Homework #1

CSE584

Darshan Chudiwal
dsc5636@psu.edu

Paper 1 : Deep Active Learning by Leveraging Training Dynamics

Wang, H., Huang, W., Wu, Z., Tong, H., Margenot, A. J., & He, J. (2022). Deep active learning by leveraging training dynamics. *Advances in Neural Information Processing Systems*, *35*, 25171-25184.

1. What problem does this paper try to solve, i.e., its motivation

Active learning tackles the problems posed by the high cost of labelling vast amounts of data required for deep learning. Theoretical analyses and insights applicable for classical active learning methods do not work well on over-parameterized deep neural networks as the functions are much more complex. The performance for the models using these insights has been mostly underwhelming.

Further, the authors mention that literature in deep active learning has relied on experimental results, without any kind of theoretical guarantee which has raised questions on its reliability.

To this end, the paper proposes a model with a theoretical guarantee backing its performance that can work on the over-parameterized regime of models.

2. How does it solve the problem?

The paper proposes a model it coins '*dynamicAL'* to solve the problem discussed. First, the authors introduce *training dynamics,* a proxy for how quickly a model's loss is reduced. It also describes how the model is trained. The model selects data points based on the impact it has on increasing *training dynamics,* essentially choosing points that will help the model find the optimal solution faster. This is made possible by using the Neural Tangent Kernel which helps in picking the correct data points to be labelled.

The goal is to train faster, as training faster leads to better generalization. The same is proven theoretically and empirically in the paper. The theoretical proof centers that faster training with a strong alignment of true labels onto the NTK space helps the model learn more general patterns in the data. Thus, leading to a better worst-case generalization error.

The model is then tested on popular datasets like CIFAR-10 with results outperforming active learning techniques - uncertainty sampling and random sampling.

Thus, a technique that works well with over-parameterized models with theoretical guarantees is provided.

3. A list of novelties/contributions

- Bridges the gap between theoretical findings in deep learning with real world use.
- It outperforms traditional methods like uncertainty sampling and random sampling in both convergence speed and accuracy.
- Provides theoretical proof for connection between faster convergence and better generalization. Thus, it provides a theory driven method to improve on generalization.
- DynamicAL-trained models generalize better to unseen data compared to models trained with traditional methods on multiple datasets (CIFAR10, SVHN, and Caltech101).
- Easy to implement with time complexity being scalable as it uses pseudo-labeling and subset approximation to remain practical for large datasets.

4. What do you think are the downsides of the work?

- Active Learning Focuses on Labeling Costs but Ignores Model Retraining Costs
- The paper proposes the use of Pseudo-Labeling, which could be inaccurate, and there is limited exploration of this uncertainty.
- While the paper introduces Pseudo-Labeling and Subset approximation for scalability, computing the NTK matrix is still an expensive task that the paper does not delve into.
- The datasets used for empirical analysis are all standard ones like CIFAR10, while in practice real-world datasets have much more noise.

Paper 2 : Deep Active Learning for Anchor User Prediction

Cheng, A., Zhou, C., Yang, H., Wu, J., Li, L., Tan, J., & Guo, L. (2019). Deep active learning for anchor user prediction. arXiv preprint arXiv:1906.07318.

1. What problem does this paper try to solve, i.e., its motivation

The paper tackles the problem of predicting anchor users across different social media. To begin with, an anchor user is a user with accounts on different social media for instance, a user could have an account on Facebook and Twitter. There are various benefits in doing this, for instance, cross-network recommendations, link prediction and information diffusion analysis all rely on these predictions.

A supervised learning model can be accurate in predicting anchor users but requires enough labeled data points which are not easy to come across. The paper aims to reduce the number of labeled examples needed to predict an anchor user via Active Learning.

2. How does it solve the problem?

First, two neural networks are used to generate vector representations of the users. These networks consist of a convolution and a deconvolution neural network that capture meaningful information about a user. These vectors are combined and classified as an anchor or a non-anchor pair.

Then, some query strategies are used to select the most informative pair of users to label. The strategies consist of – Cross Network Structure Aware Information Entropy, Cosine Similarity and Expected Error Reduction.

Finally, a pair of users is selected based on either a query strategy that worked well in the past or trying other strategies to find a better choice balanced by a multi-arm bandit algorithm approach. This flow ensures that only the most impactful pair of users is labeled.

