Choosing the Right Experts for Biomedical QA Tasks

Faizel Khan
University of Minnesota
khanx370@umn.edu

Abstract

Question Answering (QA) is a Natural Language Processing (NLP) task that allows models to predict answers to questions given question-answer pairs for a given context. Biomedical Question Answering (BioQA), as a QA sub-task, enables innovative applications to effectively comprehend complex biomedical knowledge. BioQA has been gaining attention in recent years due to its potential applications. However, it is a challenging task as biomedical questions are often complex and diverse. The biomedical questions of different types have varied focuses and need many problem-solving procedures. Generating large Question Answering (QA) models from the ground up for various biomedical datasets can be computationally expensive and raise an overfitting issue. The existing QA models use a single homogeneous model to answer all questions, which can lead to the model becoming confused when faced with different types of questions and consequently reducing performance. We introduce a Mixture of Expert (MoE) (Robert A. Jacobs and Hinton., 1991) based QA method, which aims to decouple the computation for different questions using the sparse gating method. To the best of our knowledge, our proposed QA-based MoE model is the first attempt to devise a BioQA system with a non-BioQA pretrained model. The results show that our MoE-based QA method performs almost similarly to other existing QA methods. During our analysis, we find that class imbalance impacts the overall performance, i.e., when the model overly relies on a few strong experts that tend to add noise rather than merit. This restricts us from increasing the number of experts in the model. We tried addressing this issue using loss functions sensitive to class imbalance, e.g., focal loss (Lin et al., 2017). Even after trying different loss functions, we could not find a better solution to handle the class imbalance problem. Therefore, we could not improve the performance of the MoE-based QA model.

1 Introduction

Over the past few years, Question Answering aspects of NLP models have drawn attention in the medical community for their prospective applications like nursing chatbots. This interest has led to an increase in many biomedical-specific large NLP models like BioBERT (Lee et al., 2019), and BioGPT (Luo et al., 2022). These models have attracted considerable attention with their significant performance improvement over the other generalized NLP models like BERT (Devlin et al., 2018). However, the biomedical-specific models have been trained on specific biomedical corpora that raise an issue of overfitting to the superficial correlation. In addition, these model generations require substantial computational resources; for example, BioBERT was generated after training the model for 23 days on eight NVIDIA V100 GPUs. This limits the interest of researchers with low resources to work in the biomedical domain.

BioQA dataset is challenging compared to other QA datasets because different biomedical questions have different emphases and need different problem-solving processes(Jin et al., 2021). The existing QA model, like DistilBERT (Sanh et al., 2019), uses a homogeneous model to answer biomedical questions. It is hard to determine the model parameters that are to be shared between different types of biomedical, and thus, it could negatively affect the model performance. Therefore, we introduce the MoE-based QA method, which will address the issue of the model parameter. This method aims to decouple the computation for different types of questions by the sparse gating method used in MoE. This method learns the routing strategy so that the questions will be grouped into clusters, and each expert tends to answer several types of questions it is expert in.

In this paper, we utilize the BioASQ dataset to train and evaluate this MoE-based model. While increasing the number of experts, we faced the challenge of class imbalance where the model overly relies on a few strong experts so we utilized different loss functions, like focal loss, to deal with a class imbalance when increasing the number of experts in this model. Focal loss significantly drops the training loss at the beginning of the training compared to cross-entropy loss and then comes close; however, performs poorly on model accuracy overall. In addition, we trained this MoE-based model with a mixed dataset from an out-of-domain dataset, SQUAD (Rajpurkar et al., 2018), which outperformed the existing baseline QA models. However, we do not have a better understanding of how this version of MoE works better than the MoE with only BioASQ. Therefore, we leave it as part of our future exploration.

2 Related Work

DistilBERT is a smaller, distilled version of the original BERT model. The authors of DistilBERT leverage knowledge distillation to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities while keeping it 60% faster. Since DistilBERT is smaller, faster and lighter model, it is cheaper to pre-train and useful for on-device computations.

A recent study, Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer (Shazeer et al., 2017), showed that MoE could be utilized in tasks with a large number of parameters, such as in language modeling tasks and introduced the Noisy Top-K gating consists of thousands of feed-forward sub-networks [Figure 1]. Using such gating methods could allow us to choose the optimal combination of different experts created from baseline models that are trained in different domains, including the biomedical domain.

