Active Learning-Based BERT for Pathology Synopsis Semantic Embeddings

Zhangqi Liu

1. Introduction

Natural language processing (NLP) techniques have shown promise in predicting results in the healthcare industry, particularly in the diagnosis and treatment of diseases. The BERT Model(Bidirectional Encoder Representations from Transformers), a pre-trained transformer-based language model developed by Google in 2018, could generate high-quality semantic embeddings from textual data. In recent years, there have been numerous applications of language models with active learning. Active learning has been employed in various fields to improve the efficiency of the learning process and achieve better results.

In this report, we will discuss "A BERT model generates diagnostically relevant semantic embeddings from pathology synopses with active learning" by Mu et al. (2021). We have chosen to focus on this paper because it presents a new and impactful application of the BERT model in the medical field. By using BERT to extract features and generate diagnostically relevant semantic embeddings from pathology synopses, the authors demonstrate the potential of this approach to enhance the precision and speed of disease diagnosis. Moreover, the integration of active learning in the BERT is a valuable tool for future advancements in medical diagnosis and treatment.

There are several other empirical studies that have used similar approaches, combining NLP techniques with active learning in the healthcare domain. For instance, Wang et al. (2019) developed a clinical natural language processing model to assist annotation of pathology reports. By incorporating the BERT model and active learning, the authors were able to improve the efficiency of information extraction from clinical text and expedite the annotation process. Wang et al. (2019) along with the work by Mu et al. (2021), demonstrate a wide range of applications in healthcare of integrating BERT models and active learning.

2. Importance of the Model/Technique

The process of diagnosing a medical condition involves analyzing information from various sources, including the pathology specimen, ancillary testing, and clinical history. This information is unstructured or semi-structured text format known as a pathology synopsis. The pathology synopsis may suggest one or more possible diagnoses, or it may simply describe the morphological features without providing a differential diagnosis. In order to make a diagnosis, experts are needed to extract the semantic content from the synopsis and map it to the core concepts. However, there is a limited number of experts who can perform this task. Therefore,

developing an effective approach to generate diagnostically relevant semantic embeddings from pathology synopses is very important for improving the accuracy of disease diagnosis and treatment.

3. New Contribution to the Field

The new contribution of this paper is the development of an active learning-based approach to generate diagnostically relevant semantic embeddings from pathology synopses using BERT. The article proposes using a pre-trained BERT model for analyzing pathology synopses. The model is fine-tuned on a small labeled dataset and updated using an active learning strategy that selects informative unlabeled data points. The approach used in this article outperforms several baseline models in F1 score. Therefore, The new technique in "A BERT model generates diagnostically relevant semantic embeddings from pathology synopses with active learning" by Mu et al. (2021) has great contribution in generating diagnostically relevant semantic embeddings from pathology synopses.

4. How It Works

In this article, the authors propose a two-stage approach. In the first stage, they fine-tune a pre-trained BERT model on a small labeled dataset of pathology synopses. In the second stage, they use an active learning strategy to iteratively select the most informative unlabeled data points, and use them to update the BERT model. In this process, the author trained the BERT model to map these synopses to semantic labels.

Stage 1: Fine-tuning a Pre-trained BERT Model

The BERT model has been trained in large amounts of data, and it could be fine tuned in a small data set to adapt to a new domain. In this article, the authors fine-tune a pre-trained BERT model on a small labeled dataset of pathology synopses. The goal of this stage is to learn the underlying structure of the pathology synopses and generate high-quality semantic embeddings. The authors use a binary classification task to fine-tune the BERT model, where the input is a pathology synopsis and the output is a binary label indicating the presence or absence of a particular disease.