3. A list of novelties/contributions

- The paper is one of the first ones to rely on active learning for anchor prediction.
- The paper presents a model it coins DALAUP. The DALAUP model uses Deep Learning for user representation and Active Learning for labeling the most informative pair of users.
- The paper presents a strategy that includes three query strategies to aid in making decisions on labeling. The three query strategies ensure an efficient labeling process.
- The model outperforms state-of-the-art methods such as DeepWalk, PALE and IONE.
- The model is tested on real world datasets of Facebook and Foursquare to benchmark its performance against state-of-the-art techniques often used for the task.

4. What do you think are the downsides of the work?

- The DALAUP model is a complex combination of different neural networks that share weights, a three-pronged query strategy for active learning, and a multi-armed bandit algorithm in query selection. Implementing such a model could be challenging, especially considering the nature of the data.
- The Fourspace dataset is relatively small in size when compared to social media giants like Facebook and LinkedIn. The complexity of the model makes scalability on such large datasets doubtful and requires further exploration which is not explored in the paper.
- Finally, it is doubtful that the DALAUP model can be used for other applications as the model's highly specialized nature makes it difficult to adapt to other use cases.

Paper 3 : Gone Fishing: Neural Active Learning with Fisher Embeddings

Ash, J., Goel, S., Krishnamurthy, A., & Kakade, S. (2021). Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems, 34*, 8927-8939.

1. What problem does this paper try to solve, i.e., its motivation

- The paper's motivation is to provide an active learning method that works well with deep learning techniques.
- Present research according to the paper tends to work only on convex functions which deep learning models tend to violate.
- Furthermore, these methods are also computationally expensive to use as they query single data points.
- Finally, the paper suggests there is a need for a method that is both scalable and provides reasonable performance.

2. How does it solve the problem?

The method is as follows –

1. The paper presents BAIT (Batch Active Learning via Information maTrices). The algorithm is based on exploiting the Fisher Information Matrix applied to the last layer of the neural network to query the most informative data points.
2. The most informative point is chosen for the batch of data points to be labeled. After adding the data point to it, the next data points added are chosen by sorting on the greatest reduction of uncertainty in the model. This forward greedy optimization approach also considers diversity and tries to avoid redundant data points that teach the model the same thing.
3. This is done iteratively until a batch of n data points is added to the batch to be labelled. The data isn't submodular to really allow for total diversity. To address this, a backward greedy optimization is applied. A factor of 2 is chosen to select 2N data points first. Then, N data points are removed in a backward manner where the least informative N data points are removed leaving N data points in the batch to be labeled.
4. The batch is then sent to be labeled. The model is then trained again using the newly labeled data.
5. This process is repeated until the desired performance, or the best possible performance given the resources is achieved.

The use of Fisher matrix in the last layer of the deep neural network to search for the most important data points helps the method work well with the network. This also helps make it tractable considering the linear nature of the last layer of a neural network. The use of the Fisher matrix and a greedy optimization in addition allows for batch selection which makes the technique scalable and computationally efficient.

3. A list of novelties/contributions

- The paper proposes BAIT, a batch querying algorithm for active learning designed to perform well on non-convex functions.
- The use of Fisher information matrix for pruning data points from the batch applying it to neural networks for active learning is a novel approach to the problem.
- The paper tries to bridge the gap between classical active learning research to deep neural networks.
- The paper provides a technique that is scalable in terms of when the labeling cost is much higher than the computational costs.
- The forward-backward greedy optimization approximation tries to keep the data points diverse while reducing the computing costs.

4. What do you think are the downsides of the work?

- While addressing the need for a scalable technique compatible with deep neural networks, the paper mentions that BAIT is more 'computationally intensive' than BADGE.
- BAIT focuses on the last layer which could lead to a loss of important information from the previous layers reducing the accuracy of the model.
- Large sized batches aren't explored in the paper. Large batches would allow for more scalability with datasets that are super large.
- There is a lack of theoretical guarantee despite the paper trying to provide a theoretical basis for the algorithm. The forward-backward greedy optimization is a greedy approximation that the authors claim works well in practice. Hence, there is no surety in this technique working well in different scenarios.
- Calculating the Fisher matrix can be computationally intensive which isn't addressed properly in the paper. Finally, memory for the matrix for large datasets can be a limiting factor as well.