**Ensemble Deep Active learning**

**: With connecting Batch Strategy, SSL and BNN**

Active learning(이하 AL)은 Labeling 했을 때 모델 학습에 효과적인 Unlabeled Data를 Sampling 함으로써 Labeling cost를 줄인다. 하지만 최근Deep learning(이하 DL)모델은 많은 양의 Labeled Data를 필요로 하여, 기존의 AL의 방식으론 제한이 있다. 따라서 DL모델의 성능을 유지하면서 Active learning를 적용하고자 하는 Deep Active learning(이하 Deep AL)이 부각된다.

Deep AL에서 중요하게 다뤄지는 문제는 'Insufficient data for labeled samples' 이다. 많은 경우 시간, 금전적인 비용 문제로 인해 충분한 양의 Labeled Data를 얻기란 어렵다. 즉, 매번 1개의 unlabeled data를 Sampling 하는 기존의 AL 방식으론 DL 모델의 요구를 맞추기란 비현실적이다. 더불어 DL 모델에서 매번 Label 데이터를 1개씩 추가하며 학습하는 것은 비효율적이다. 따라서 Labeled Data 외의 다른 방식으로 추가적인 정보를 구하는 것과, Unlabeled data을 Batch 단위로 Sampling 할 필요가 있다.

최근 연구에선 AL과 DL을 같이 활용함으로써 부족한 Label Data의 빈자리를 채우고자 한다. 예로 들어 Cost-Effective Active learning(CEAL)(Wang, Zhang, Li, Zhang, & Lin, 2016)에서는 Unlabeled Data에 대해 Pseudo label을 부여했다. 또한 Weighted Incremental Dictionary Learning(WI-DL)(Liu, Zhang, & Eom, 2016)에서는 Unlabeled data을 통한 feature learning을 통해 학습 모델을 fine-tuning 하였다. 한편 Generative Adversarial Active Learning(GAAL)(Zhu & Bento, 2017)에서는 GAN 모델을 활용하여 정보가 많은 sample들을 만들었다. 한편으론 어떻게 Batch pool을 선택할 것 인가에 대한 연구되고 있다. 주요 방법으로는 1) Uncertainty-based and Hybrid Query Methods, 2) Deep Bayesian Active learning(이하 DBAL) Methods, 3) Density Based Methods가 있다(Ren et al., 2021). 첫번째 방식은Traditional 모델(ex- SVM, KNN)을 통해 uncertainty를 계산하며 이를 기준으로 Sampling한다. 하지만 traditional 방식이 모델의 uncertainty를 잘 반영하지 못한다는 한계가 있다. DBAL은 Bayesian Neural Network을 통해 모델의 Uncertainty를 보다 정확하게 측정한다. 마지막으로 Density Based method는 Original dataset의 feature space의 분포를 고려하여 Sampling 하며, 대표적으로 Core set(Sener & Savarese, 2017)가 있다.

지금까지 알아본 바에 의하면 Data expansion과 연관된DL과 Batch sampling을 함께 고려하는 연구는 거의 없다. Simeoni의 연구(Siméoni, Budnik, Avrithis, & Gravier, 2021)에서 pseudo labeling 및 pre-training 방법과 다양한 Batch Sampling 방식을 orthogonal하게 적용하여 모델의 성과를 향상시켰다. 그 외의 연구는 찾지 못했다. Simeoni의 연구 또한 한계점이 있다. 첫째, Data Expansion 목적의 DL과 Batch Sampling 방식을 Orthogonal 하게 구성한 점이다. 이로 인해 각 과정이 별개로 수행되어 계산량이 가중되었으며, 각 방법 간의 시너지를 고려하지 못했다.

이 논문에서는Batch Sampling, Data Expansion DL, 더 나아가 Learning Model까지 연결할 것이다. 먼저 Batch Strategy로는 Coreset(Wang et al., 2016)의 Batch Strategy를 차용한다. 이후 Coreset의 결과물인 Batch Sampling point에 대한 Adjacency Matrix w를 활용하여 Pseudo label을 부여할 것이다. Pseudo labeling은 Learning model 학습 후 초기화할 것이다. 이후 Labeled Data를 통해 Bayesian Neural Network(이하 BNN)를 학습한다. BNN 학습 간 계산한 각 Data별 Epistemic uncertainty를 가중치로 고려하여 다음 Batch sampling 간 Coreset의 Sampling 방식과 함께 고려한다.

위의 방식을 통해서 각각의 방식을 독립적으로 시행하는 것에 비해 확연히 계산양이 줄어들 것은 분명하다. Batch Sampling과 Learning Model의 계산양만으로 Data Expansion 방법까지 적용할 수 있다. 또한 성능이 향상될 것으로 보인다. Batch Sampling의 방식은 크게 Uncertainty을 고려하는 방법과 Data feature을 고려하는 방법으로 나뉜다. 하지만 내 지식 안에선 현재까지 이 두가지를 함께 고려한 연구는 없는 것으로 안다. Simeoni의 연구(Siméoni et al., 2021)에서 Batch 방식은 크게 상관이 없고 오히려 Random이 더 좋았다고 하나, 이는 각 단계별로 orthogonal 하게 방법을 적용해서 그런 것으로 판단된다. 따라서 Coreset Batch sampling 방식이 데이터의 특성을 잘 반영하는 것으로 알려진 바(Ren et al., 2021), Coreset의 결과물을 통해 Pseudo labeling 했을 때 보다 좋은 성능을 낼 것으로 기대한다.

**<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*.