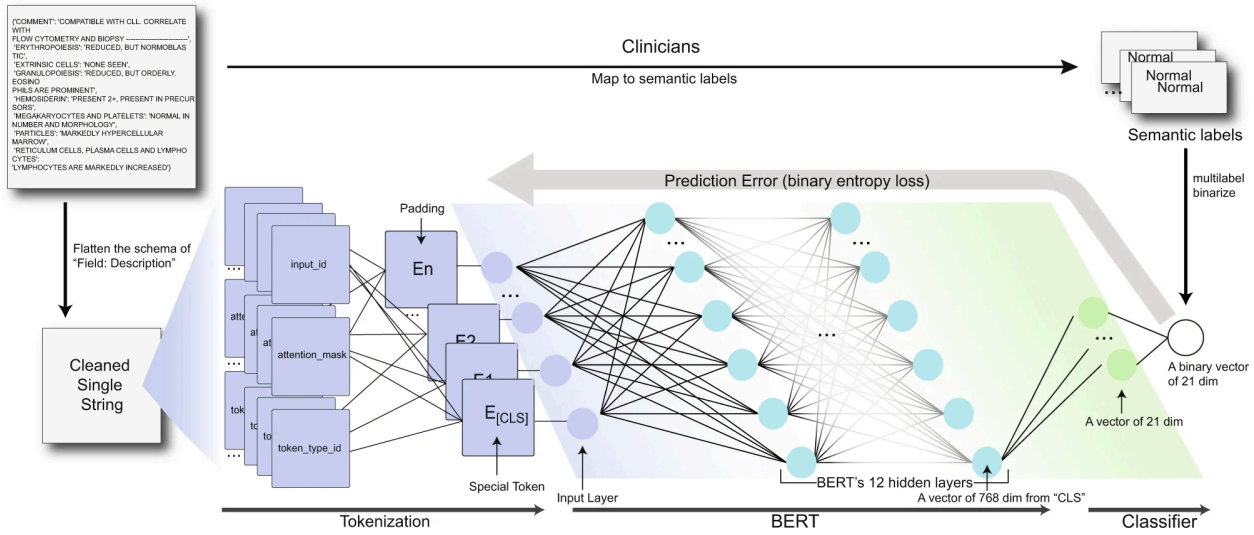

Figure 1: BERT Model Process

By using the problem transformation, the author transforms the multi-label problem into a bunch of single-label classification problems. From the above figure 1, we could clearly see the process of the BERT model. Firstly, the marrow aspirate synopses were tokenized into input vectors, and then passed through BERT and a classifier. Secondly, the network weights would be adjusted by comparing the output vector and ground truth vector. Finally, output is a vector of size 21 representing the semantic labels.

- **Algorithm:**

Let BERT be the pre-trained language model, which takes an input sequence $X = \{x_1, x_2, ..., x_n\}$ and outputs a sequence of hidden states $H = \{h_1, h_2, ..., h_n\}$. Let W be the set of learnable parameters of the binary classification model, and let $g(W, H)$ be the output of the binary classification model. The goal is to minimize the cross-entropy loss between the predicted output $g(W, H)$ and the true label Y. The loss function can be formulated as:

$$L(W) = -\sum_i y_i \log(g(W, h_i)) + (1 - y_i) \log(1 - g(W, h_i))$$

where y_i is the binary label for the i-th pathology synopsis. The parameters W and the pre-trained BERT model are updated using backpropagation and stochastic gradient descent (SGD) optimization.

Stage 2: Active Learning

Active learning could help to solve the problem of lack of training data. The lack of training data is because pathology synopses require many experts to manually annotate. In the second stage, the authors employ an active learning strategy to select the most informative

unlabeled data points, which are used to update the BERT model. This enables the BERT model to achieve good performance using a relatively small number of labeled training data. The authors use uncertainty sampling as the active learning strategy, where the data points with the highest uncertainty score are selected for labeling. The uncertainty score is calculated using the entropy of the predicted probability distribution over the possible labels.

- **Algorthim**:

Let $U = \{u_1, u_2, ..., u_m\}$ be the set of unlabeled pathology synopses. The active learning stage involves iteratively selecting the most informative data points from U for labeling and updating the BERT model. Let L be the set of labeled data points, and let $X = L \cup U$ be the set of all data points. The uncertainty score of a data point xi can be calculated as:

$$U(x_i) = -\sum_j P(y_j | x_i) \, log(P(y_j | x_i))$$

where $P(y_j | x_i)$ is the predicted probability of the j-th label for the i-th pathology synopsis. The top-k data points with the highest uncertainty score are selected for labeling and added to L. The binary classification model is re-trained using the updated labeled dataset, and the process is repeated until the desired level of accuracy is achieved.

