**Classification by Coreset selection**

GSDS 석사과정 20224314

강현구

1. **Abstract**

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

본 연구에서는 Active learning의 coreset Selection을 활용한 새로운 Classification 방법을 제안한다. Coreset Selection을 통해 Dataset의 Diversity를 보장한다. 또한 Coreset Selection 간 형성한 subgraph들은 기하적 관점에서 모델의 성능에 의존하지 않고 높은 신뢰도의 Classification을 가능하게 한다. 즉, SSL 모델에 의거하지 않은 Classification으로 Pseudo labelling 함으로써 Label Data 부족 문제를 해결한다. 본 방법은 클래스의 수가 적고 데이터 분포가 밀접해 있는 데이터셋에 대해선 높은 효율 및 정확도를 갖출 것이다. Distance base method가 적용될 수 있는 모든 분야에 적용할 수 있다. 또한 Represent learning과의 시너지가 높다. 또한 적절한 Sampling size를 제안할 것이다. 이로써 현장에서 Labeling cost를 측정할 여지를 준다.

\* 표현 수정

Labeled sample, unlabeled sample

**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. Knowledge tranfer은 Labeling이 된 Dataset과 그렇지 않은 Target 데이터셋으로 나뉜 시나리오를 가정한다. Labeling dataset으로부터 Transferable한 representation을 학습하고, Learning pipeline을 통해 전달함으로써 target dataset의 Label data 부족 문제를 해결한다. Semi-supervised learning(이하 SSL)은 부족한 Label data와 매우 많은 unlabeled data 시나리오를 가정한다. 본 연구에서는 최근 활발히 연구되고 있는 이미지 classification에 대한 SSL에 초점을 두겠다.

SSL은 크게 Consistency regularization 방법과 Pseudo Labeling 으로 나뉜다. Consistency regularization 방법은 모델의 Input에 변형이 가해져도 모델의 결과는 일정하게 유지돼야 함을 가정한다. Label Data에 변형을 가한 데이터를 생산(augmentation) 함으로써 동일한 Class의 Label Data를 생성한다. Pseudo labeling은 학습된 모델을 통해 신뢰도가 높게 나온 Unlabel Data에 대해 Pseudo labeling을 부여한다. 이후 pseudo labeling 된 sample을 모델 학습에 활용한다. 그 외에도 두 방법은 Confirmation bias를 해소하기 위해 Unlabeled data를 활용하는 방법이 다르다. Confirmation bias이란 모델을 학습하는 과정에서 unlabeled sample에 대한 잘못된 예측값이 모델의 성능악화를 가중시키는 것을 의미한다. Confirmation bias를 해결하기 위해 warm-up phase using labeled data, uncertainty weighting, adversarial attacks, graph-consistency 방법 등을 적용한다.

본 paper에서는 SSL을 통한 이미지 Classification에서 pseudo labeling 방법에 초점을 잡는다. 최근 이미지 분류 연구에 있어 Consistency regularization의 높은 성능으로 인해 선호받고 있다. **[점검 필요]** 단, Consistency regularization과 Pseudo labeling을 통해 얻을 수 있는 dataset의 diversity는 성격이 달라 둘 모두 필요하다. Consistency regularization은 label data를 다양한 방법으로 augmentation 하나, unlabeled 데이터를 직접적으로 활용하는데 제약이 있다. 반면 Pseudo labeling의 경우 Confirmation bias가 있으나 unlabeled data의 정보를 직접적으로 활용할 수 있다. 즉, Pseudo labeling은 Consistency regularization과 상반된 방법이 아닌 병행 가능한 방법이다. 각각의 방법에 대한 연구는 SSL 모델 전체의 성능 향상을 위해 필요하다. 또한 Confirmation bias를 잘 해소하여 Consistency regularization 보다 높은 성능을 낸 연구{Arazo, 2020 #9}로 보아 Pseudo labeling의 잠재력을 엿볼 수 있다.

