CSE 584 - MACHINE LEARNING
HOMEWORK - 1

Yashita Watchpillai
PSU ID: 951837360
Email ID: yxw5822@psu.edu

Paper I - Deep Bayesian Active Learning with Image Data

1. **Motivation**

   This paper dives into how deep learning can be improved for active learning with image data, focusing on two key challenges:
   - Need for Large Datasets: Deep learning models need a lot of labeled data to work effectively, but collecting and labeling this data especially in fields like medical diagnosis can be expensive and time-consuming.
   - Uncertainty Handling: Active learning is about choosing the most valuable data points to label, and to do that, we need to understand where the model is uncertain. But traditional deep learning models usually just give a single prediction without showing how confident they are, which is a problem.

   The aim of the paper is to make deep learning models more efficient for active learning with images and improve their ability to handle uncertainty.

2. **Solution**
   - Bayesian CNNs: Each anomaly detection method learns distributions over model parameters using Bayesian CNNs to model uncertainties.
   - MC Dropout: Monte Carlo dropout with dropout at train and inference time provides uncertainty estimation. This avoids overfitting by using dropout during testing and sampling predictions at test time.
   - Active Learning: The work investigates the selection of points based on model uncertainty, including approaches that consider the most uncertain predictions.
   - Applications: The approach is applied to both MNIST and ISIC 2016 datasets. CNNs are used since they work well with image data.
   - Results: For MNIST, the method achieved a 5% test error with just 295 labeled images, compared to 835 with random sampling. In skin cancer detection using ISIC 2016, a fine-tuned VGG16 CNN identified melanoma with fewer labeled samples.

3. **Novelties**
   - Introduces Bayesian CNNs within an active learning setup, offering a way to estimate uncertainty something traditional CNNs don't inherently provide.
   - Uses MC Dropout to efficiently approximate Bayesian inference, enabling practical uncertainty estimates without complex Bayesian methods.

- Shows that the approach can achieve comparable or improved performance with fewer labeled images, such as a 1.64% test error on MNIST with just 1000 labeled images.
- Validates the method's effectiveness in skin cancer diagnosis, demonstrating its practical use with a minimal number of labeled samples.

4. **Downsides**
    - MC dropout requires multiple stochastic forward passes during testing, increasing computational demands. For example, training a melanoma diagnosis model took 20 hours per run.
    - The model is reset to its original pre-trained state after each acquisition, which can extend the training time. This resetting might be unnecessary in some cases and could be optimized.
    - The method depends on CNNs and may not be effective for other data types or models, such as text or time series.
    - Acquisition functions have only been tested on a limited number of datasets. It's important to assess their performance across a broader range of tasks and models.

Paper II: Deep Active Learning For Named Entity Recognition

1. **Motivation**

    Deep learning models heavily rely on large labeled datasets. The paper focuses on addressing this issue for NER (Named Entity Recognition). Generally deep learning models outperform traditional methods on large datasets but their performance shrinks when training data is limited. This is a major issue for NER tasks. These tasks need explicit labeling by experts which can be laborious and expensive. The key challenges are:
    - Data shortage: NER tasks need esoteric knowledge that makes labeling data costly.
    - High costs: Active learning when applied to Deep Learning could be expensive as it requires iterative retraining.

    Thus, the main motivation of this paper is to make deep learning models more efficient by reducing their dependency on large labeled datasets. Combining deep learning with active learning is suggested by authors to evidently reduce the amount of labeled data required while still achieving state of the art performance.

2. **Solution**
    - Character level Encoder: CNN is the first component that processes character-level information. An encoder is used to extract features from each word.
    - Word level Encoder: Another encoder is used to capture the meaning of each word in a sentence. compared to RNN's that are slower CNNs are chosen for their computational efficiency.
    - LSTM Tag Decoder: After the NN is used for encoding the input, an LSTM network is used for assigning labels to words in a sequence. To understand the dependencies of sequences in NER an LSTM can be used.

- For sequence tagging, an LSTM combined with CNNs at the word and character levels offers a useful balance between speed and accuracy. Active learning scenarios involving regular model retraining require an architecture that is both lightweight and fast, while still producing results that are on par with state-of-the-art.
- Using an active learning strategy reduces the amount of heavily labeled data that is required. Selecting the most informative examples for labeling rather than randomly categorizing data is the fundamental concept of active learning. The following active learning strategies are examined in the paper:
    - Sentences that the model is least confident in their predictions for are given priority in samples (sentences) using the uncertainty sampling approach. These ambiguous samples are thought to offer the most useful data for enhancing the model.
    - Heuristic Approach: Using the normalized log-likelihood of the model's current predictions, the authors give a straightforward heuristic. Since it only needs to make one forward pass through the data, this method is computationally efficient because the model only chooses words with the lowest average confidence ratings.