In another BioQA study, MoE was utilized to create MoEBQA (Dai et al., 2022), a Multiple Choice QA method. This study uses a pretrained PubMedBERT (Gu et al., 2022) model. Their experiments show that their MoE extension significantly boosted the performance of their question-answering models and . Another study, Build a Robust QA system with a transformer-based Mixture of Experts (MoE) (Zhou et al., 2022), drew inspiration from Shazeer et al. (2017) to apply MoE on out-of-domain datasets. It uses DistilBERT as the baseline model and then trains different experts on out-of-domain datasets like Relation Extraction (Levy and Specia, 2017), DuoRC (Saha et al., 2018), and RACE (Lai et al., 2017). It uses a combination of Data Augmentation and MoE techniques to adapt the baseline model to the out-of-domain datasets. This study showed a 9.52% performance gain over the baseline model. This is a reassurance that the MoE technique can be used to empower the generalized model on out-of-domain topics. We will utilize some of their training methods to optimize our QA task on the biomedical dataset. We also draw inspiration from their code base to implement our models and apply them in the biomedical domain dataset.

3 Approach

We utilize MoE in our prediction model. The MoE consists of a number of experts and a trainable gating network. In this model, multiple expert learners are used to divide a problem into subproblems, with each expert tasked to excel at one of the subproblems. Then the gating network selects a sparse combination of experts to process each input.

We create a neural network with k number of experts at the beginning of the training. Each expert contains one hidden layer containing thousands of GELU-activated (Hendrycks and Gimpel, 2016) units. Then, we create different batches of data. Each expert layer receives a combined batch consisting of the relevant training examples from all the data-parallel replicas and outputs. We also create a gating network in parallel that connects to each of the expert layers at their output side. This gating network takes the same input given to the expert layers. The gating network uses a loss function to evaluate the contribution that each expert should have in the decision-making. Therefore, the gating network is a key instrument in this model to choose the experts that can be trusted to make a strong prediction.

For an input x, let us denote G(x) be the output of the gating network for expert i and $E_i(x)$ be the output for expert i. Then the output y of the MoE model can be defined as,

$$y = \sum_{i=1}^{n} G(x)_i E_i(x) \tag{1}$$

Traditionally, a softmax gating network is used for MoE, but we choose to adopt the Noisy Top-K gating network (Shazeer et al., 2017), defined as below, that adds sparsity and noise to the model through expert selection and noisy weight terms.

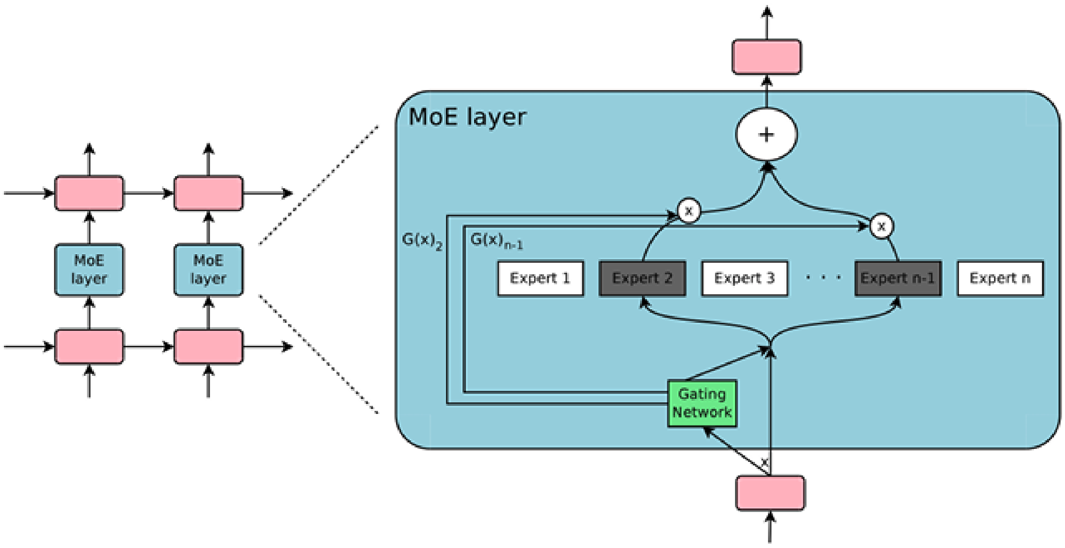

Figure 1: **The Sparsely-Gated Mixture-of-Experts architecture**

$$G(x) = Softmax(KeepTopK(H(x), k) \quad (2)$$

$$\begin{aligned} H(x)_i = &(x \cdot W_g)_i + \\ &StandardNormal() \cdot \\ &Softplus((x \cdot W_{noise})_i) \end{aligned} \quad (3)$$

$$KeepTopK(v, k)_i = \begin{cases} v_i, & \text{if } v_i \text{ is in the top k} \\ -\infty, & \text{otherwise} \end{cases} \quad (4)$$

