Current Data Science: Active Learning

Yash Manne
DATA 590
11/23/2023

Historically, machine learning models were trained on predefined, annotated datasets. However, annotating data is often a resource-intensive endeavor. Consequently, a concerted effort has been made to streamline the process by training only on valuable samples. Enter active learning, a dynamic paradigm that adeptly identifies the most informative data points for annotation and uses this new set to continuously update a model initially trained on a small dataset. Rather than passively relying on randomly chosen labeled instances, active learning algorithms intelligently identify ambiguous or challenging images, prompting human annotators to label those specific cases. By iteratively refining the model with progressively informative examples, active learning enhances performance and maximizes the annotation process's efficiency.

The impact of active learning relies heavily on the **acquisition function** that determines what data points from a pool of unlabeled data points are chosen for annotation. Traditionally, these acquisition functions aimed to approximate the informativeness or the uncertainty of a new data point – essentially finding images for which the current model doesn't have "good" predictions. Metrics include maximum entropy, least confidence, variation ratio, and margin-based uncertainty.[1] However, focusing solely on informativeness/uncertainty leads to excessive sampling of outliers or biased data for solving a specific problem in the ML context – improving class A prediction at the expense of class B or C. Another popular option is to approximate the representativeness of data points (diversity sampling) with clustering & discriminative learning.[2] A hybrid combination of the two has been the dominant approach for active learning for deep learning.

Active learning (AL) has been successful in low-throughput model training with traditional, non-deep learning machine learning models. However, it is innately challenging to meld AL into deep learning workflows for a variety of reasons:

1. Most uncertainty measures rely on accurate probability estimates for each class. Deep learning probability estimates from softmax functions are often overconfident as it is hard to represent uncertainty in deep learning models.
2. Current AL approaches have been optimized to pick the best next image. However, the computational complexity of deep learning pushes for selecting batches of images at a time – while accounting for relationships between the chosen set of images.
3. Few AL approaches have focused on high-dimensional data.

The survey of active learning methods in deep learning summarizes a variety of methods to abate some of the challenges listed above. The paper categorizes acquisition strategies into three main groups:

1. **Data & Model Independent**: Most classical AL methods, Entropy, CoreSet, and Batch Active learning by Diverse Gradient Embeddings (BADGE)[4] can be applied to various contexts because they only impact the data selection process and not the storage & training. However, the performance of these methods is limited in that it might not exceed the performance of training on full data.
2. **Enhancing Via Data:** Methods such as Cost-Effective AL (CEAL) use pseudo-labeling while other methods like AdvDeepFool & BGDAL augment the existing data set with synthetic data. It should be noted that data augmentation might waste computing because it can generate samples that are not guaranteed to be informative.
3. **Enhancement Via Model Training**: Some methods like Learning Prediction Loss (LPL), Wasserstein Adversarial AL (WAAL), and Bayesian Active Learning by Disagreement (BALD)[2] involve modifying the training process to predict the confidence/loss of the model's predictions & incorporate it directly into the model's objective function. Another potential enhancement is to use Monte Carlo Dropout – in which uncertainty in the weights model weights induces prediction uncertainty by marginalizing over the approximate posterior distribution using Monte Carlo integration – to approximate the class probabilities for classical methods such as Entropy.

While AL enhancement strategies lead to significant improvements in overall model performance, an evaluation of the various deep active learning strategies shows that a simple Entropy-based approach without any enhancements still outperforms random sampling and is within 5% of the best-performing complex strategies such as WAAL & LPL.[1] Additionally, in the case of good pre-training, the Entropy-based approach actually performs better than most other approaches with a limited number of new labels.

As such, for our project, which requires selecting subsets of images to validate our pre-trained CNN model performance, we must consider whether we would over-engineer our solution by applying these new novel methods. Instead, we plan to initially stick to the simple Entropy-based approach, which doesn't require additional computing. If this doesn't work, we can consider other data & model-independent strategies such as BADGE & margin-based uncertainty before slowly leaning into enhanced strategies like CEAL[3]. We plan to use WAAL & LPL strategies if and only if required. While active learning may be a viable solution for our project, we must ultimately remember the goal of our project is not to simply boost model performance but to serve the desires of marine researchers who may desire a more customizable image sample selector for validation

References:

1. Zhan, X., Wang, Q., Huang, K.-H., Xiong, H., Dou, D., & Chan, A. B. (2022). A Comparative Survey of Deep Active Learning. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2203.13450

Others:

2. Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep Bayesian Active Learning with Image Data. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/1703.02910
3. Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2017). Cost-Effective Active Learning for Deep Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology, 27*(12), 2591–2600. doi:10.1109/tcsvt.2016.2589879
4. Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., & Agarwal, A. (2020). Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/1906.03671