Pseudo labeling에 있어 가장 큰 난관은 confirmation bias를 잘 해소하는 것이다. 이때 핵심은 ‘어떻게 Pseudo labeling 하는가’ 이다. 기존의 연구들은 CNN 모델 예측 값에 대해 Uncertainty weight 를 부여**[어떤 논문?]**, 신뢰도에 대한 Threshold**[어떤 논문?]** 설정 또는 초기 수렴을 위한 규제항을 도입한다. 대부분의 조치들은 CNN 모델의 성과에 의존하고 있다. **[진짜 인지 점검 필요함]** 단, CNN모델을 학습하기에도 Label data가 부족하여 모델의 성과를 신뢰할 수 없을 때가 많다. Label data가 dataset의 diversity를 반영하지 못할 수 있다. 모델이 학습하기에도 부족한 Label data는 confirmation bias를 방지하기 위한 조치들의 효과 또한 악화시킨다. 모델의 성과에 의존하지 않고 데이터 셋의 사전 지식을 활용한 경우도 있다{Arazo, 2020 #9}. 하지만 실제 현장에선 Dataset에 대해 사전 지식을 알 수 있을 것이라 기대하기 어렵다. 따라서 모델의 성능에 의존하지 않는 Pseudo labeling 방법과 데이터셋의 사전 지식에 의존하지 않는 SSL 방법이 필요하다.

Active learning의 일환인 Coreset-selection 방법을 통해 모델 성능에 의존하지 않는 pseudo labeling 방법을 제안한다. Active learning은 모델을 학습함에 있어 유용한 Data를 Sampling 하는 방법이다. 어떤 Acquisition Strategy을 활용하냐에 따라 Sampled Data가 Dataset의 특정 정보를 반영시킬 수 있다. Coreset selection은 Dataset의 Diversity를 보장하는 Data를 Sampling 한다. 또한 Sampling 간 Data들의 거리 정보를 활용하여 CNN 모델에 의존하지 않는다. 즉, Coreset selection은 Dataset의 diversity 반영 및 CNN 모델의 성능에 의존하지 않는 방법이다. 더 나아가 본 연구에서 각 Sampling point들은 기하적 관점에서 Unlabeled Data에 대해 대표성을 띄고 있음을 보일 것이다. 또한 기하적 관계를 토대로 모델의 성능에 의존하지 않고 높은 신뢰도의 Classification이 가능함을 보일 것이다. 이로써 Coreset selection을 통한 Dataset의 diversity 반영, 모델에 의존하지 않는 높은 정확도의 classification을 통한 모델 초기 학습 Confirmation bias 발생을 방지할 것이다.

본 연구는 Neural network를 활용하지 않는 distance based Classification 방법이라는 점에서 Contribution이 있다. 또한 SSL에 적용하여 Label Data 부족 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.

본 연구는 Active learning과 SSL의 연계 연구 맥락에서 contribution이 있다. Active learning과 SSL의 연계에 대한 연구는 드물다. 일부 있는 연구조차 Active learning과 SSL을 병렬적으로 활용한다는 한계가 있다. 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). Semi supervised 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 있는 다양한 영역에 폭 넓게 활용되고 있다.

SSL의 주요 방법 중 하나는 Consistency regularization 방법이다. 모델의 Input에 변형이 가해져도 모델의 결과는 일정하게 유지돼야 함을 가정한다. Confirmation bias를 효과적으로 제거하여 Pseudo labeling 보다 적극적으로 활용되고 있다**[이 말이 맞는지 점검 필요]**.

Randomize data augmentation, dropout, and random max-pooling에 활용한다. [15]

각각의 epoch에 perturbation을 적용한다.[30]

Temporal ensembling[18] : Teacher – student, add uncertainty weight[11]

Adversarial training by perturbation [20]

Contrastive loss [22] : similarity in unlabeled data

Interpolation consistency training [28] : interpolated unlabeled sample

Co training [21] : two model predict same thing , [31] calculate consistency

또 다른 SSL의 방식은 Pseudo labeling이며, 본 연구에서 주로 다룬다. Pseudo labeling은 Unlabeled Data에 대한 예측을 leverage함으로써 모델의 성능을 향상시킨다.

