A Mixture-of-Experts Deep Learning Model for Fake News Detection

Farzana Yasmin Ahmad, Zeyu Zhang, Shohaib Mahmud, Sudipta Saha Shubha
{fa7sa,qxc4fh,sm6re,ss7krd}@virginia.edu

1 Introduction

With the ever-increasing popularity of online social media, people are now more inclined to express their thoughts and feelings freely. On one hand, the right to free expression has resulted in a massive increase in internet information that, when mined methodically, can be utilized for citizen journalism and public awareness. On the other hand, its abuse has led to an alarming increase in online fake news. The spread and repercussions of fake news are becoming more severe due to insufficient fact-checking or third-party filtering in online social media.

According to a Pew Research Center survey, 50% of Americans consider fake news to be a major issue, ranking it above violent crime (Nakamura et al., 2019). Furthermore, the report found that 68% of Americans believe fake news has a substantial impact on their faith in the government and 54% believe it has a significant impact on their trust in one another (Nakamura et al., 2019). Hence, extensive research on fake news detection is critical for society. Significant research effort has been put into automatically detecting fake news using deep learning (DL) models (Alam et al., 2021).

Recent developments in DL have demonstrated that when a DL model's number of parameters increases, its capacity typically improves, resulting in higher model accuracy. This fact has led researchers in the fields of computer vision and natural language processing to frequently use enormous models with billions of parameters. Although the accuracy of huge models has improved significantly, the cost of computing to train them is very expensive (Liu et al., 2023). This prevents huge models from being used more frequently (Liu et al., 2023).

To address this issue, sparsely activated models have recently been proposed, which adhere to the fundamental idea of using huge parameters while maintaining constant computing cost (Liu et al., 2023). The Mixture-of-Experts (MoE) structure is now one of the most common approaches to perform sparse activation since it doesn't reduce model capacity as quantization and knowledge distillation do (Liu et al., 2023). An MoE model purposefully chooses just a handful of the parameters for computation for each input rather than activating all of them (i.e., all of the experts, which are often Feed-Forward Networks). Due to this, FLOPS require a sub-linear scaling as the model size increases. The potential of sparse MoE models over their dense counterparts has been demonstrated in recent literature (Liu et al., 2023).

In this project, we aim to investigate how accurately MoE model detects fake news compared to other models (e.g., BERT) that have been widely used for fake news detection. A major challenge for this project is to train and fine-tune an MoE model for the domain of fake news detection, which has not yet been explored much in existing literature.

2 Dataset Description

For this project we are going to work on two existing datasets. We are giving a brief description of these datasets in the following subsections.

2.1 Dataset 1

For the dataset 1, we are considering LIAR (Wang, 2017) which is a publicly available dataset for fake news detection. The sources of the instances of LIAR are political debate, TV ads, Facebook posts, Twitter post, interviews etc. It has 12.8k short statements in various context from POLITIFACT.com. They are all manually labeled. These short statements are labeled for truthfulness, subject, context/venue, speaker, state, party, and to-

tal credit history count. By credit history counts, we mean the historical counts of inaccurate statements for this speaker including the current statement. The truthfulness for each statement is evaluated by a POLITIFACT.com editor. There are six fine grained labels for the truthfulness. They are pants-fire, false, barely true, half true, mostly-true and true. The general statistics of LIAR is shown in the Table 1.

Train Set Size	10269
Test Set Size	1284
Validation Set Size	1283
Vocab Size	4358

Table 1: LIAR Dataset Statistics

Labels in the LIAR dataset is relatively well balanced. Except for the pants-fire label, the number of instances for all other labels range from 2063 to 2638. For pants-fire case there are only 1050 instances.

2.2 Dataset 2

We are also considering the Fake-and-Real-News (FARN) dataset from (Ahmed et al., 2018, 2017), where the fake news are collected from websites that Politifact (a fact-checking organization in the USA), and the real news are from the news website Reuters. The FARN dataset has 23481 and 21417 examples for fake and real news, respectively. We split the dataset into a training set, a validation set, and a test set with a ratio of 8:1:1 and build a vocabulary that has 66725 tokens based on the training set.

