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.

4-2) Choose Proper

Coreset selection 기반의 Pseudo labeling에서 가장 중요한 것은 적절한 subgraph의 반지름 을 선정하는 것이다. Pseudo labeling의 신뢰도는 Subgraph의 반지름 크기에 크게 영향을 받는다. 반지름 의 크기가 충분히 작을수록 클래스 데이터 분포간 주변부, 중심부를 명확히 구분할 수 있다. 반대로 반지름 의 크기가 크다면 subgraph들이 주변부와 겹쳐 조건을 만족하는 경우가 줄 것이다. 더불어 Minor Class 데이터가 분포한 영역이 한 subgraph에 포함되는 등 오차가 커질 것이다.

반지름 의 크기는 Sampling size b(= ||)에 달려 있다. 즉, b의 크기가 작을 땐 Labelling Cost는 줄어들지만 이 커져 유의미한 pseudo labeling을 적용하기 어렵다. 반대로 b의 크기가 크다면 이 줄어들어 전체 데이터를 덮는 subgraph의 밀도가 높아져 높은 정확도로 pseudo labeling을 부여할 수 있다. 따라서 Labeling Cost와 pseudo labeling 정확도 사이에 Trade off 관계가 형성된다. 따라서 Labeling Cost를 고려하면서도 충분히 작은 반지름 의 subgraph를 형성할 수 있도록 적절한 Sampling size을 고를 가이드라인이 필요하다. 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. 한 Subgraph에 속하는 데이터의 수가 많을수록 Confirmation bias 확률이 증가한다. 따라서 한 subgraph의 속하는 데이터의 개수가 일정 이상 넘지 않도록 규제하겠다.

임의의 Sampling size로 Coreset selection을 수행하여 subgraph들을 형성한다. 이후 가장 밀도가 높은 subgraph에 속하는 Data에 한정하여 다시 Coreset selection을 수행한다. 이때 각 subgraph마다 일정 이상의 데이터의 수(ex- M)보다 적게 포함하도록 하는 반지름 을 찾는다. 반지름 은 가장 밀도가 높았던 subgraph에 속하는 데이터의 개수를 M개로 줄였다. 따라서 을 전체 데이터 셋에 적용할 때 모든 subgraph에 속한 데이터 수를 M개 아래로 고정할 수 있다.

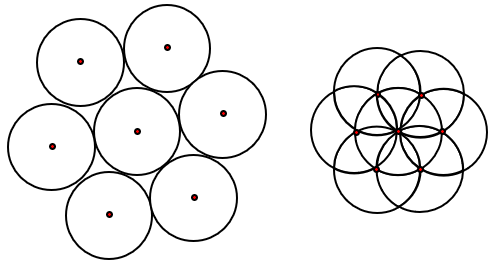
이제 반대로 반지름의 크기를 보다 작게 만드는 Sampling size n’을 찾을 것이다. Subgraph의 넓이를 통해서 Batch size의 Bound를 형성할 수 있다. 하나의 Outlier가 데이터 분포에서 벗어나 있는 경우에 대해 robust을 부여하기 위해서 Cover 하고 있는 subgraph의 넓이를 통해 Dataset S의 넓이를 정의한다. Dataset의 데이터 분포의 넓이를 S라고 할 때 S값은 고정되어 있다. 이때 S는 각 Sampling size n에 따라 n개의 subgraph들로 덮여야 한다. 이때 Subgraph안에

Sampling size n이 고정되어 있을 때 S Subgraph의 분포 형태에 따라 Upper bound와 Lower bound가 형성된다. 이에 앞서 증명해야 할 것이 하나 있다.

Lemma 1. 각 subgraph의 중점들은 최소 이상의 간격을 두고 있다.

Coreset selection은 k-center greedy 알고리즘과 방법이 동일하다. 첫 Sampling Data는 Random하게 Sampling 한다. 그리고 은 Sampled Data x\_1 와 가장 거리가 먼 Unlabeled Data 거리로 정의한다. 이후 의 거리를 가진 Unlabeled Data를 x\_2로 Sampling 한다. 이 과정을 Sampling size b에 도달할 때까지 반복한다. 이로써 이며, for all i,j in [1, b] 이 성립한다. 의 거리를 가지는 점이 2개 이상일 때 등호가 된다. 을 성립한다.

이로써 Subgraph를 넓이의 관점에서 볼 때 모든 Dataset을 Cover 하고 있는 subgraph들의 넓이는 가장 클 때는 겹치지 않았을 때이며, 각 중점 간의 거리가 이며 최대한 겹쳐 있을 때 최소가 된다.



첫째 subgraph의 반지름을 줄이기 위해 적절한 양의 batch size를 선택한다.

Thm.

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.

**1. Task Description**

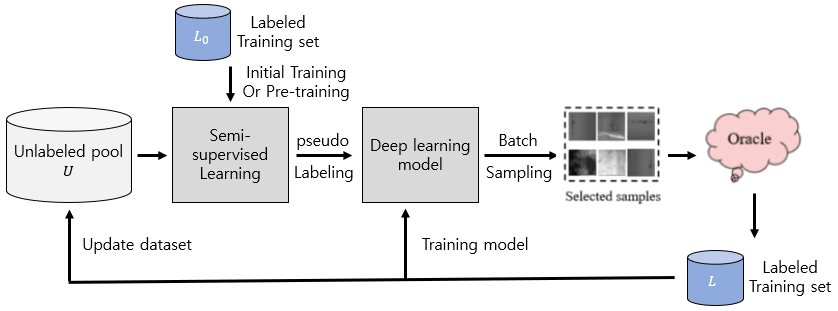
**1-1) Data set : MNIST, SVHN, CIFAR-10, CIFAR-10**

* 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.

**<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*.