초기에는 pseudo labeling을 label로서 활용했다. 이때 모델의 fine tuning에 한정했다[16]

최근의 pseudo labeling은 Hard label로 활용했다. [19]. Uncertainty weight 또한 부여했다.

Graph label propagation 활용한다. [17] uncertainty score 도 활용

Threshold를 설정하여 일정 이상의 신뢰도를 가진 값에만 Pseudo labeling을 부여하는 방안을 쓰고 있다. [읽었던 것 중에 있는데 - check]

각각의 조치들은 CNN 모델의 성과에 의존하고 있다. **[진짜 인지 점검 필요함]** 단, CNN모델을 학습하기에도 Label data가 부족하여 모델의 성과를 신뢰할 수 없을 때가 많다. Label data가 dataset의 diversity를 반영하지 못할 수 있다. 모델이 학습하기에도 부족한 Label data는 confirmation bias를 방지하기 위한 조치들의 효과 또한 악화시킨다. 즉, labeled 데이터가 부족한 상황 속에서 uncertainty 및 신뢰도를 통한 confirmation bias 해소는 큰 의미가 없다. 한편 {Arazo, 2020 #9}에서는 Soft pseudo label, Mix-up data augmentation & minimum batch, Dataset의 Class 비율 활용을 통해 Confirmation bias를 효과적으로 방지하였다. CIFAR 10/100, SHVN 데이터셋에 대해 SOTA 성능을 내는 등 Consistency regularization 보다 높은 성능을 냈다. 단, [9] 연구에서도 한계가 있다. 먼저 Class-balance scenario를 가정함으로써 데이터셋의 클래스 비율정보를 규제항으로 적용한다. 즉, Dataset의 클래스 비율에 대한 사전 지식을 얻을 수 없을 경우 초기 confirmation bias를 피할 수 없다는 한계가 있다. 초기 학습에 대한 규제 없이는 Confirmation bias가 발생하여 CNN 모델의 결과를 통해 적용하는 Soft pseudo label을 신뢰할 수 없다.

2-2) Active learning [최근 논문 참고하여 내용 보강할 것]

Active learning(이하 AL) 또한 Labeling cost를 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 Labeling을 요청할 데이터를 생성한다. Stream-based Selective sampling은 새로운 sample Data에 대해서 레이블이 필요한지를 판단한다. 마지막으로 Pool-based sampling은 주어진 Dataset에서 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 모델 성능을 가장 향상시킬 것으로 기대되는 Data들을 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 를 가지게 만드는 데이터를 Sampling 하는 것과 동일하다{Sener, 2017 #4}. 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. Pseudo labeling**

Notation은 [9] 연구를 참고하였다. 우린 SSL 모델을 from a set of training samples 로 수식화한다. 이 Sample들은 unlabeled set 과 labeled set s.t. , being 은 에 대응하는 C Class에 대한 one-hot encoding label이다. 이때 은 CNN 모델을, 은 모델의 파라미터를 의미한다. 우리가 Pseudo labeling을 하는 만큼, 모든 개의 unlabeled data는 모두 pseudo label 가 가능하다고 가정한다. labeled sample와 라고 할 때, 우린 SSL을  로 재정의할 수 있다.

CNN의 파라미터 는 categorical cross-entropy를 통해 최적화할 수 있다. 이때 Loss 함수는 [9]와 유사하게 구조화 할 것이다.

