**Pseudo labeling with Coreset selection**

GSDS 석사과정 20224314

강현구

1. **Abstract**

Semi-supervised learning(이하 SSL)은 Labeling Cost 문제를 해결할 수 있는 주요 방안 중 하나이다. 단 모델을 학습하거나, Dataset의 Diversity를 반영하지 못할 정도로 Label Data가 부족할 경우 Confirmation bias가 발생할 수 있다. 이를 막기 위해 Pseudo labeling간 클래스별 Prior 확률을 활용하나, 이는 비현실적이다. 따라서 Label 데이터를 선정할 때부터 Dataset의 Diversity를 고려하며, Dataset에 대한 사전 지식 없이 Label Data 문제를 해결할 필요가 있다.

본 연구에서는 Active learning의 coreset Selection을 활용한 새로운 Pseudo labelling 방법을 제안한다. Coreset Selection을 통해 Label Data의 Diversity를 보장한다. 또한 Coreset Selection 간 형성한 subgraph들은 기하적 관점에서 모델의 성능에 의존하지 않고 높은 신뢰도의 Classification을 가능하게 한다. 즉, 모델에 의거하지 않은 Classification으로 Label Data 부족 문제를 해결한다. 본 방법론은 클래스의 수가 적고 데이터 분포가 밀접해 있는 데이터셋에 대해선 높은 효율 및 정확도를 갖출 것이다. 또한 Consistency 방법과 Represent learning과 병행할 수 있어 적용 범위가 넓다.

**1. Introduction**

최근Deep learning(이하 DL)모델은 많은 양의 Labeled Data를 기반으로 다양한 분야에서 성과를 내고 있다. 하지만 모델이 필요로 하는 데이터가 증가함에 따라, Labeling Cost를 어떻게 해결할 것인가 중요한 화두가 되었다. 이와 관련하여 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 supervised learning이 제한되는 분야에 폭 넓게 사용되고 있다.

SSL 은 SL 방법과 달리 오류에 취약하다. Labeled 데이터의 규모가 커 Dataset을 잘 반영하며 학습 모델의 성능을 이끌어 낼 수 있다면 괜찮다. 하지만 Label 데이터가 적어 Dataset의 Diversity을 반영하지 못하며, 또한 모델을 학습하기에 충분하지 않다면 Confirmation bias가 발생한다. 초기에 발생한 Confirmation bias을 해결하지 못한다면 모델 전체의 성능을 크게 악화시킬 수 있다.

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. 한편 Pseudo labeling은 학습된 모델을 통해 신뢰도가 높게 나온 Unlabel Data에 대해 Pseudo labeling을 부여한다. Pseudo labeling 한 Unlabel data를 모델 학습에 활용하여 보다 높은 모델 성능을 보장한다. 단, Confirmation bias을 해소하지 못했을 경우 오히려 성능이 악화된다. 이를 해소하기 위해 모델 Fine tuning 과정에만 한정적으로 활용하거나, 신뢰도에 대해Threshold를 설정, 또는 규제항을 도입한다.

Consistency 방법은 Label data 부족 문제는 해소하나, 주어진 Label Data가 Dataset의 Diversity를 반영하지 않을 경우 그 외 Class Data에 대해 오차를 가질 수 있다. 이를 해소하기 위해선 Label Data을 선별하는 과정부터 Diversity를 고려해야 한다. 한편 Pseudo labeling은 Unlabel에 대한 판단을 학습 모델 성능에 의존한다. 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, Dataset에 대한 사전 확률을 안다는 것은 비현실적이다.

본 연구에서는 Dataset의 Diversity 미반영 및 SSL 모델 학습 간 Label Data부족 문제를 해결할 것이다. 이를 위해 Active learning – Coreset selection 방법을 도입한다. Active learning은 최소한의 Data를 Sampling하여 전체 데이터셋을 활용했을 때와의 성능을 유사하게 만든다. 이때 Acquisition Strategy에 따라 Sampling Data가 Dataset의 특정 정보를 반영시킬 수 있다. 본 연구에서는 Sampling 된 Label Data가 데이터 셋의 Diversity를 반영할 것이다. 이후 기하적 관점에서 각 Label Data는 Unlabel Data에 대해 대표성을 띄고 있음을 증명할 것이다. 이후 Coreset Selection 간 생성한 subgraph의 기하적 관점에서 모델의 성능에 의존하지 않고 높은 신뢰도의 Classification이 가능함을 보일 것이다. Classification을 통해 추가적인 Label Data를 확보함으로써 Label Data 부족 문제를 해결한다.

본 연구는 Dataset diversity 미반영 및 Label Data 부족으로 인한 Confirmation bias 문제를 해소할 수 있다. 또한 Subgraph를 활용한 기하적 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 적용 범위가 넓을 것이라 판단한다. .

