Application of Mixture of Experts in Domain-agnostic QA

Jiayi Li, Stanford University
jiayili@Stanford.edu

Introduction: *Single vs Multi datasets?*

Many datasets have been created for training 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. Generalizability is a challenging task as different datasets come from distinct domains and have distinct features.

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

Fig 1. Comparison of the average number of words in questions, contexts, and answers in each dataset

One straightforward approach to utilize data from multiple sources is to combine data from several training datasets and sample uniformly from the large domain. This enables the model to learn general patterns of question answering but 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. This inspires our approach to ensemble several models each representing an expert of a specific in-domain dataset to improve transfer learning to agnostic domains.

Problem Setup

The objective of question answering is to model the distribution p(a|q, c), where q, c, a ∈ D represent a question, context, and answer respectively from a dataset D. We make the standard assumption that the start indices are independent with end indices, i.e. p(span(start = i, end = j)|q, c) = p(start = i|q, c) · p(end = j|q, c). We have a collection of source datasets D = {D1, D2, ...Dk}

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

where θ refers to the parameters of an encoder model (pretrained BERT-based model in our case) and ϕ is the classifier weights used to predict the start and end indices of tokens.

Methods

- **Pre-trained encoding model**

In this project, we leveraged on pretrained ALBERT (A Lite BERT) as the encoder part of the model considering memory constraints, training speed, and scaling ability. On top of ALBERT, we used a linear layer to output the probability of each token being selected as answer start or answer end.

- **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 by sampling 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 Di to get dataset expert θi and ϕi, i.e.

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

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.

- **Ensemble**

Now with a collection of dataset experts {θi, ϕi : 1 ≤ i ≤ k} whose predictions are viewed as independent from each other. We ensemble them at test time by selecting the start and end indices i and i subject to i ≤ i such that

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

Where p-l is the prediction made by the l th expert.

Experiments

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

Model	oo-domain F1	oo-domain EM	RACE F1	RACE EM	DuoRC F1	DuoRC EM	RelationExtraction F1	RelationExtraction EM
Distilbert	49.88	34.55	37.44	24.22	45.68	37.30	66.46	42.19
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

Model	SQuAD F1	SQuAD EM	NewsQA F1	NewsQA EM	Natural Question F1	Natural Question EM
Distilbert-multi-dataset	77.31	63.12	57.75	40.38	69.51	53.17
Distilbert-NewsQA-expert			56.25	38.91		
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

Here, we list some important findings from the experiment:

- In table3, distilbert-NewsQA-expert is trained directly on NewsQA without seeing other datasets. It underperforms the multi-dataset model, indicating that even experts need more data to learn better and the first step in our method is necessary.
- The expert ensemble method consistently improves the performance of multi-dataset baseline on both in-domain and oo-domain datasets even with the scaling of model size.
- Through expert ensemble, the albert-base model is able to achieve a F1 score comparable to that of albert-large-multi-dataset model which is much larger in size.

Analysis

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 that even ground truth fails to do.

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.