4 Experimental Setup

4.1 Hardware

We setup our virtual environment in the cloud service provided my the Minnesota Supercomputing Institute (MSI) (of Minnesota Twin Cities) at the University of Minnesota - Twin Cities campus. We specifically utilize their "Agate A40 GPU session" which provides us with 1 AMD EPYC GPU, 16 cores CPU processor, 60 GB RAM and 80 GB local scratch. Although, it limits our session time to 4 hours at a time.

4.2 Baseline Model

We utilize a standard pre-trained DistilBERT model available through huggingface as our baseline model. We also fine-tuned it to average across the batch and used AdamW (Loshchilov and Hutter, 2017) optimizer to average the loss.

4.3 Hyperparameter

We used the number of experts from 1 to 7 and found that all of our model converges within 7 epochs. So, we use 5 epochs with a learning rate of 3e-5 and a batch size of 16 to train all of our models.

4.4 Evaluation Methods

To measure the performance of our model, we compare it with BioBERT and DistilBERT models. Exact match (EM) and F1-score will be used as evaluation methods. EM is a binary classification that returns 1 only if predicted and actual values are totally identical. For example, if the output is ‘Cancer’ but the labeled answer is ‘Kidney Cancer,’ the EM score will be zero. F1-score is a harmonic mean of two other evaluation methods: precision and recall. Precision measures the ability of a model to predict relevant points. While recall measures how well the model classifies all the relevant cases. The range of F1-score values is from 0 to 1. As bigger the F1-score the better performance of a model. For example, in the Cancer example, the precision will be 1 (as the output is subset of labeled answer) and the the recall will be 0.5 (as the output included one out of the two words in the labeled answers). This will give an F1 score of 0.67.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

Table 1: Results

Model		F1	EM
DistilBERT		51.9	35.1
Fine-tuned DistilBERT		85.28	78.97
BioBERT		88.72	84.33
MoE Based QA	1 Expert	83.99	78.66
	2 Experts	82.92	76.97
	3 Expert	83.79	77.6
	4 Experts	83.25	77.81
	5 Experts	84.35	78.34
	6 Experts	82.52	75.6
	7 Experts	82.45	75.92

4.5 Visualization

It is challenging to keep up with the results running in the cloud as we would have to export them from MSI to our storage and then compare them with the old models manually. So, we utilized an online visualizing platform, WandB (WandB), for our models. It allowed us to easily track our training loss and accuracy for each model and compare it between different models.

5 Results

The base model, DistilBERT, performed F1 of 51.9 and EM of 35.1. We finetuned it on BioASQ to perform F1 of 85.28 and EM of 78.97. Then we trained MoE models and looked at the number of experts from 1 to 7 as shown in Table 1. We also evaluated BioBERT on the same test set and got F1 of 88.72 and EM of 84.33.

6 Analysis and Discussion

We conduct a brief analysis of test results from the best-performing systems in this section.

- **Difficulty in maintaining context**: In many cases, the model tends to predict the answer with a longer context window than the provided answer (ground truth label), causing the EM score to drop, but not affecting the F1 score as much. For example,
  Question: Which ploidy-agnostic method has been developed for estimating telomere length from whole genome sequencing data?
  Context: (long context)
  Answer: Fidaxomicin

- **Number of Experts**: Our best MoE model with 3 experts performed F1 of 83.79 and EM of 77.6 Compared to the model with 1 expert, we observe that the model does better as we add more experts. However, as the number of experts increases, we think that the model overly relies on a few strong experts compared to the other experts that tend to add noise rather than merit, which reflects in steady performance after 5 experts. We tried to alleviate this problem using Focalloss, which is sensitive to class imbalance and discussed in-depth below.