본 연구는 Active learning과 SSL의 연계 연구라는 점에서 차별성이 있다. 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

SSL은 Labeling Cost 문제로 일부의 Label Data만을 구할 수 있는 현장에서 널리 활용된다. SSL은 크게 3가지 가정을 기반으로 한다. 첫째, 확률 밀도가 높은 지역의 입력값 x1, x2가 가깝다면 각각의 연관된 Label y1, y2도 가깝다. 둘, 모델의 결정 경계가 데이터의 확률 밀도가 높은 곳을 지나지 않는다. 마지막으로 고차원의 입력 데이터가 저차원의 Manifold를 따라 놓여 있다. 즉, SSL은 Data 간의 거리가 Data 특성의 유사성을 반영하며, 각 Class 별로 밀도를 통해 경계 지점을 파악할 수 있을 때 활용할 수 있다. 그 결과 Distance-based 방법들이 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 Data을 Permutation하거나 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 Unlabel Data에 대한 예측을 leverage함으로써 모델의 성능을 향상시킨다. 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. Confirmation bias는 Label Data가 모델을 학습하기에도 부족할 때 심화된다. 바로 Label Data가 적은 학습 초기의 모델의 성능은 좋지 않다. 따라서 Pseudo labeling 한 결과가 실제와 달라, 실제 Dataset 에서의 클래스 비율과 상이하게 달라질 수 있다. 이를 막기 위해 Dataset의 클래스의 사전 확률을 규제항으로 도입하나, Dataset의 사전 정보를 규제항으로 사용한다는 것은 비현실적 방안이다.

그럼에도 Pseudo labeling을 연구할 필요성이 있다. 일부 몇 논문에서는 Pseudo labeling에서 Confirmation bias를 방지하였을 경우 Consistency regularization 만큼의 성능을 확보한 연구가 있다. 무엇보다 Pseudo labeling은 Consistency regularization과 상반된 방법이 아닌 병행 가능한 방법이다. Pseudo labeling의 성과를 향상시키는 것은 추후 Consistency regularization 방법과의 연계에서 좋은 결과를 가져올 것이다. 이에 본 연구에서는 Pseudo labeling의 문제점인 Confirmation bias를, 특히 Label Data가 극히 적은 상황 속에서 해결할 방안을 제안할 것이다. .

2-2) Active learning

Active learning(이하 AL) 또한 Labeling cost 줄일 수 있는 방법이다. AL은 적은 수의 Labeling으로도 전체 데이터셋이 Labeling 되었을 때와 유사한 성능을 가져올 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 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 Label Data가 부족할 때엔 적용하기 어렵다. 마지막으로 Diversity-based approach는 데이터셋의 Diversity를 보장할 수 있는 데이터를 선별하는 방법이다. 대표적으론 Coreset selection이 있다.

Coreset Selection은 모델 성능의 기댓값을 최대화하는 Data를 Sampling 한다. 이는 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}. K-Center greedy 알고리즘을 통해 모든 데이터를 덮으면서 반지름을 최소로 하는 subgraph를 형성할 수 있다. 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 Data 간의 거리를 기반으로 Sampling 하여 SSL 모델의 성능에 의존하지 않는다. 따라서 Label Data가 부족한 상황에서도 동일한 Sampling 성능을 낼 수 있다.

**3. Method**

3-1) Pseudo labeling

모델의 성능에 의존하지 않고 신뢰도 높은 Classification을 통해 Pseudo labeling을 부여하겠다. Label Data 부족하기 때문에 모델의 성능을 신뢰할 수 없다. 따라서 본 연구에서는 Coreset selection 간 형성된 subgraph를 활용하여 신뢰도 높은 Classification을 진행한다. subgraph 의 중점이자 i번째로 Sampling 된 data는 라 한다. 각 들은 Coreset selection 간 반지름 을 최소화함으로써 Sampling할 수 있다. 확률 밀도가 높은 지역의 입력값 x1, x2가 가깝다면 각각의 연관된 클래스 y1, y2도 가깝다고 가정한다. 확률 밀도는 각 subgraph에 포함된 Data의 개수로 측정할 수 있다. 즉, 확률밀도가 높은 subgraph 에 속한 Data s.t. j ∈ {1,…, | }들은 높은 확률로 중점인 와 클래스가 동일할 것이며, 은 subgraph에 속한 unlabel Data를 대표할 수 있다.

