CS224n - Building a Robust QA system

Yuanzhe Dong, Xixuan Julie Liu, Yu Qing Zhou
Computer Science Department, Stanford University

Summary

In this project, we build a robust question answering system that can adapt to out-of-domain datasets. We incorporate the Mixture-of-Experts (MoE) and Switch Transformer architectures to our model, and EDA and back translation for data augmentation. A combination of our best model architecture and techniques achieves 53.477 F1 and 37.696 EM in the out of domain evaluation.

Background

While a single network may overfit to the superficial distribution in the in-domain training data, with a meaningful number of expert sub-networks, a gating network that selects a sparse combination of experts for each input example, and careful balance on the importance of expert sub-networks, a Mixture-of-Experts (MoE) model [3] can train a robust learner that can be generalized to out-of-domain datasets. However, the paper [3] does not touch on how well MoE applies to the QA task.

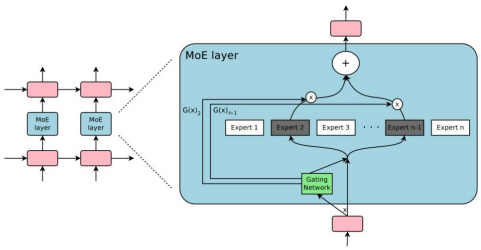

Figure 1: The architecture of Sparsely-gated Mixture-of-Experts Layer [3]

Inspired by the success of large-scale Transformer [4], while seeking greater computational efficiency, Switch Transformer [1] is proposed as a sparsely-activated expert model. It activates a subset of the neural network weights for each incoming example. Switch Transformer simplifies the MoE routing algorithm with reduced communication and computational costs.

In addition to novel architectures, data augmentation can also boost performance and robustness of training. Easy data augmentation (EDA) techniques [5], including synonym replacement, random deletion, random swap, and random insertion, have shown effectiveness on small datasets, despite their simplicity. Back translation is another technique that has also been shown to improve reading comprehension performance [6], thus gaining popularity.

Technical Methods

As figure 1 shows, after the output layer of the DistilBERT [2], we adds n single fully-connected layer in parallel as experts and another linear layer that serves as the gating function, before producing the final output. Given a input x, the output y of the model is $y = \sum_{i=1}^{n} G(x)_i E_i(x)$, where $G(x)_i$ is output of the gating function and and $E(x)$ is the output of ith expert network.

For the Switch Transformer, we bring the MoE layers up to the middle of the DistilBERT model [2] and replace the dense feed forward network with a sparsely-activated switch FFN layers, as figure 2 shows. Through testings, we find that 8 switch FFN layer work the best. We choose 8 switch FFN layer in the following experiments.

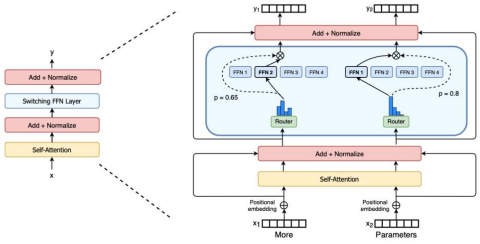

Figure 2: Switch transformer architecture [1]

For EDA, after data augmentation of each context paragraph, we rematch the answers within the augmented context. In order to reduce failure of rematch, we avoid operations on words within contexts that also appear in the answers.

Similarly, for back translation, we only translate context before and after answers. We use Google Translation API for its better speed and accuracy. We use Spanish, French, and German as intermediate languages.

We train our models for 5 epochs with a learning rate of 3e-5, and a batch size of 16. For each expert, we use hidden dimension of 3,072. We trained the DistilBERT Baseline only on the in-domain training data, same as the MoE with One-Expert-per-dataset (one expert per in-domain dataset), which uses out-of-domain data only for training the gating network. We use a mixture of in-domain and out-of-domain training data to train the Sparly gated MoE model and the Switch Transform. Data augmentation is only used on the out-of-domain training examples, as they are disproportionately small compared to in-domain datasets.

Experiments Analysis

Unexpectedly, comparison across different number of experts (figure 3) shows that models with 1 and 2 expert(s) perform the best. This is likely due to the lack of strength of our loss function $L_{importance}$ in equalizing importance of experts, which makes gating network tend to produce large weights for the same few experts.

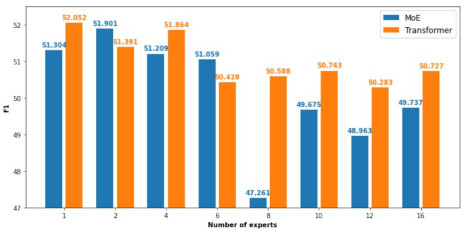

Figure 3: Performance of MoE vs Transformer, with different number of experts

We also explore training each expert within our MoE architecture with one of the three in-domain datasets, which does not yield as good performance as self-supervised experts trained on a mixture of training data (table 1).

Table 1: Ablation across model architectures and data augmentation (out-of-domain validation performance and improvement over baseline)

Treatment	Experiment	F1	Improvement
Baseline	DistilBERT Baseline	48.83	-
	DistilBERT +OOD	51.330	2.5
Explore MoE Architecture	One Expert per Dataset	47.096	-1.734
	Self-supervised Experts	51.901	3.071
	Switch Transformer	**52.052**	**3.222**
Data Augmentation (with Switch Transformer)	EDA	52.396	3.566
	Back translation	52.905	4.075
	EDA + back translation	**53.477**	**4.647**

To conclude, our combination of best architecture and data augmentation achieves a 53.477 F1 score, which is a 9.52% performance gain. We also succeed in demonstrating that MoE architecture can be applied to QA tasks.

References

[1] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[3] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[5] J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.

[6] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension, 2018.