- **FocalLoss effects on MoE Training**: We noticed that there was some class imbalance in our MoE model while using the Crossentropy Loss function. So, we introduced Focalloss, which tries to handle the class imbalance problem by assigning more weights to hard or easily misclassified examples and to down-weight easy examples. This significantly drops the training loss at the beginning of the training. However, the model performs poorly on F1 and EM [Figure 2]. This is why we had to continue using cross-entropy for the final version of the model.

- **Mixing with out-of-domain dataset**: We also trained an MoE model with a mixed dataset of BioASQ and SQUAD while training the MoE model and tested with only BioASQ. The resulting model outperformed the existing QA methods, F1 86.16 and EM 79.92. However, we do not have a better understanding of this version of MoE works better than the MoE with only BioASQ. We intend to look into variations of such mixtures of datasets in the future.

- **Comparing with other models**: We first wanted to compare our model with BioBERT using Mean Reciprocal Rank (MRR). It is a measure to evaluate systems that return a ranked list of answers to queries. However, we decided to implement BioBERT through the transformers library and evaluated it on our BioASQ dataset using the F1 and EM method(as shown in Table 1).

7 Conclusion

In this paper, we explore methods to handle diverse biomedical questions, devise ways to avoid the homogeneous nature of existing QA models, and re-

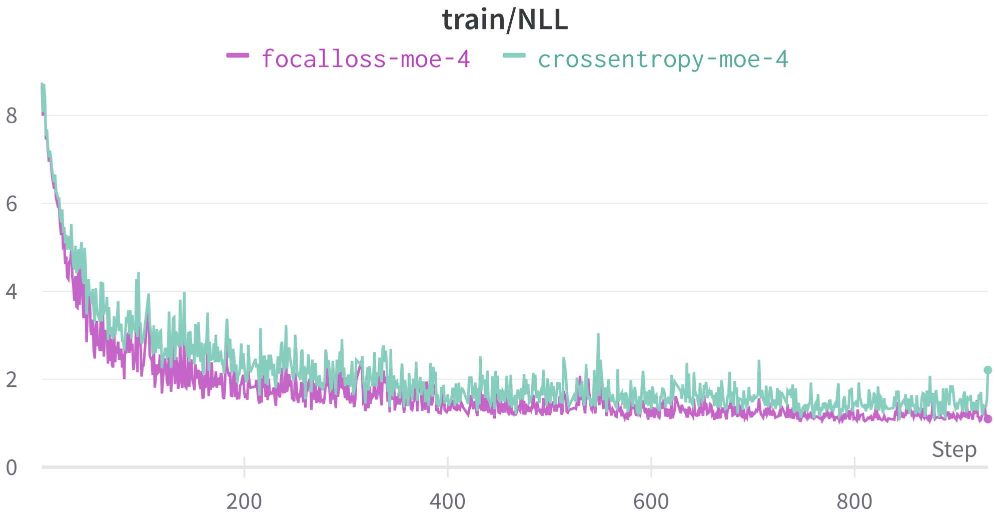

Figure 2: **Focal Training Loss Vs Cross-entropy Loss**

duce substantial computational resources employed to build models like BioBERT. In order to solve the aforementioned issues, we propose an MoE-based QA method using a non-BioQA pre-trained model, which aims to decouple the computation for different questions using the sparse gating method. We utilized BioASQ to train and evaluate this MoE-based QA method. The method, however, faces a class imbalance issue where a few strong experts are overly relied upon, which tends to add noise. We tried to alleviate this issue of class imbalance using focal loss but it did not perform any better than cross-entropy.

References

Damai Dai, Wenbin Jiang, Jiyuan Zhang, Weihua Peng, Yajuan Lyu, Zhifang Sui, Baobao Chang, and Yong Zhu. 2022. Mixture of experts for biomedical question answering.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus).

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2021. Biomedical question answering: A survey of approaches and challenges.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Roger Levy and Lucia Specia, editors. 2017. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*. Bbac409.

The University of Minnesota Twin Cities. The minnesota supercomputing institute (msi). `https://www.msi.umn.edu/`. Online; accessed 17 October 2022.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Steven J. Nowlan Robert A. Jacobs, Michael I. Jordan and Geoffrey E. Hinton. 1991. Adaptive mixture of local experts. *Neural computation 3, 79-87*.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.

WandB. Weights amp; biases – developer tools for ml. https://wandb.ai/site. Online; accessed 25 October 2022.

Yu Qing Zhou, Xixuan Julie Liu, and Yuanzhe Dong. 2022. Build a robust qa system with transformer-based mixture of experts.