더 나아가 Active selection을 통해 형성한 subgraph들이 Dataset을 촘촘하게 덮을 수 있을 정도로 충분히 작은 반지름 을 가졌다고 가정한다. s.t. k ∈ {1,…, | } 를 와 접하고 있는 subgraph 들이라고 하자. 이때 각 subgraph의 중점 , 에 대한 클래스는 알고 있다. 이때 기하적 관점에서 의 Class가 다를 경우 서로 다른 클래스 경계값 주변부에 위치할 것이라고 유추해볼 수 있다. 반대로 모든 와 의 클래스가 동일할 때 특정 클래스 데이터 분포의 중심부에 위치할 것이다. 즉, Figure1과 같이 subgraph 간의 기하적 관계와 각 중점의 클래스를 통해서 각 subgraph들이 클래스별 데이터 분포간 어디에 위치해 있는지 유추할 수 있다. 이며, 을 성립하는 subgraph 에 에 대해, 에 속한 unlabel Data 모두를 와 동일한 Class로 Pseudo labeling을 한다. 는 자연수로 hyperparameter이다. 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 데이터들을 한번에 높은 신뢰도로 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.



[Figure 1. Subgraph 간 기하적 관계]

Coreset selection을 다회차 진행하여 subgraph들이 중첩되었을 때 또 다른 방식으로 Pseudo labeling이 가능하다. 계산상 효율을 높이기 위해 Pseudo labeling되지 않은 Unlabel Data에 한해 진행한다. 각 Coreset selection 간 subgraph들은 모든 Data을 cover 해야 한다. 따라서 Coreset selection이 p번 진행된다면 모든 Unlabeled Data들은 최소 p번의 subgraph안에 속하게 된다. 한번의 Coreset selection간 다수의 subgraph에 속할 수 있으므로 p번 이상 속할 수 있다. Unlabel Data들은 특정 Class의 중점을 가진 subgraph에 많이 속할수록 동일한 클래스일 확률이 높아질 것이다. 또한 각 subgraph들의 반지름 에 반비례하여 확률이 높아질 것이다. p번째 Coreset selection 간 각 클래스 별로 subgraph에 속하는 횟수와 , 반지름을 기반으로 softmax를 통해 확률화한다.

(3.1)

는 p번째 Coreset selection의 대상이 되는 unlabel Data의 개수이다. 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 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.

[Figure 2. P 회차 Coreset selection을 통해 중첩된 subgraph]

4-2) Choose Proper

Coreset selection 기반의 Classification의 신뢰도를 보장하기 위해 충분히 작은 반지름 을 적용해야 한다. 반지름 의 크기가 충분히 작을수록 클래스 데이터 분포간 주변부, 중심부를 명확히 구분할 수 있다. 반대로 반지름의 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 (= ||)에 달려 있다. 즉, n의 크기가 작을 땐 Labelling Cost는 줄어들지만 이 커져 신뢰도 높은 Classification을 할 수 없다. 반대로 n의 크기가 크다면 이 줄어들어 subgraph 간 간격이 줄어들어 높은 정확도의 Classification이 가능하다. 따라서 Labeling Cost와 classification 정확도 사이에 Trade off 관계가 형성된다. 따라서 적절한 Sampling size을 고를 가이드라인이 필요하다. Pseudo labeling 간 가장 피해야할 문제는 Confirmation bias이다. 따라서 Labelling cost가 있음에도, Confirmation bias을 피하는 것에 우선순위를 둔다. 각 subgraph의 속하는 데이터의 개수가 일정 이상 넘지 않도록 규제함으로써 Confirmation bias를 방지하겠다.

임의의 Sampling size 으로 Coreset selection을 수행한다. 가장 밀도가 높은 subgraph에 속하는 Data에 한정하여 다시 Coreset selection을 수행한다. 이때 각 subgraph마다 일정 이상의 데이터의 수 M 보다 적게 포함하도록 하는 반지름 을 찾는다. 반지름 은 가장 밀도가 높았던 subgraph에 속하는 데이터의 개수를 M개로 줄인다. 따라서 을 전체 데이터 셋에 적용할 때에도 모든 subgraph에 속한 데이터 수를 M개 아래로 고정할 수 있다.

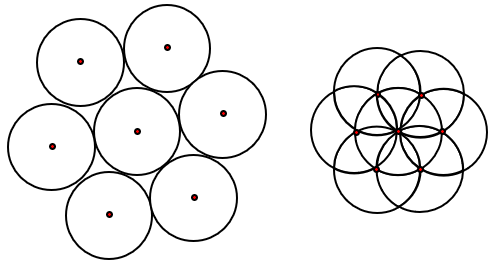
이제 반대로 반지름의 크기를 보다 작게 만드는 Sampling size n’을 찾을 것이다. Subgraph의 넓이를 통해서 Batch size의 Bound를 형성할 수 있다. 하나의 Outlier가 데이터 분포에서 벗어나 있는 경우에 대해 robust을 부여하기 위해서 Cover 하고 있는 subgraph의 넓이를 통해 Dataset S의 넓이를 정의한다. Dataset의 데이터 분포의 넓이를 S라고 할 때 S값은 고정되어 있다. 이때 S는 각 Sampling size n에 따라 n개의 subgraph들로 덮여야 한다. 이때 Subgraph안에

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.

**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*.