Name: **Rohan Prasad** Date: 10/15/2024
Email: rpp5524@psu.edu
PSU ID: **980707395**

CSE 584 Homework 1

Paper 1: Variational Adversarial Active Learning

https://arxiv.org/abs/1904.00370v3

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

Answer: The research study attempts to solve the problem of conventional active learning strategies and heuristics where it **selects a small subset of the dataset** to label but most of these heuristics are task specific and may not apply for other tasks.

The goal is to provide a technique that reduces the expense and labor associated with training models by identifying the most useful data points for labeling, especially in situations when labels are hard to come by or prohibitively expensive. **Variational Adversarial Active Learning** is unlike conventional active learning techniques as it is task agnostic and applied to a plethora of tasks like image captioning and semantic segmentation.

2. How does it solve the problem?

Answer: The authors present Variational Adversarial Active Learning (VAAL),a **pool-based semi-supervised active learning method**. This method makes use of an **adversarial network** that distinguishes between labeled and unlabeled input, and a **Variational Autoencoder (VAE)** to learn a latent representation of the data. The representativeness of data points in the latent space, which the VAE attempts to make indistinguishable to the adversarial discriminator, is the system's primary focus instead of task-specificity.

It selects instances for labeling from the unlabeled pool that are sufficiently different in the latent space learned by the VAE to maximize the performance of the representation learned on the newly labeled data.

By mapping the sets of **labeled** and **unlabeled** data into a shared embedding, the VAE learns a latent representation. In this domain, they employ an adversarial network to accurately distinguish one from the other. The discriminator and the VAE are framed as two-player mini-max games, where the discriminator network learns **how to discriminate** between the two sets of data points, while the VAE is trained to learn a feature space to fool the adversarial network into predicting that all data points, from both the labeled and unlabeled sets, are from the labeled pool.

Implementation details

The authors perform image classification and semantic segmentation tasks on VGG16 and dilated residual network (DRN) with an unweighted cross-entropy loss function.

They evaluated VAAL on CIFAR10 and CIFAR100 both with 60K images of size 32 × 32, and Caltech-256 which has 30607 images of size 224 × 224.

3. A list of novelties/contributions

Answer: Some of the contributions that the authors present in the paper are as follows:

- The authors introduce an active learning model that is **independent of task** performance. It does not interfere with the process of selecting samples.
- They use an adversarial learning strategy in conjunction with a VAE to control the latent space of data representations in order to achieve effective sample selection.
- The authors **introduce an algorithm for a sampling strategy**. In order to present this oracle which non-adversarially provides erroneous labels, realistically, they applied a targeted noise on visually similar classes. They use the probability associated with the discriminator's predictions as a score to collect b number of samples in every batch predicted as "unlabeled" with the lowest confidence to be sent to the oracle. This demonstrates the robustness against different levels of **labeling noise**.

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

Answer:

The caliber of the latent space that the VAE learns has a significant impact on how effective VAAL is. A considerable degradation in the model's performance could occur **if the latent representation is not meaningful** or discriminative enough.

The simultaneous use of adversarial networks and VAE may necessitate significant **computational resources**, which renders it less feasible for contexts with limited resources or for very big datasets.

Although the work presents comprehensive findings on image datasets, it **does not show how this method may be applied to non-visual data** or other kinds of jobs (not just classification and segmentation). This might make it less applicable in a variety of real-world situations.

Paper 2: Active Learning for Convolutional Neural Networks: a Core-Set Approach

https://arxiv.org/abs/1708.00489v4

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

Answer: The motivation for this research is that the cost and effort needed to manually label the vast amounts of data is very high. Many active learning heuristics like Settles (2010) and MacKay(1992), selectively label the most informative data points which is **ineffective when applied to CNNs in a batch setting** due to **correlated sampling**. It is imperative to query labels for a large subset of the dataset in each loop which results in correlated samples even for moderately small subset sizes.

This is why the authors explored a more effective approach to active learning specifically for CNNs. They define it as a core-set selection problem, which means choosing a set of points such that a model learned over the selected subset is competitive for the remaining data points.

So the problem they are trying to solve is what the best way is to select data points to label so that the highest accuracy can be achieved given a fixed labeling budget.

