EECS 6322 Week 7, Paper 1
Improving Language Understanding by Generative Pre-Training

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a semi-supervised learning approach for improving natural language understanding using generative pre-training and discriminative fine tuning. The generative pre-training makes use of the unlabelled data to tune the parameters of the model which perform much better than random initializations. This unlabelled data is currently present in abundance, especially in language processing, from sources such as wikipedia, novels, articles, etc, which enables the model to capture the general task-related information and helps it perform better. Then, the model parameters are fine-tuned during back propagation using task-specific labelled data, which is usually sparsely available due to the manual labelling requirements. Thus, the model benefits from large unlabelled data and hence improves the accuracy. The pre-training phase of the model is unsupervised and the fine-tuning phase of the model is supervised.

The architecture proposed by the authors uses a Transformer due to its ability to handle long range dependencies. It uses a new variant of the original transformer architecture known as the Transformer Decoder that applies multi-headed self-attention operations. The training of the model is done in two phases. The first phase is the unsupervised pre-training and the second phase is the supervised fine-tuning. In the first pre-training phase, the objective function used is given by $L1(U) = \sum_i logP(u_i|u_{i-k}, \cdots, u_{i-1}; \Theta)$*, where* Θ *represents the network parameters,* P *is the conditional probability, and* k *is the size of the context window. The output distribution of this pre-training phase is given by* $h_0 = UW_e + W_p$*,* $h_l = transformer\_block(h_{l-1})\ \forall i\ \epsilon\ [1, n]$*, and* $P(u) = softmax(h_n W_e^T)$*, where* n *is the number of layers,* W_p *are the positional embeddings,* U *is the context vector tokens, and* W_e *is the token embedding matrix.*

In the second fine-tuning phase, the parameters are adapted according to the specific tasks. For a labelled dataset C*, with input tokens* $x^1, \cdots, x^m$ *and labels* y*, the objective function to maximize is given by* $L2(C) = \sum_{(x,y)} logP(y|x^1, \cdots, x^m)$*. Additionally, to improve generalization of the model and to help the model converge faster, the objective function is modified to an auxiliary objective given by* $L3(C) = L2(C) + \lambda * L1(C)$*. The input transformations for some specific tasks that have structured inputs are pre-processed and separated by delimiters before feeding into the model since the pre-trained model is trained on continuous data. Hence, for tasks like Textual Entailment, Similarity, and Question Answering/ Commonsense Reasoning, the inputs have been separated in a specific order separated by delimiters in cases where the order is specified. In other cases, where the order of the sentence is not specified, all combinations of the sentence orderings are fed to the system.*

The model has been trained on the BooksCorpus dataset which contains over 7,000 unique unpublished books in various different genres. Its pre-training specifications include a 12-layer decoder-only transformer with masked self-attention heads, the Adam optimizer and a max learning rate of $2.5e{-}4$ *gradually increased from* 0*. Layer normalization is used in the architecture with a weight initialization of* $N(0, 0.02)$*. The Gaussian Error Linear Unit (GELU) activation function was used along with a L2 regularization with* $w = 0.0$ *on all gain weights/ non-bias weights. For fine tuning, the model reuses the hyperparameters from the pre-training phase, with an added dropout rate of* 0.1*. Additionally, a batch size of 32,* $\lambda = 5$ *and a learning rate of* $6.25e{-}5$ *was used in this phase.*

The model has been tested on several tasks along with their specific datasets. For the Natural Language Inference task, the model had to predict the relationship between a pair of sentences and classify them as 'entailment', 'contradiction', or 'neutral'. The proposed model was compared with the current state of the art models on six datasets namely MNLI-m, MNLI-mm, SNLI, SciTail, QNLI, and RTE where the proposed model outperformed its counterparts on all five datasets except the RTE dataset. In the Question Answering/ Commonsense Reasoning task, the model had to output the correct answer from English passages associated with questions. The proposed model again provided a better accuracy against all of its state of the art counterparts. Similarly, in the Semantic Similarity task the model had to identify if the sentences were semantically similar and in the Classification tasks the model had to classify if the sentence was grammatically correct or not. In both tasks, the proposed model had a greater accuracy against the other state of the art models in four out of six datasets and comparable accuracies in the other two datasets for both the tasks combined. This proves that the proposed model is able to perform better over a wider range of applications as compared to other architectures which are designed specifically for a given application.

Finally, the effect of the number of layers transferred from the pre-trained phase was also studied and the authors found that an increase in the number of layers transferred resulted in a better accuracy. Hence, this study concludes that the pre-trained phase helps the model capture useful task-related information and contains useful functionality for solving the specific target tasks. An ablation study on the model was also conducted where different phases of the models were skipped to measure their relative accuracies. It was found that the full model (Transformer with auxiliary LM) performed better on larger datasets, whereas the model without the auxiliary LM performed slightly better on the smaller datasets.

II. Strengths

The framework presented by the authors uses a universal representation that transfers information over a wide range of tasks with little adaptation according to the specific task. The model was also able to outperform the state of the art models in different applications, concluding that the proposed model not only generalized better due to the pre-training phase, but was also able to make better decisions, hence increasing the accuracy of the model on most datasets. The transformer used in the architecture of the model also enables the model to capture the long-range dependencies in the data, hence improving its decision making abilities.

III. Weaknesses

On the RTE dataset, a multi-task biLSTM model outperformed the proposed model by a significant margin, hence concluding that the proposed model does not perform very well with multi-task training. The current model also does not perform very well on datasets with short term dependencies due to the use of a transformer in its architecture. Although this does not prove as a problem for a majority of the tasks since they require long-range dependencies, for the few applications that require to capture the short-range dependencies, this model would not be the best choice.

IV. Correctness

The claims and the empirical methodology of the authors are correct since the model combines parts of the existing architecture to enhance the model performance on language processing. Compared to its prior work, the architecture comprises a transformer that helps the model to capture long-range information in the data hence significantly increasing its accuracy. The correctness of the methods can also be seen by the results compared by the authors against the current state of the art architectures, where the model outperforms its counterparts in nine out of twelve datasets.

V. Clarity

The paper is clear and well written. Each phase of model training, along with its objective functions have been clearly mentioned, along with its theoretical grounding. The paper also includes proper tables and figures required for the architecture and results which supplement the comparison and understanding of the reader.

VI. Relation to prior work

In semi-supervised learning for the natural language processing domain, previous work include models that use small range dependencies such as phrase-level or word-level structures that are then used in a supervised model. The model proposed by the authors of this paper however aims to capture the higher-level or the long-range dependencies of the data.

In unsupervised pre-training, previous work on natural language processing included models that performed unsupervised pre-training and supervised fine-tuning according to the specific task. However, the earlier models used LSTM architectures that still captured the short-term dependencies of the data. In contrast, the model proposed by the paper uses a transformer in its architecture that enables the model to capture the long-range dependencies and improves its efficiency over a wider range of tasks.

A few other approaches used earlier also used auxiliary features from the hidden representations of the pre-trained model during the fine-tuning phase of the model. This increased the number of new parameters for each target task, whereas the model proposed in the paper uses minimal changes in the model architecture during transfer, that results in the introduction of a very few parameters into the architecture.

VII. Reproducibility

The majority of the work proposed by the authors can be reproduced through the model specifications provided on the paper's github link. The code implementations, along with the architecture details in the paper can be modified according to the available resources and the majority of the results can be obtained or reproduced.

VIII. Additional comments

Under heading 4.2, the last sentence that mentions 'Figure 1' actually refers to 'Table 1' in the paper. Hence, the sentence should be " Table 1 provides an overview of all the tasks and datasets ".