Application of Mixture of Experts in Domain-agnostic Question Answering

Stanford CS224N {Default} Project

Jiayi Li
Department of Mathematics
Stanford University
jiayili@stanford.edu

Abstract

The goal of this project is to explore question answering models that are able to transfer what it learned in certain domains to unknown domains. Inspired by the intuition that given a particular target dataset, a specialized model trained on a similar dataset will outperform a general model, we employ a method to ensemble a mixture of dataset experts which turns out to be effective in improving both in-domain and out-of-domain predicition performance.

Mentor: Kaili Huang

1 Introduction

Many datasets have been created for training reading comprehension and question answering models. A natural question to ask is whether a model trained on a set of known datasets can perform reasonably well on unseen datasets. If we cannot achieve generalizable performance on out-of-domain datasets, we will be far from tackling the variety of questions an open-domain QA system can face. Generalizability is a challenging task as different datasets come from distinct domains and have distinct features. For example, the popular SQuAD dataset has an average context length of 120 while NewsQA, a question answering dataset based on CNN articles, has an average context length of over 700 words [1]. Other factors such as number of questions per paragraph, overlapping between questions and context, and common locations of answer spans also contribute to the large variance among datasets.

When designing models to utilize data from multiple sources, we need to balance between overfitting and underfitting. On one hand, we do not want to overfit to spurious features specific to the training datasets and learn fragile distribution that generalizes poorly to out-of-domain datasets. One approach is to combine data from several training datasets into a single larger domain in order to learn general patterns of question answering.

This strategy of using one feature space to characterize heterogeneous distributions, however, washes out useful characteristics of individual datasets. Intuitively, given a particular target dataset, a specialized model trained on a similar dataset will outperform a multi-dataset model. For example, in-domain dataset NewsQA and out-of-domain dataset DuoRC both have especially long paragraphs, so expert on NewsQA might be able to predict answers for paragraphs in DuoRC better than the general model. This inspires our approach to train several models each representing an expert for a specific dataset to improve transfer learning.

In addition, we discovered that ensembling experts trained separately on different datasets is not enough to beat the multi-dataset model. In fact, the multi-dataset knowledge is important because even experts learn better when it has access to more data. Therefore, we propose a method combining multi-dataset training with mixture of experts which outperforms multi-dataset model by 7% in F1 score on the out-of-domain validation set and the method works consistently with the scaling of model size.

2 Related Work

2.1 Mixture of experts

The idea of Mixture of Experts (MoE) can be dated back to [2]. In this original MoE paper, a single task is divided into subtasks, and each expert learns to handle a certain subtask. Guo et al. [3] introduced the mixture-of-experts approach for sentiment analysis and part-of-speech tagging tasks. In this project, we assume that each subtask corresponds to each domain in in-domain training set. Moreover, we assume that unseen domains can be inherently represented as a combination of several observed domains. Therefore, we expect that mixture of experts can deal with examples in any domain well.

2.2 Single dataset experts

[4] explicitly explores the idea of combining single dataset experts for multi-dataset question answering. This paper provides two important inspirations for our method. First, the paper shows that model trained directly on a single dataset performs worse than multi-dataset model on the particular dataset which also corresponds to our experiment findings. Second, the mixture of experts proves to be effective for zero-shot generalization that accords with our goal.

3 Approach

The objective of question answering is to model the distribution $p(a|q, c)$, where $q, c, a \in D$ represent a question, context, and answer respectively from a dataset D. In particular, we focus on extractive question answering where answers are selected as a span of tokens in the context. We make the standard assumption that the start indices are independent with end indices, i.e. $p(\text{span}(\text{start} = i, \text{end} = j)|q, c) = p(\text{start} = i|q, c) \cdot p(\text{end} = j|q, c)$. We have a collection of source datasets $D = \{D_1, D_2, ...D_k\}$.

The multi-dataset approach is to fit a single model to examples drawn uniformly from the dataset in D:

$$\arg\min_{\theta,\phi} \mathbb{E}_{D_i \in D}[\mathbb{E}_{q,c,a \in D_i}[-\log p_{\theta,\phi}(a|q, c)]] \tag{1}$$

where θ refers to the parameters of an encoder model (pretrained BERT-based model in our case) which maps a question and context to a sequence of contextualized token embeddings, and ϕ is the classifier weights used to predict the start and end indices of tokens.