2. How does it solve the problem?

Answer:. They perform the core-set selection without using the labels.They provide a rigorous bound between an average loss over any given subset and the remaining data points via the geometry of the data points to attack the **unlabeled core-set problem for CNNs**. Some related work on core-set approach are methods like core-sets for SVM (Tsang et al., 2005) and core-sets for k-Means and k-Medians (Har-Peled & Kushal, 2005) but no such method for CNN exists.

By connecting the issue to the **k-Center geometric problem** and offering an effective greedy approximation strategy to address it, the research provides a theoretical basis for this method. Using a training error, a core-set loss, and a generalization error as upper bounds on the active learning loss, this approach selects samples.

$$E_{\mathbf{x},y\sim p_{\mathcal{Z}}}[l(\mathbf{x},y;A_{\mathbf{s}})] \leq \underbrace{\left|E_{\mathbf{x},y\sim p_{\mathcal{Z}}}[l(\mathbf{x},y;A_{\mathbf{s}})] - \frac{1}{n}\sum_{i\in[n]} l(\mathbf{x}_i,y_i;A_{\mathbf{s}})\right|}_{\text{Generalization Error}} + \underbrace{\frac{1}{|\mathbf{s}|}\sum_{j\in\mathbf{s}} l(\mathbf{x}_j,y_j;A_{\mathbf{s}})}_{\text{Training Error}}$$

$$+ \underbrace{\left|\frac{1}{n}\sum_{i\in[n]} l(\mathbf{x}_i,y_i;A_{\mathbf{s}}) - \frac{1}{|\mathbf{s}|}\sum_{j\in\mathbf{s}} l(\mathbf{x}_j,y_j;A_{\mathbf{s}}),\right|}_{\text{Core-Set Loss}}$$

Algorithm 1 k-Center-Greedy

Input: data $\mathbf{x}_i$, existing pool $\mathbf{s}^0$ and a budget b
Initialize $\mathbf{s} = \mathbf{s}^0$
repeat
  $u = \arg\max_{i \in [n] \setminus \mathbf{s}} \min_{j \in \mathbf{s}} \Delta(\mathbf{x}_i, \mathbf{x}_j)$
  $\mathbf{s} = \mathbf{s} \cup \{u\}$
until $|\mathbf{s}| = b + |\mathbf{s}^0|$
return $\mathbf{s} \setminus \mathbf{s}^0$

Although the authors suggest an **advanced approach using a Mixed Integer Program (MIP)** to further optimize this answer, the greedy algorithm offers a good starting point. The greatest distance between any data point and its nearest chosen center from the subset is limited by a value δ, which characterizes the MIP. By conducting a binary search between the greedy algorithm's result and half of its value, this parameterization enables the algorithm to iteratively test whether it is possible to arrive at a better answer than the first greedy approach. This ensures that the best solution is within this range.

Implementation details
The authors use the L2 distance between activation functions of the final FC layer. For all experiments they used VGG-16. They optimized all models using RMSProp with a learning rate of 1e−3 using Tensorflow They also trained CNNs from scratch after each iteration.

3. A list of novelties/contributions
Answer:
Some of the contributions that the authors present in the paper are as follows:

- The authors reframe active learning by considering it to be a core-set selection. Instead, they select a subset of the data **(the "core-set")** that represents the entire dataset, compared to employing conventional active learning heuristic methods that might select data points ineffectively, especially in batch settings for CNNs. This method makes the labeling process more effective and less expensive.

- This study makes a substantial theoretical addition by establishing a bound that, **using the geometry of the data points**, links the average loss over a selected subset to the possible loss across the entire dataset. This bound gives one a way to express in mathematical terms the degree to which a selected subset will generalize to the full data space. It explicitly addresses the dispersion and coverage of the data points by the core-set and is based on the geometric features of the data points.

- The authors create an effective greedy algorithm based on the well-known combinatorial optimization problem, **the k-Center problem**, to apply the core-set technique. In order to ensure that the core-set is as representative as possible, this technique aims to minimize the maximum distance that any data point has to its nearest point in the selected subset. The algorithm's greedy character makes it computationally feasible while maintaining a

good approximation to the optimal solution, which makes it useful for the big datasets that are commonly used to train CNNs.

4. What do you think are the downsides of the work?
Answer:
The authors experiment with many of these active learning heuristics and find that **they are not effective when applied to CNNs.** They argue that the main factor behind this ineffectiveness is the correlation caused via batch sampling.