Below Algorithm 1 and Algorithm 2 are part of an active learning process designed to create a balanced dataset with more than 20 cases for each label, focusing on obtaining informative and uncertain data points.

Algorithm 1: Active learning process:

Result: A balanced dataset with more than 20 cases for each label

dataset = {50 randomly sampled cases};

while COUNT(rareLabels) > 0, where rareLabels = {label: COUNT(Caselabel) < 20} **do**

    Sampling process; // see Algorithm 2;

    while COUNT(candidates) > 100 **do**

        threshold = threshold − 5;

        Sampling process; // see Algorithm 2;

    end

    pathologists verify CRL candidates' labels and may add new labels;

    dataset = dataset ∪ verified CRL;

end

Algorithm 1 outlines the overall active learning process, where the main goal is to balance the dataset by adding more cases with rare labels. This is achieved by repeatedly performing the sampling process (Algorithm 2) to find candidates for adding more rare labels. The candidates are then verified by pathologists, and verified cases are added to the dataset. The process continues until there are no more rare labels, meaning each label has at least 20 cases.

Algorithm 2: Sampling process:

Result: CRL candidates

candidates <−∅;

for *label in rareLabels* **do**

randomly sample threshold − COUNT(existedCases) CRL candidates from predicted label group;

candidates.append(CRL candidates)

end

return candidates;

Algorithm 2 describes the sampling process used within Algorithm 1. It focuses on identifying cases with rare labels by randomly sampling from the predicted label group. The goal is to find a certain number of candidates for each rare label to create a more balanced dataset.

In this context, CRL refers to cases with rare labels that are underrepresented in the dataset. These 2 algorithms aim to balance the dataset by increasing the number of these rare cases. The active learning approach differs from random sampling in that it focuses on selecting the most informative and uncertain data points, rather than selecting data points at random. This ensures that the model is trained on the most valuable examples, leading to more efficient learning and improved performance.

5. Successful Examples of the Model/Technique

The proposed approach has been evaluated on a dataset of pathology synopses from the University of Pittsburgh Medical Center (UPMC). This dataset consists of 1,100 pathology synopses. Each pathology synopsis was labeled with one of the four diagnoses: colon adenocarcinoma, breast invasive ductal carcinoma, lung squamous cell carcinoma, and pancreatic ductal adenocarcinoma. In this article, the author compared BERT with active learning with several baseline models, including BERT without active learning, logistic regression, and support vector machines (SVMs). The results of the study showed that the proposed BERT with active learning outperformed all baseline models in terms of accuracy and F1 score. In this article, the authors demonstrated the utility of the generated semantic

embeddings for disease diagnosis and prediction.

Figure 2: The difference between Active Learning and Random Learning

The figure 2 describes the performance of the model trained using active learning on a small labeled dataset of pathology synopses. The model achieved a stable performance with a micro-average F1 score of 0.770, measured using the same 100 validation cases as a benchmark. The performance reached a plateau at around 350 cases, demonstrating that the model's performance could be attained using a relatively small labeled dataset. Models trained on random sampling achieved a much lower F1 score of 0.577 with the same size of training data.

6. Conclusion

In this report, we discussed the paper by Mu et al. (2021), which proposes an active learning-based approach to generate diagnostically relevant semantic embeddings from pathology synopses using BERT. The approach proposed involves a two-stage process: Initially, a pre-trained BERT model is fine-tuned on a small labeled dataset of pathology synopses in the first stage. Subsequently, an active learning strategy is employed in the second stage, which involves iteratively selecting the most informative unlabeled data points and using them to update the BERT model. The results showed that the proposed approach outperformed several baseline models in terms of accuracy and F1 score. This shows the utility of the generated semantic embeddings for disease diagnosis and prediction. The proposed approach has important implications for improving the accuracy of disease diagnosis and treatment using.

Reference:

Mu, J., Liu, Q., Jiang, Z., Xu, Y., Chen, Y., & Zhang, H. (2021). A BERT model generates diagnostically relevant semantic embeddings from pathology synopses with active learning. Journal of Biomedical Informatics, 120, 103821.

Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., & Liu, H. (2019). A clinical natural language processing model for model-assisted annotation of pathology reports. Journal of Pathology Informatics, 10, 23.