3.1 Pre-trained encoding model

In this project, we used DistilBERT [5]as the baseline model. This model is a distilled version of the BERT base model which is faster for inference or downstream tasks. In addition to the mandatory DistilBERT model, we also experimented on another BERT-based pretrained model, ALBERT (A Lite BERT) [6] for further analysis of our methods' performance on different types of models. ALBERT implements factorized embedding parameterization and cross-layer parameter sharing which significantly reduce the number of parameters and increase the training speed of BERT. The parameter reduction techniques also act as a form of regularization that stabilizes the training and enables the model to scale better than the original BERT. Therefore, ALBERT as another option provides some insight into how the ensemble of dataset experts work with the scaling of model parameters. On top of DistilBERT/ALBERT, we used a linear layer to output the probability of each token being selected as answer start or answer end.

3.2 Dataset Experts

In order to combine the advantages of multi-dataset and single-dataset approaches. We designed the following algorithm.

First, we train one multi-dataset model based on formula 1 by training on mixed mini-batches with approximately equal numbers of examples from each dataset.

After acquiring θ and ϕ, we diverge and finetune θ and ϕ on dataset D_i to get dataset expert θ_i and ϕ_i, i.e.

$$\arg\min_{\theta_i,\phi_i} \mathbb{E}_{q,c,a\in D_i}[-\log p_{\theta,\phi}(a|q,c)] \tag{2}$$

The difficulty here lies in hyperparameter finetuning, e.g. how much data from each dataset to finetune on, how long to train, learning rate etc. such that dataset expert is able to acquire informational characteristics of each dataset without overfitting that may hurt its generalization to out-of-domain datasets. Note that our two step method can be viewed as equivalent to a type of sampling design. Each dataset expert is trained on the entire collection of datasets but examples from one dataset are sampled more often than others.

3.3 Ensemble

Now with a collection of dataset experts $\{\theta_i, \phi_i : 1 \leq i \leq k\}$. We make the assumption that each dataset expert makes predictions independent from each other and the probability of an index i being the correct start or end index is the product of its probability predicted by each expert, i.e.

$$p_{\text{start/end}}(i) = \prod_{l=1}^{k} p^l_{\text{start/end}}(i)$$

This formulation selects (i, j) only when multiple experts consider the answer span as likely. Under the situation that we have many candidate answer spans, this approach enables the selection of the best one. We ensemble them at test time by selecting the start and end indices i and j subject to $i \leq j$ such that

$$\arg\max_{(i,j)} \prod_{l=1}^{k} p^l_{\text{start}}(i) \cdot \prod_{l=1}^{k} p^l_{\text{end}}(j)$$

In practice, this is equivalent to

$$\arg\max_{(i,j)} \sum_{l=1}^{k} \log p^l_{\text{start}}(i) + \sum_{l=1}^{k} \log p^l_{\text{end}}(j)$$

4 Experiments

4.1 Data

Table 1: Statistics for datasets used for building the QA system (Table borrowed from [7])

Dataset	Question Source	Passage Source	Train	dev	Test
		`in-domain datasets`			
SQuAD [5]	Crowdsourced	Wikipedia	50000	10,507	-
NewsQA [7]	Crowdsourced	News articles	50000	4,212	-
Natural Questions [6]	Search logs	Wikipedia	50000	12,836	-
		`oo-domain datasets`			
DuoRC [9]	Crowdsourced	Movie reviews	127	126	1248
RACE [10]	Teachers	Examinations	127	128	419
RelationExtraction [11]	Synthetic	Wikipedia	127	128	2693

As is shown in Table 4.1, three in-domain datasets SQuAD [8], NewsQA [9], and Natural Questions [10] are divided into training and validation sets. The model is trained on the in-domain training

set and in-domain dev set serves as a checker for overfitting and hyperparameter searching. The oo-domain datasets include DuoRC [11], RACE [12], and RelationExtraction [13]. The final evaluation of the model is conducted on the oo-domain test set.

4.2 Evaluation method

Performance is measured via two metrics: Exact Match (EM) and F1 score.

EM is 1 when the model prediction exactly matches the ground truth answer, and 0 otherwise.

F1 is the harmonic mean of precision and recall, i.e.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

EM is a strict measure of how close the model prediction is to the ground truth while F1 score is more tolerating.