- The model present in this paper is updated by the authors via incremental training, which combines freshly labeled samples with previously labeled data. The active learning process computing overhead is greatly decreased.

3. **Novelties**
    - CNN-CNN-LSTM Architecture: The study suggests a lightweight model architecture for neural network encoding (NER) that mixes convolutional networks for word- and character-level encoding with an LSTM tag decoder in between. With its computationally efficient design, this architecture performs on par with larger models. Particularly on larger datasets like OntoNotes-5.0, it is better than many other intricate architectures like CNN-LSTM-CRF (Conditional Random Fields) in terms of speed while maintaining competitive accuracy.
    - Active Learning for NER: Specifically for the NER challenge, deep active learning is applied in this paper. This work is one of the first to apply active learning to sequence tagging in NLP, as previous work mostly concentrated on image classification problems.
    - Heuristic-Based Active Learning: "Uncertainty sampling technique for active learning" which is a computationally efficient heuristic-based method was suggested in this paper. Their approach is faster while maintaining comparable performance because it only requires one forward pass, in contrast to Bayesian active learning methods that call for numerous passes.
    - Less Labeling Work: One important empirical addition is the proof that, with active learning, cutting-edge NER performance can be attained with only 25–30% of the initial training set. For example, they match 24.9% of the labeled data in the OntoNotes-5.0 English dataset with 99% of the performance of the best model.

4. **Downsides**
    - Incremental retraining doesn't significantly reduce costs and can cause delays with large datasets or complex models.
    - Heuristic Limitations: The heuristic for uncertainty sampling may not generalize beyond specific tasks, and more advanced methods like Bayesian approaches could be more effective but are computationally intensive.
    - Limited Advanced Methods: The paper does not explore advanced techniques such as Bayesian active learning by disagreement (BALD) in depth, despite their potential for better performance.
    - Outlier Over-Sampling: The uncertainty sampling method might lead to focusing too much on outliers, with only limited mitigation offered.

Paper III: Consistency-Based Semi-supervised Active Learning: Towards Minimizing Labeling Cost

1. **Motivation**

    Making the process of labeling data for machine learning cheaper and simpler. This is especially tricky for things like image classification, where getting enough labeled data is both hard and costly. Our aim is to make active learning (AL) work better by using some cool techniques from semi-supervised learning (SSL).

    Motivation: Getting labeled data for supervised learning takes a lot of time and money. Active learning tries to help by focusing only on the most useful data points for labeling. However, the usual methods have their downsides. For instance, they often don't do well at the start when there's barely any labeled data (the cold start problem) and sometimes don't pick the best samples. Using SSL can fix these problems, guiding active learning to make smarter and more efficient choices.

2. **Solution**
    - Consistency-Based Active Learning: How consistent a model's predictions are on slightly changed data is used to decide which samples to label. This is different from the usual methods that focus on how uncertain the model is (like using entropy).
    - Combining AL and SSL: Active learning (AL) was merged with semi-supervised learning (SSL) into one system. Using SSL's consistency loss, makes the process of picking which data points to label much more efficient.
    - Solving the Cold Start Problem: A major benefit here is that it helps the model perform better right from the start. Traditional active learning often struggles when there's little labeled data, but adding SSL gives it a better start.
    - Experiments and Results: This method was tested on popular datasets like CIFAR-10, CIFAR-100, and ImageNet. The results showed that this approach does a better job at getting value from fewer labeled data points compared to traditional active learning methods.

3. **Novelties**
   - A Combined Approach: The paper presents a framework that effectively merges semi-supervised learning (SSL) and active learning (AL). This method allows the model to leverage both labeled and unlabeled data, enhancing efficiency and reducing the dependency on large amounts of labeled data.
   - New Data Labeling Strategy: The authors introduce an innovative technique for choosing which unlabeled data should be labeled. They use a consistency-based approach where the model targets data with unstable predictions after minor changes, ensuring that only the genuinely challenging data is prioritized.
   - Tackling Early Challenges: The paper also addresses the cold start problem in active learning, where a model may struggle with a very small initial set of labeled data. They propose a method to determine the optimal amount of labeled data needed initially to give the model a stronger start.

4. **Downsides**
   - Dependence on Augmentations: This method relies heavily on how well the data is augmented. If the augmentations aren't done carefully, they can create misleading results and affect which samples get labeled.
   - Computational Complexity: The approach needs multiple passes for each sample because of the augmentations, making it more computationally demanding compared to traditional active learning, which only needs one pass per sample.
   - Applicability to Other Domains: While this method works great with image data, it might not be as effective for other types of data, like text or time-series, where useful augmentations are harder to come by.
   - Uncertainty Not Fully Addressed: The focus is on prediction inconsistencies, but it doesn't tackle other types of uncertainty, like model-based uncertainty, which might also be helpful in choosing the best samples.