3 Proposed Method

The recent trend in machine learning has been "The bigger, the better". In the last three years, the size of the largest model trained increased 1000 folds. During this time, the number of parameters of trained models jumped from a few hundred millions to half a trillion parameters (Megatron-Turing NLG 530B). However, increasing the model size comes at the cost of increased computational cost. To combat this issue, mixture-of-experts (MoE) based transformer models have been proposed in recent time (Fedus et al., 2021). MoE based models allow us to increase the model size without incurring linear increase in computational cost. Motivated by the successes of the MoE-based model, we will explore the potentials of the MoE-based transformer model in solving fake news detection problem. Given the scope of this project, we tentatively plan to utilize the MoE model called SwitchTransformer (Fedus et al., 2021) with 8 experts per layer. There are 12 MoE layers in the model, equally divided among the encoder and decoder. There are approximately 0.6 billion parameters in the model. We will use a tokenizer pre-trained on C4 dataset (Raffel et al., 2019) which has about 750GB of English-language text sourced from the public Common Crawl web scrape. The tokenizer has approximately 32000 words in its vocabulary. We will train the model with MLP based classification layers on the aforementioned datasets. We will evaluate the model's performance in classifying fake news based on the accuracy and the training time metrics.

4 Implementation

We compared our SwitchTransformer-based classifier with the BERT-based classifier for fake news detection. This section elaborates on how we implemented the SwitchTransformer-based classifier and the BERT baseline.

4.1 Data Pre-processing

Since we used the pre-trained encoder module of both SwitchTransformer and BERT, we directly adopted their corresponding tokenizers without training our tokenizer. We used AutoTokenizer.from_pretrained("google/switch-base-8") for SwtichTransformer and BertTokenizer.from_pretrained("bert-base-uncased") for BERT, respectively, from the package transformers. The following depicts how we pre-processed the data on each dataset.

Liar. Liar is a 6-class classification dataset. Each example in Liar is a description sentence along with a label. The tokenizer encoded each sentence into a list of token IDs by tokenizing the sentence into words and converting words to their corresponding IDs. The maximum sequence length was set during this process, and the padding was enabled. If the length of a sentence was shorter than the maximum sequence length, it was padded with padding tokens; if the length was longer than the maximum sequence length, it was truncated. Finally, token IDs and labels were converted to PyTorch tensors. BERT's tokenizer additionally provided a mask tensor for each encoded

sentence, which served as one of the model inputs.

FARN. FARN is a 2-class classification dataset slightly different from Liar regarding the data format. Each example in FARN has a tile, a paragraph body, a date, and a label. We combined the title and the body into a string for each sentence, in which a space separated the title and the body. The date was discarded in our experiment. The same strategy (e.g., the padding and the truncation) was applied for FARN as Liar when encoding the example.

4.2 Vocabulary Size

The tokenizer of SwitchTransformer has 32100 tokens; the tokenizer of BERT has 30522 tokens.

4.3 Third-party Packages and Coding

We implemented all methods in PyTorch. In addition, some python packages were used in the implementation, e.g., datasets, numpy, pandas, torch.nn, torch.utils.data, and transformers. Specifically, we mainly adopted load_dataset from datasets, DataLoader and Dataset from torch.utils.data, and AutoTokenizer, SwitchTransformersForConditionalGeneration, BertModel, and BertTokenizer from transformers. The function load_dataset("liar") loads the Liar dataset.

We did not adopt other people's code besides the Python packages above. We independently completed the implementation.

4.4 GitHub Code Repository

We published our code on our GitHub repository (Authors). The URL is `https://github.com/Zeyu-ZEYU/UVA-NLP-2023`.

4.5 Prediction Model

The prediction model consists of two distinct parts: a transformer module that learns linguistic features and a classifier module that predicts fake news label based on the learned linguistic features. This project utilized a pretrained SwitchTransformer encoder module to extract the aforementioned linguistic features. For the baseline method, we used a pretrained BERT model. The SwitchTransformer encoder was loaded using SwitchTransformersForConditionalGeneration.from_pretrained("google/switch-base-8"), and the encoder of BERT was loaded from BertModel.from_pretrained("bert-base-uncased"). We removed the pooler layer from BERT and only adopted the last hidden state.

We attached a classifier layer to the SwitchTransformer Encoder module and the BERT Encoder module. The classifier accepts the last hidden state from the encoder as the input and consists of a Linear(max_seq_len * model_dim, 768), a Dropout(0.2), a ReLU(), and a Linear(768, num_classes), followed by a Softmax().

5 Experimental Evaluation

In this Section, we present the details on our experimental evaluation.

5.1 Experiment Setting

The experiment was performed on a CS department server equipped with 4 GPUs (NVIDIA GeForce RTX 2080 Ti). However, only one GPU was used. The batch size was 128, the learning rate was 1e-5 (using Adam optimizer), and we trained each model for 10 epochs. We fixed the parameters of the encoder module of each model and only trained the classifier's parameters.