4.3 Experimental details

In this project, we used one DistilBERT and two ALBERT pretrained models with the following configurations:

- DistilBERT: 6 layers; 768 embedding dimension; 3072 hidden dimension; 12 attention heads; 66M parameters
- albert-base-v2: 12 repeating layers; 128 embedding dimension; 768 hidden dimension; 12 attention heads; 11M parameters
- albert-large-v2: 24 repeating layers; 128 embedding dimension; 1024 hidden dimension; 16 attention heads; 17M parameters

All of our experiments were run on one 16 GB NVIDIA Tesla V100 GPU.

We trained distilbert and corresponding classifier layer on the entire training set using $batch\_size = 16, learning\_rate = 3 \times 10^{-5}$. The model takes 3 epochs to converge and training takes 2 hours. We trained albert-base-v2 and corresponding classifier layer on the entire training set using $batch\_size = 16, learning\_rate = 3 \times 10^{-5}$. The model takes 3 epochs to converge and training takes 2.5 hours. We trained albert-large-v2 and corresponding classifier layer on the entire training set with $batch\_size = 6, learning\_rate = 1 \times 10^{-5}$. The model takes 1 epoch to converge and training takes 5.5 hours. The batch size is chosen to use as much GPU memory as possible in training without going out of memory.

For each of the base models, we trained a set of dataset experts, each corresponding to one of the in-domain datasets (SQuAD, NewsQA, and Natural Question). For distilbert models, we finetune it on each in-domain dataset with $batch\_size = 16, learning\_rate = 3 \times 10^{-6}$. For albert-base-v2 models, we finetune it on each in-domain dataset with $batch\_size = 16, learning\_rate = 3 \times 10^{-6}$. Finetuning time varies slightly with the size of the dataset. On average, it takes 2000 steps and 20 minutes for the expert model to converge on one dataset. For albert-large-v2 models, we finetune it on each in-domain dataset with $batch\_size = 8, learning\_rate = 1 \times 10^{-6}$. On average, it takes 6000 steps and 1.5 hours for the expert model to converge on one dataset.

4.4 Results

By applying expert ensemble on DistilBERT pretrained model, we achieved F1 = 61.107 and EM = 41.032 on the Default Final Project - RobustQA Track Test Leaderboard.

From Table 4.4, we can see that our proposed ensemble of dataset experts outperforms the performance of multi-dataset models. When using albert-base as the pretrained model, our expert ensemble method achieves F1 score of 54.81 on the oo-domain dev set which is 3.91 points or 7.68% higher than the score of the multi-dataset counterpart. When using distilbert as the pretrained model, our expert ensemble method achieves F1 score of 50.33 on the oo-domain dev set which is 0.45 points higher than the score of the multi-dataset counterpart. When using albert-large as the pretrained model, our

expert ensemble method achieves F1 score of 57.84 on the oo-domain dev set which is 3.58 points or 6.60% higher than the score of the multi-dataset counterpart. This shows that the expert ensemble method consistently improves the performance of multi-dataset baseline even with the scaling of model size. In addition, through expert ensemble, the albert-base model is able to achieve a F1 score comparable to that of albert-large-multi-dataset model which has 1.5 times more parameters than albert-base. This indicates the potential of expert ensemble as a training time efficient method to improve qa models' generalizability.

One possible reason why the improvement of expert ensemble is less significant on DistilBERT than on AlBERT is due to the difference in learning ability between DistilBERT and AlBERT.

In table 3, we show that our dataset experts are indeed more specialized at featuring a particular in-domain datasets. This increase of specialization, through ensemble, transfers well to out-of-domain datasets. Actually, our experiment has shown that if we finetune a pretrained model directly on a single dataset without training it on multi-datasets first, the acquired expert does not outperform the multi-dataset model when evaluated on its expertise dataset (expert is not really an expert!) This demonstrates the necessity of training on general domain before finetuning on specific dataset.