(1)

이때 최초의 는 모델에 의존하지 않은 새로운 방식으로 계산한 Softmax probability를 적용한다. 이후 warm training 이후 second epoch 부터는 는 CNN을 통해 구할 것이다. 는 element-wise 하게 적용된다. 이때 핵심은 어떻게 CNN 모델에 의존하지 않고 Pseudo labeling과 Softmax probability를 계산하는 가에 달려있다. 이에 대해선 먼저 전체 구조를 설명한 뒤 구체적으로 기술하겠다. Pseudo labeling 방법은 2가지이다. 첫번째를 CS1 방법, 두번째를 CS2 방법이다 하겠다. 첫번째 방법은 Coreset selection을 한번 수행할 때마다 조건을 달성하는 특정 Data에 한해 높은 신뢰도의 Classification이 가능하다. 두번째 방법은 Coreset selection을 다회차 iteration 할 때 subgraph의 중첩 정도에 따라 클래스별 softmax probability를 계산할 수 있다. 이에 주어진 label sample과 CS1 방법으로 labeling 한 은 Hard label로 고려하며, 그 외의 의 unlabel data는 soft labeling을 적용할 것이다(s.t. ). Soft pseudo labeling은 label noise와 있는 상황 속에서 학습할 때 confirmation bias를 방지할 수 있다. [24 – check 필요]. 모델 학습 간 2회에 나눠 warm-up training을 진행한다. 첫번째에는 Labeled Data와 1차 hard pseudo labeling Data로만 10번의 epoch를 진행한다. 그 후 2차 warm-up training을 그 외 unlabeled Data에 대해 coreset 기반의 구한 softmax값을 통해 5 epoch 진행한다. [적절한 횟수는 연구를 통해 확인 필요]. 이후 softmax의 값은 CNN을 통해 산출한다.

우리는 각 single class에 대한 soft pseudo label 분포를 집중시키기 위해 규제항을 도입한다.

은 의 C class에 대한 softmax output을 의미한다. 이 규제항은 average per-sample entropy로 SSL에서는 잘 알려져 있는 규제항이다.[32] 이로써 SSL Loss는 다음과 같이 표현할 수 있다.

(2)

는 regularization의 영향력을 조절한다. 이때 은 hyper parameter test를 통해 계산하겠다.

Confirmation bias를 방지하기 위해 [9] 연구와 동일하게 Mix-up data augmentation과 setting a minimum number of labeled samples per mini-batch를 적용한다. Mix-up data augmentation[25]은 data augmentation과 label smoothing을 결합한 강력한 규제 방법이다[25]. Mix-up은 sample pair ( 와 대응하는 labels (간의 convex combination을 학습한다. 이 결합을 통해 각 샘플 간의 linear behavior을 학습하여 oscillation을 감소시킨다. [26]에서 보였듯 sample pair 간 Label을 smoothing 하는 것은 모델에 대한 과신뢰를 방지해줄 수 있다.

[9] 논문에서 보였듯 Mixup data augmentation만 적용하는 것은 confirmation bias를 다루기에 충분하지 않다. Labeled data에 대해 setting a minimum number of labeled samples for mini-batch k를 적용하여 oversampling 하겠다. 또한 높은 신뢰도를 가진 CS1 방법을 통해 구한 개의 sample에 대해서 차등적으로 k를 적용하겠다. 이로써 label data가 적어 의 경우에 대해서 pseudo labeled data가 labeled data보다 더욱 영향을 미치는 경우를 방지하겠다.

고민할 점

* 언제 softmax의 값을 새로운 방법에서 CNN으로 전환할 것인가?
* 흠.. 최소한 2차 pseudo labeling data가 CNN 보다 성능이 좋을 때까지는 활용해야 하늗데. 어떻게 이를 측정할 수 있지?