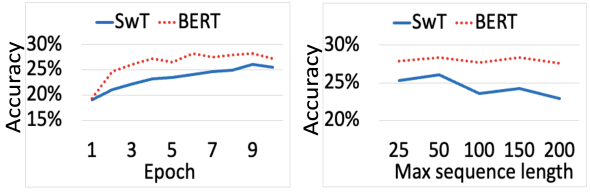
We performed the experiment on Liar with the maximum sequence lengths of 25, 50, 100, 150, and 200 and on FARN with the maximum sequence lengths of 50, 100, 200, and 300.

5.2 Experimental Results

Our key results include: (i) SwitchTransformer achieves comparable or slightly lower accuracy than BERT, however, (ii) SwitchTransformer requires significantly lower training time than BERT.

5.2.1 Liar dataset

Accuracy. Figure 1 shows the accuracy comparison between SwitchTransformer and BERT on the Liar dataset.



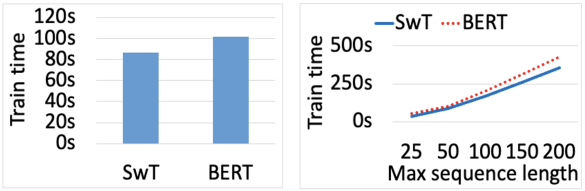
(a) Accuracy with maximum sequence length=50

(b) Accuracy with varying maximum sequence lengths

Figure 1: Accuracy comparison between SwitchTransformer (SwT) and BERT on Liar dataset.

We see that BERT achieves 2%-4% higher accuracy than SwitchTransformer. This is because BERT considers both forward and backward contexts in its prediction, whereas SwitchTransformer only looks into the backward context. This result shows the importance of considering both contexts during prediction.

Training Time. Figure 2 shows the training time comparison between SwitchTransformer and BERT on the Liar dataset.



(a) Training time with maximum sequence length=50

(b) Training time with varying maximum sequence lengths

Figure 2: Training time comparison between SwitchTransformer (SwT) and BERT on Liar dataset.

We see that SwitchTransformer achieves up to 28% lower training time than BERT. This is because the feed forward network is divided horizontally in SwitchTransformer and not all feed forward networks are activated for each input, thus achieving sparse activation and more parallelization. This result indicates that the sparsely activated models can reduce training cost even when the number of parameters are huge as discussed in Section 1.

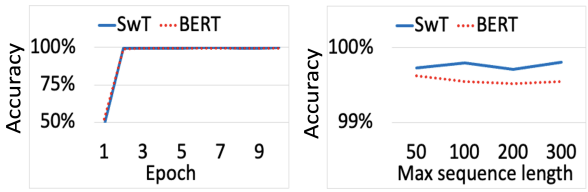
With the increase of the maximum sequence length, both methods show the trend of increasing training time as each method needs to process more tokens concurrently with the increase of maximum sequence length.

5.2.2 FARN dataset

Figure 3 and Figure 4 show the accuracy and training time comparisons between SwitchTransformer and BERT on the FARN dataset, respectively. In this dataset, SwitchTransformer achieves comparable accuracy with BERT. However, SwitchTransformer requires up to 50% less training time than BERT for the same reasons described above.

6 Conclusion

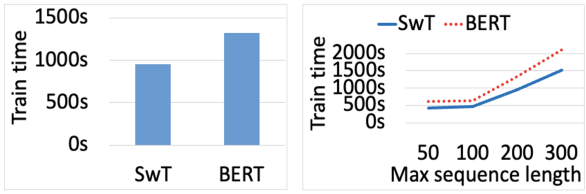
In this project, we have developed an MoE-based DL model for fake news detection. The model requires significantly lower training time than BERT with comparable or slightly lower accuracy. In the future, we will evaluate how increasing the number of parameters in the MoE model impacts the accuracy and training time.



(a) Accuracy with maximum sequence length=200

(b) Accuracy with varying maximum sequence lengths

Figure 3: Accuracy comparison between SwitchTransformer (SwT) and BERT on FARN dataset.



(a) Training time with maximum sequence length=200

(b) Training time with varying maximum sequence lengths

Figure 4: Training time comparison between SwitchTransformer (SwT) and BERT on FARN dataset.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *ISDDC*.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1).

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, et al. 2021. A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541*.

Authors. Source code for implementation.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961.

Juncai Liu, Jessie Hui Wang, and Yimin Jiang. 2023. Janus: A unified distributed training framework for sparse mixture-of-experts models. In *Proc. of SIGCOMM*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.