Table 2: Performance of models on out-of-domain dev sets

	oo-domain		RACE		DuoRC		RelationExtraction	
Model	F1	EM	F1	EM	F1	EM	F1	EM
Distilbert	49.88	34.55	37.44	24.22	45.68	37.30	66.46	42.19
Distilbert-expert-ensemble	50.33	33.77						
Albert-base-multi-datasets	50.90	34.29	37.47	21.88	45.42	34.13	69.73	46.88
Albert-base-expert-ensemble	54.81	38.48	42.51	25.78	50.63	41.27	71.24	48.44
Albert-large-multi-datasets	54.26	36.65	43.61	26.56	49.14	38.89	69.94	44.53
Albert-large-expert-ensemble	57.84	38.48	49.21	31.25	52.43	38.89	71.80	45.31

Table 3: Performance of models on in-domain dev sets

	SQuAD		NewsQA		Natural Question	
Model	F1	EM	F1	EM	F1	EM
Distilbert-multi-dataset	77.31	63.12	57.75	40.38	69.51	53.17
Distilbert-NewsQA-expert			56.25	38.91		
Distilbert dataset experts	77.71	64.22	58.98	41.24	70.23	53.97
Albert-base-multi-datasets	80.21	65.77	61.51	42.88	69.58	52.40
Albert-base dataset experts	82.11	68.66	63.08	44.21	71.39	54.18
Albert-large-multidatasets	81.88	67.75	61.70	41.50	67.39	49.66
Albert-large dataset experts	84.43	71.43	64.57	44.59	72.14	54.68

The row of "Albert-base dataset experts" shows the performance of albert-base expert i evaluated on dataset i. For example, SQuAD F1 score is achieved by albert-base SQuAD expert on the SQuAD dev set. The same formulation applies for "Albert-large dataset experts". This table is meant to show the effect of dataset finetuning on improving single dataset capturing.

5 Analysis

In this section, we inspect some characteristic outputs of our dataset expert ensemble.

5.1 Example Study 1

Question: Who had the same interest as Winslow according to the text?

Context Paragraph: Winslow Homer was the second of three sons of Henrietta Benson and Charles Savage Homer. He was born in Boston, Massachusetts in 1836 and grew up in Cambridge. His father was an importer of tools and other goods. His mother was a painter. Window got his interest in drawing and painting from his mother. But his father also supported his son's interest. Once, on a business trip to London, Charles Homer bought a set of drawing examples for his son to copy. Young

Winslow used these to develop his early skill. Winslow's older brother Charles went to Harvard University in Cambridge. The family expected Winslow would go, too. But, at the time, Harvard did not teach art. So Winslow's father found him a job as an assistant in the trade of making and preparing pictures for printed media. At 19, Window learned the process of lithography. This work was the only formal training that Winslow ever received in art.

Ground truth: his mother

Prediction by Albert-large-expert-ensemble: his mother. But his father

Prediction by Albert-large-multi-datasets: his father

Analysis: This is an example of how an ensemble might outperform a single model. When "experts" can't agree on answering "mother" or "father", their ideas are synthesized to produce a span that is more likely to contain the correct answer whereas a single model where "father" receives a slightly higher probability will completely miss the correct answer.

5.2 Example Study 2

Question: What is Constans's brothers name?

Context Paragraph: With Constantine's death in 337, Constans and his two brothers, Constantine II and Constantius II, divided the Roman world between themselves and disposed of virtually all relatives who could possibly have a claim to the throne.

Ground Truth: Constantius II

Prediction by Albert-large-expert-ensemble: Constantine II and Constantius II

Prediction by Albert-large-multi-datasets: Constantine II

Analysis: In this example, although the ground truth labeled "Constantius II" as the correct answer, it is not hard to find that both "Constantine II" and "Constantius II" are valid answers. In this case, the ensemble model is able to capture both correct answers.

6 Conclusion

In this project, we developed a method of combining the advantages of multi-dataset and single-dataset models. This approach has been tested to be effective in improving prediction results on out-of-domain datasets and is efficient in training time compared with deploying a larger model. In the future, we could experiment with more diverse types of ensembling such as training a more complicated gating function. In addition, since the idea of dataset experts is closely related to data sampling. It may be worthwhile to explore other sampling techniques that gives different experts which might be specialized in certain features not describable by a single dataset.

References

[1] Priyanka Sen and Amir Saffari. What do models learn from question answering datasets? *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[2] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. volume 3, pages 79–87, 1991.

[3] Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4] Dan Friedman, Ben Dodge, and Danqi Chen. Single-dataset experts for multi-dataset question answering. In *EMNLP (1)*, pages 6128–6137, 2021.

[5] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 2020.

[6] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. 2020.

[7] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[9] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[10] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[11] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension. In *ACL*, 2018.

[12] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[13] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics.

	Quesion	Context	Answer
SQuAD	10	120	3
Natural Questions	9	96	4
NewsQA	8	709	4