Although the approach is demonstrated to be successful, **there was not a thorough discussion of the computational complexity** or viability of scaling to very big datasets.Choosing the right initial data subset/cire-set can have a big impact on how well the core-set strategy works. The problem of “cold start" arises again here. The suggested method's application to other deep learning models or tasks beyond picture classification may be limited due to its exclusive focus on CNNs.

The study only focuses on CNNs which include tasks like image captioning, which makes it incomplete as CNNs are capable of much more than just vanilla image captioning. It has been demonstrated that the Core-set technique is a successful representation learning approach for large-scale picture classification problems and that it functions optimally in small class sizes. However, **performance declines as the number of classes increases**. Furthermore, it appears that distance-based representation techniques, such as Core-set, are useless for high-dimensional data because high-dimensions p-norms suffer from the distance concentration phenomenon, also known as the “curse of dimensionality”

Paper 3: ActiveLLM: Large Language Model-based Active Learning for Textual Few-Shot Scenarios

https://arxiv.org/abs/2405.10808

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

Answer: The study discusses the drawbacks of conventional active learning techniques, particularly the "**cold start**" issue, which necessitates a sizable starting dataset in order to produce insightful selection results. In circumstances involving few-shot learning, when data is sparse, and model mismatch, where the model used for training and querying differs, this issue is made worse. These difficulties mean that traditional active learning techniques frequently fall short of improving pre-trained models, like BERT, in few-shot environments. The idea is to use Large Language Models (LLMs) such as GPT-4 to address these drawbacks and create an active learning technique that is more successful.

Model mismatch scenarios arise where the instance selection model (query model) differs from the model used for final tasks (successor model), active learning can yield limited gains, The work in the paper attempts to

2. How does it solve the problem?

Answer: It proposes ActiveLLM, a pool-based sampling method that operates in batch mode, it selects a subset from a pool of unlabeled data for querying an oracle. This method is particularly suited for scenarios involving **model mismatch**. It utilizes instruction-tuned LLMs as query models, while allowing the choice of a successor model to be independent of these models. ActiveLLM does not train the instruction-tuned LLMs during the active learning process, enabling direct application to the unlabeled dataset without encountering the **cold-start** problem.

Some **earlier work** on Active learning includes strategies like Least Confidence, Core-Set and Discriminative Active Learning (D-AL) which are shown to improve BERT based classifiers when the training data is small. Some hybrid active learning strategies combining uncertainty-based and diversity methods, utilizing SentenceBERT embeddings are also introduced.

A novel approach to the cold start problem is using active learning by utilizing BERT's pre-existing masked language modeling objective along with clustering to select instances. This leads to better confidence scores in the initial stages.

Some other Active learning strategies include matching the query model and the successor model. Other explored strategies are Prediction Entropy (PE), Breaking Ties, LC, and Contrastive Active Learning, using Active learning with MCDO (Monte Carlo Dropout)

3. A list of novelties/contributions

Answer: Some of the contributions that the authors present in the paper are as follows:

- The authors introduce **ActiveLLM** which is a unique approach to active learning that uses LLMs like GPT-4 for instance selection, which can be used for few-shot learning scenarios and it is good at overcoming the **cold start** problem.
- The methodology they used **decouples the successor model dependency and the query model**, enhancing scalability and practicality. It employs instruction-tuned LLMs as query models, while allowing the choice of a successor model to be independent of these models. Also ActiveLLM does not require training the LLM during the active learning process, making it efficient in terms of time and resources.
- The authors introduce 2 variations of ActiveLLM -
    - one that does not incorporate feedback, addressing few-shot learning scenarios,
    - another that does incorporate feedback, suitable for general scenarios involving iterative querying.

4. What do you think are the downsides of the work?
Answer: Even Though GPT-4 has vast capabilities, it is extremely expensive & resource intensive to deploy & APIs may be limiting because of privacy concerns. As explained in the paper, BERT-like models have benefited more from this approach in Active Learning.

In the research work and experiments in the paper, the authors have used the chat version of the models like GPT 4. They also claim that different results might be obtained if they make use of the n models respective API Calls for the experiments. This difference might be insignificant for models offered by Llama and Mistral but providers like Google and OpenAI which hide parameters, preceding prompts, and exact version declarations.

This approach uses prompts which might hog the context window of the LLM. LLMs tend to forget the task or are not able to reason over the various instances.

Although the authors did not receive better results incorporating the true labels during iterated querying, this behavior may be because they only evaluated ActiveLLM on at most three-class multi-classification tasks. But actual real world applications will have more than 3 classes of course.