3-1) Coreset selection based pseudo labeling

Confirmation bias를 막기 위해선 CNN 모델에 대한 과신뢰를 줄이거나, 새로운 높은 성능의 Classification 방법을 고안할 필요가 있다. 먼저 전자의 방법으로서 mix-up augmentation을 적용했다. 이에 본 연구에서는 후자의 관점에서 Coreset selection 간 형성된 subgraph를 활용하여 신뢰도 높은 Classification을 제안한다. 본 연구에서 다루는 개의 Labeled data들은 개의 data들을 P회차의 Coreset selection을 통해 sampling했다고 가정하자. 이때 를 p번째 Coreset selection을 통해 Sampling 한 data set이라고 하며, 을 각 Sampling point 라고 하자. 이때 을 성립한다. 각 을 중점으로 삼는 subgraph 의 반지름을 라고 하자. Coreset selection은 K-centers algorithm과 동일하여 subgraph 의 반지름 을 최소화시키는 unlabeled data 를 로 Sampling한다. 이로써 는 을 만족한다. 각 subgraph의 밀도는 해당 subgraph에 포함된 Data 의 개수로 측정할 수 있다. 본 연구에서 제안하는 Pseudo labeling 방법은 한번의 Coreset selection 간 subgraph 간의 기하적 관계로 Classification 하는 방법과, p회차의 coreset selection을 iteration 할 때 중첩되는 subgraph을 통해 각 Class 별로 softmax probability를 계산하는 방법 2가지로 나뉜다. 편의상 전자의 방법을 first Coreset based classification(CS1), 후자를 second coreset based classification(CS2)로 부르겠다.

첫번째 CS1 방법은 subgraph 간의 기하적 관계와 밀도를 통해 Classification 한다. 편의 상 이번 coreset 간 표기를 로 단순화하겠다. 확률 밀도가 높은 지역의 입력값 x1, x2가 가깝다면 각각의 연관된 Label y1, y2도 가깝다는 SSL의 가정을 통해, 밀도가 높은 subgraph 에 속한 Data s.t. j ∈ {1,…, | }들은 높은 확률로 중점인 와 클래스가 동일할 것이다. 즉, 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 subgraph라고 하자. 각 subgraph의 중점 , 들의 클래스는 알고 있다. 이때 기하적 관점에서 의 Class가 다를 경우 와 가 서로 다른 클래스들이 겹치는 곳에 있다고 유추해볼 수 있다. 반대로 모든 와 의 클래스가 동일할 때 특정 클래스의 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. 즉, 이며, 을 성립하는 subgraph 에 속한 unlabeled Data 에 대해 와 동일한 Class로 Classification을 할 수 있다. 는 hyperparameter이다.

[Figure 1. Subgraph 간 기하적 관계]

본 Classification 방법은 클래스의 종류가 적고, Unlabeled data가 많아 특정 클래스의 분포가 명확할 수록 성능이 향상된다. 충분히 작은 반지름 , 클래스별 데이터 밀집이 보장된다면 각 클래스별 중심부에 해당하는 데이터들을 높은 신뢰도로 Classification을 할 수 있다. 추후 실험 결과로 밝히겠지만 subgraph의 밀도는 Classification 정확도에 영향을 미치지 않는다. 오히려 밀도 조건에 추가함에 따라 Classification이 가능한 Data의 수를 줄이는 악영향만을 미친다. 또한 의 크기에 상관없이 높은 classification 성능을 낸다. 은 의 크기가 클수록 작아진다. 뒤의 결과에서 밝히겠지만 의 크기에 상관없이, 즉 의 크기에 상관없이 항상 높은 성능의 classification이 가능했다. 마지막으로 Labeled data을 보다 활용할 수 있도록 다음 회차의 Coreset selection과 CS1은 그 외의 Unlabeled data에 한해 진행한다.

3-2) Calculate probability

Coreset selection을 다회차 진행하여 subgraph들이 중첩되었을 때 Cs1으로 pseudo labeling 되지 않은 에 대해 class 별 probability를 계산할 수 있다. 1번의 Coreset selection을 통해 모든 들은 1개 이상의 subgraph에 포함된다. 따라서 Coreset selection이 총 P번 진행된다면 모든 들은 최소 P개의 subgraph안에 속하게 된다. 은 특정 Class의 중점을 가진 subgraph에 많이 속할수록 동일한 클래스일 확률이 높을 것이다. 또한 각 subgraph들의 반지름 에 반비례하여 확률이 높을 것이다. 각 클래스 별로 subgraph에 속하는 횟수 와, 반지름을 기반으로 softmax를 통해 확률화한다.

