**Ensemble Deep Active learning**

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

Active learning(이하 AL)은 모델 학습에 효과가 좋은 Unlabeled Data를 Sampling 함으로써 Labeling cost를 줄인다. 하지만 최근의 Deep learning(이하 DL)모델은 많은 양의 Labeled Data를 필요로 하여, 기존의 AL의 방식으로는 제한이 있었다. 따라서 DL의 성능을 유지하면서 Active learning를 적용하고자 하는 Deep Active learning(이하 Deep AL)이 부각되고 있다.

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

최근 연구에선 AL과 DL을 같이 활용함으로써 부족한 Label Data의 빈자리를 채우고자 했다. 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 연구가 진행되었다. 주로 사용되고 있는 방법으론 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 Uncertainty을 aleatoric uncertainty와 Epistemic uncertainty로 구분함으로써 보다 정밀하게 Batch를 Sampling한다. 마지막으로 Density Based method는 Data set 관점에서 Sampling을 한다. 대표적으로 Core set(Sener & Savarese, 2017)이 있으며, 전체 Original dataset의 feature space의 분포를 잘 표현한다.

지금까지 알아본 바에 의하면 기존의 연구에선 Batch Strategy와 ML 방법을 함께 고려한 연구는 거의 없다. Simeoni의 연구(Siméoni, Budnik, Avrithis, & Gravier, 2021)에서 다양한 Batch Strategy와 Semi-supervised learning 및 Unsupervised learning을 orthogonal하게 적용하여 모델의 성과를 향상시켰다. 그 외로 Batch Strategy와 다른 ML 방식을 적용한 연구는 아직까지 찾지 못했다. Simeoni의 연구에서는 크게 2가지 문제점이 있다. 첫째, Batch strategy 와 Semi-supervised learning을 Orthogonal 하게 구성하여 필요로 하는 계산양이 많으며, 각 방법 간의 시너지를 고려하지 못했다. 둘째, Pseudo labeling을 학습 이후에도 유지하여 오차가 중첩되는 것을 허용한다.

이 논문에서는 Batch Strategy와 ML 방법 뿐만 아니라, Learning Model과 Batch Strategy도 연결할 것이다. 구체적으로는 Batch Strategy로는 Coreset(Wang et al., 2016)의 Batch Strategy를 차용한다. 이후 Batch Strategy의 결과물 중 하나인 Batch pool data들에 대한 Adjacency Matrix w를 활용하여, 모든 Unlabeled Data에 대해 p-distribution 형태로 Pseudo label을 부여할 것이다. 이때 Adjacency matrix가 이미 구해져 있기에 Pseudo label 부여에는 계산을 거의 필요로 하지 않는다. 마지막으로 learning model로 Bayesian Neural Network(BNN)를 사용한다. BNN을 학습한 이후 pseudo label을 초기화한다. 마지막으로 BNN 학습 간 계산한 Epistemic uncertainty의 값을 활용하여, 다음 Batch sampling 간 우선순위를 가질 Dataset을 추린다. 이후 다음번의 Batch Sampling간 높은 Epistemic uncertainty를 가진 data를 같이 고려한다.

먼저 Batch Strategy의 결과를 Pseudo labeling에, 또한 BNN의 결과를 Batch sampling에 재활용하기 때문에 각 방법을 독립적으로 쓰는 것에 비해 계산양이 대폭 감소할 것이다. 더불어 Batch sampling 간 높은 Uncertainty Data와 Coreset의 representation of feature space를 같이 고려하여 성능이 향상될 것으로 보인다. Simeoni의 연구(Siméoni et al., 2021)에서 Batch 방식은 Random이 더 좋았다고 하나, 이는 각 단계별로 orthogonal 하게 방법을 적용해서 그런 것으로 판단된다.

<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*.