IRE Assignment 3

Snehal Kumar
2019101003

Q1. What is the problem that the authors are trying to solve and why is it significant?

The main problem that the authors are trying to solve is the detection of whether a search query is well-formed or not. This is an important problem as the well-formedness of a query significantly impacts the efficiency of the processing and understanding of the query. The rise in lack of well-formed queries due to rise in verbose queries cause issues as they are usually lack proper structure and grammar, making the NLP tasks more difficult. Formally: Given a query, they learn whether a query is a well-formed natural language question or not through binary classification for modelling the task. There have been several attempts to detect well-formed questions such as word and Part-of-Speech n-grams with neural networks and using grammar parsing. The authors focus on checking the well-formedness of the web search queries.

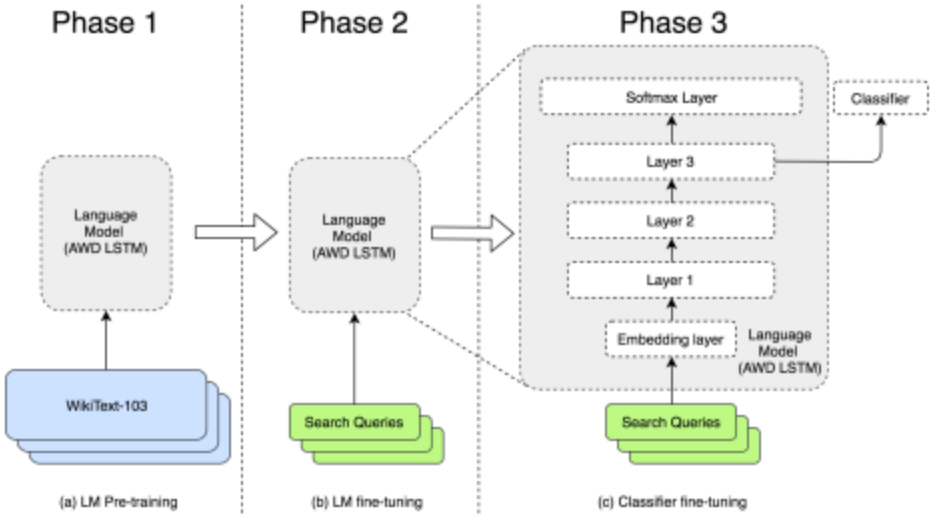
Q2. What is transfer learning?

Transfer learning is a technique which seeks to improve the training phase of a machine learning model. Similar to how humans use the knowledge of solving one task and transfer it to solve related tasks, transfer learning aims to use some knowledge of a particular task to solve other tasks. The knowledge gained while solving one problem is stored and applied to a different but related problem. By storing the weights, features etc, we need lesser amount of data to train for a new problem as this information can be leveraged by the model to speed up the training phase. Here, the efficacy of learning of a new task is dependent on the tasks previously learned as it will determine the amount of data or time required to make an accurate model. This method is especially useful in training deep learning models as they require huge datasets to train to get a reasonable accuracy.

Q3. Explain the architecture of the paper in your own words

They adapt the `ULMFiT` model in their approach for inductive transfer learning (ITL). The ITL mechanism uses `Averaged-SGD Weight-Dropped Long Short Term Memory` (AWD-LSTM)

network due to their efficiency with the same hyperparameters as that of an LSTM. The architecture of the paper is primarily divided into 3 phases.



- Phase 1: (LM Pre-training)

  In order for the model to learn the general domain, structure and language dependencies, a language model is taken and then pre-trained on a huge dataset, i.e `Wikitext-103` which consists of preprocessed Wikipedia articles.

- Phase 2: (LM Fine-tuning)

  Since the language model is now trained on the general domain, they use the task-specific data to fine-tune the model for the distribution specified. For fine-tuning, they use `discriminative fine-tuning` (DFT), which uses different learning rates for each layer, and `slanted triangular learning rates` (STLR), which does not maintain a constant learning rate for the layer and instead increases the rate and gradually linearly reduces it with the rise in training samples, approaches. This is done to avoid the issue of catastrophic forgetting tendency in language models.

- Phase 3: (Classifier Fine-tuning)

  In this phase they take the model and add two fully connected layers at the end of the architecture. These layers make the final classification, while the last softmax layer gives the rating prediction. They use the `gradual unfreezing heuristic` for fine-tuning the classifer which freezes certain layers to prevent simultaneous fine-tuning

of the layers. It starts from the last layer and unfreezes the previous layer after the layer is fine-tuned. This approach also prevents catastrophic forgetting in the model.

Q4. Discuss and describe the dataset in not more than one paragraph

The dataset consists of 25,100 queries, collected from users on WikiAnswers, annotated with human ratings between 0 and 1 of whether they are a well-formed natural language question. The tab separated values contain well-formed questions along with constructs of search queries for annotation. Each query is annotated by five raters with binary ratings after which the final rating is calculated as the average of the five binary ratings. A query is considered as well formed if its final rating is greater than or equal to 0.8.

Column	Content
1	The European Union includes how many ?
2	0.2

Column 1: Query, Column 2: Final Rating

Q5. Describe the results and takeaways from the paper

The final result shows that the implemented ITL model had an improvement of the previous SoTA from 70 to 75.03% accuracy. For further analysis, they assess the perfomance and effect of each phase which overall show improvements over the previous models. From the result, they have proved that using inductive transfer learning by fine-tuning the model does indeed help in detecting well formed queries and that the new model outperforms the baseline by a significant margin. This methodology of transfer learning shows its efficiency in this paper and that this technique can be applied to other languages as well with limited resources and dataset.

Q6. Coding Q: BERT fine-tuning

The code can be found in the code.ipynb file. Results after running BERT:

	precision	recall	f1-score	support
0	0.92	0.78	0.84	2369
1	0.71	0.89	0.79	1480
accuracy			0.82	3849
macro avg	0.81	0.83	0.82	3849
weighted avg	0.84	0.82	0.82	3849

The given dataset was used in fine-tuning the model which was taken from huggingface transformers. The classification report shows the accuracy metrics for both classes (0-not well formed, 1-well formed). Both having f1-score higher than 75.03% of the paper shows significant improvement in the binary classification task.

The results show that the model trained using BERT performs better than the ITL model implemented in the paper.

Q7. Using BERT to predict query well formedness rating

The architecture of binary classification of well formed queries can be extended to instead predict the ratings of the well formedness of the queries. This can be done by using a multi-class classification in the final layer with 6 classes from 0 to 1 in steps of 0.2. Once we have divided the final layer into the 6 nodes indicating 6 classes, we can include the distance between the classes as a feature in the loss function. The class with the maximum probability output from the softmax layer will then be chosen as the final rating prediction of the query.