(3.1)

CS1 방법이 이미 높은 정확도의 Classification이 가능함으로 Pseudo labeling되지 않은 개의 Unlabeled data에 한해 진행한다. 데이터의 클래스를 가장 높은 확률 값으로 고정할 경우 Confirmation bias가 발생할 수 있다. 따라서 soft labeling을 통해 예측에 대한 과신뢰를 방지한다. 또한 충분한 횟수의 subgraph의 중첩이 생기기 전엔 틀린 pseudo label을 부여할 확률이 높다. 따라서 Coreset selection의 회차인 P를 일정 수() 이상 진행한다. 적절한 값은 hyper parameter 값을 통해 확인하겠다.

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

마지막으로 주어진 Sampling size 에 대해서 어떻게 P 번의 Coreset selection을 나눌 것인지 점검하겠다. 각각의 가 클수록 Classification 가능 조건을 달성하는 경우가 클래스의 경우와 Classification이 가능한 데이터가 증가한다. 반면 P의 크기가 늘어날수록 한 데이터에 대해 중첩하는 subgraph의 개수가 늘어나 softmax probability의 정확도가 올라간다. 또한 한 Coreset selection을 끝낸 후, Label data을 보다 효율적으로 사용하기 위해 다음 Coreset selection은 아직 labeling 되지 않은 unlabeled data에 한해 진행한다. 따라서 = … = 로 가정했을 때에도 각각의 CS1을 통한 결과가 달라질 수 있다. 이에 어떻게 주어진 에 대해 P번의 Coreset selection으로 나누면 좋을지 경험적으로 검증하도록 하겠다.

4. 연구 환경

- Dataset : 이 논문 만큼 구체적으로 기술해야 한다!

- CIFAR 10/100

- SVHN

- Mini-Image Net

- Figure : Two moon 그림처럼 내 방법의 효과를 시각화해주면 좋을 듯!

- CNN 모델 : 13-CNN architecture [29]

- Hyperparameter 설정

- 해당 논문 처럼 매우 구체적으로 기술할 것

- 일단 이 논문의 Hyper parameter을 그대로 적용합시닷

- Certainty of incorrect prediction r\_t 를 지표로 성과 확인을 해보면 좋을 듯.

- Comparison with the state-of-the-art

- Π model [30], TE [30], MT [18], Π model-SN [22], MADNN [31], Deep-Co [21], TSSDL [19], LP [17], CCL [11], fast-SWA [29] and ICT [28]

**5. 예상 기대효과**

본 연구는 Coreset selection의 Subgraph를 기하적인 관점에서 활용한 Classification 방법을 제시한다. 이를 Pseudo labeling으로 활용함으로써 Label Data 부족으로 인한 confirmation bias를 해소할 수 있다. 더 나아가 Classification의 성능을 보장하기 위해 필요로 하는 Sample size에 대한 bound를 제시한다. 이는 현장에서 Classification 성능 확보까지 필요로 하는 Labeling Cost를 계산할 수 있는 여지를 제공한다.

본 방식은 Dataset의 특성에 따라 성능이 변한다. 데이터셋의 크기가 크고 클래스가 적어 밀도가 높은 Dataset일수록 Classification 성능이 뛰어날 것이다. 한편 본 방법은 Distance을 측정할 수 있으며, distance가 데이터 간의 특성의 유사성을 반영하는 모든 Dataset에 적용할 수 있다. 더불어 representation learning과 함께 활용한다면 시너지를 낼 수 있다는 점에서 가치가 있다.

**<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*.