BoBA: Battle of BERTs with Data Augmentation

Ishira Fernando   Rishi Desai

Department of Computer Science, Stanford University

Problem

Building QA models that generalize well to unseen data distributions that are distinct from the models' training distribution is a difficult problem. Humans are innately good at this task, which is known as domain generalization, but even state of the art models on QA benchmarks such as SQuAD are known under-perform on out-of-domain (OOD) data. Domain generalization is especially important to QA systems as they are expected to transfer well to different applications, which may involve structurally and contextually different language use.

We present **BoBA**: **B**attle **o**f **B**ERTs with Data **A**ugmentation. BoBA combines Data Augmentation and Mixture of Experts (MoE) to improve domain generalization, by outperforming a DistilBERT baseline by 5.17 F1 points and 6.55 EM points.

Background

One of the most widely utilized NLP models is DistilBERT, a knowledge distilled version of BERT. DistilBERT remains comparable to BERT in its performance, despite being over 40% smaller than BERT. Due to it's performance and ease of development, DistilBERT is used as an atomic component in many QA systems today.

Augmenting training data through random transformations is well known to help with domain generalization and robustness. Due to the rule-based nature of language data, augmentation can be very complicated. In *EDA*, the authors propose a range of simple techniques by which to perform data augmentation on language/text classification models. We adapt this approach in our project to implement data augmentation to improve the domain generalization of BoBA.

Approach

Mixture of Experts

Mixture of Experts are a class of ensemble models consisting of several individual expert models, each trained on one domain, which are "gated" by a learned gating function. The outputs of each expert is linearly combined according to the output of the gating function, which learns which expert to weigh more by conditioning on the input.

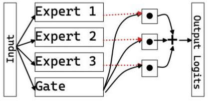
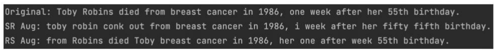


Figure 1. BoBA uses another DistilBERT as the gating function.

Data Augmentation

Adapting *EDA*'s approach we implemented two different types of data augmentation: synonym replacement (SR) and random swapping (RS). SR replaces words that are not stop words with a random synonym with some probability β_{sr}. RS swaps words in the context acording to a probability γ_{rs}. Some examples are shown below.



```
Original: Toby Robins died from breast cancer in 1986, one week after her 55th birthday.
SR Aug: toby robin conk out from breast cancer in 1986, i week after her fifty fifth birthday.
RS Aug: from Robins died Toby breast cancer in 1986, her one after week 55th birthday.
```

Figure 2. An example of SR and RS augmentation. Augmentation may become nonsensical with very high β_{sr} or γ_{rs}.

Experiments

Dataset and Metrics

- 3 *in-domain* datasets: SQuAD, NewsQA, Natural Questions
- 3 *out-of-domain* datasets: DuoRC, RACE and Relation Extraction
- EM (Exact Match) and F1 metrics to evaluate performance

Other Mixture of Experts Approaches

Model	In-Val F1	In-Val EM	Out-Val F1	Out-Val EM
DistilBERT Baseline	70.35	54.54	**46.86**	**30.89**
Squad Only	**75.67**	**61.93**	42.83	27.49
NewsQA Only	55.54	38.25	38.85	25.92
NaturalQ Only	66.82	50.79	36.70	20.68
MoE with MLP	62.13	45.12	41.85	25.65
Adversarial Training	20.62	N/A	12.11	N/A

Table 1. Baseline results for DistilBERT experts trained on *one* of the in-domain training sets.

Training BoBA

1. Train model M on the 3 in-domain train sets with data augmentation.
2. Let expert E_i be M after finetuning on the i-th ood-train set with data augmentation.
3. Train MoE model $B = f(E_1, E_2, E_3)$ on the 3 in-domain train sets without augmentation, where f is the gating function.
4. Finetune and validate B on the three ood-train sets with data augmentation.

Model	Batch Size	Learning Rate	Epochs	γ_{rs}	β_{sr}
Race Only	64	8e-7	3	0.00	0.30
Relation Extraction Only	32	1e-5	3	0.40	0.90
Duorc Only	32	1e-5	3	0.50	0.70
DistilBERT Gate (out-of-domain)	16	3e-6	1	0.90	0.80

Table 2. Hyperparameters for training each expert. γ_{rs} and β_{sr} are the random sequence percentage and synonym replacement percentage, respectively. We use the AdamW optimizer and Cross-Entropy Loss.

Results

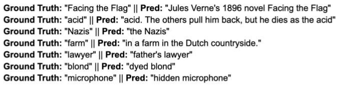
Model	Out-Val F1	Out-Val EM
DistilBERT Baseline*	46.86	30.89
Unfinetuned Expert (M)	49.54	34.55
Race Only (E_1)	49.50	34.55
Relation Extraction Only (E_2)	50.06	35.86
Duorc Only (E_3)	48.91	32.98
MoE with Frozen Experts	43.21	26.71
MoE with Unfrozen Experts*	50.14	36.65
MoE with Unfrozen Experts	**52.03**	**37.44**

Table 3. The performance of the unfinetuned expert, the three experts, and two gating function variants of our MoE model. The astrix * denotes no data augmentation was used.

- BoBA scored F1 of **59.03** and EM of **40.69** on the test set.
- Having unfrozen experts improves accuracy because the experts are optimized in sync with the gating function.

Analysis

When looking at model outputs on the OOD validation set, we found that a significant proportion (31.5%) of BoBAs answers were overlapping: they contained the ground truth answer but were not exact matches. Figure 3 shows some sample overlapping answers. This highlights that the EM score may not be a very valuable metric when evaluating the efficacy of a QA system, as the overlapping answer can still be useful in many cases.



Ground Truth: "Facing the Flag" || **Pred:** "Jules Verne's 1896 novel Facing the Flag"
Ground Truth: "acid" || **Pred:** "acid. The others pull him back, but he dies as the acid"
Ground Truth: "Nazis" || **Pred:** "the Nazis"
Ground Truth: "farm" || **Pred:** "in a farm in the Dutch countryside."
Ground Truth: "lawyer" || **Pred:** "father's lawyer"
Ground Truth: "blond" || **Pred:** "dyed blond"
Ground Truth: "microphone" || **Pred:** "hidden microphone"

Figure 3. Some sample overlapping answers

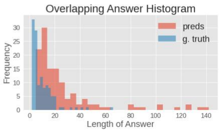


Figure 4. Answer-length histogram for overlapping answers in the OOD validation set.

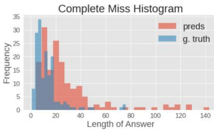


Figure 5. Answer-length histogram for complete misses in the OOD validation set.

To further investigate causes for this we plotted a histogram of answer lengths for both overlapping and completely missed answers in Figures 4 and 5. We find that the distribution of answer lengths predicted by the model has both a longer tail and has a right shifted centroid in comparison to the ground truth answers. This implies our model is failing to find the minimum-span answer as it appears to prefer guessing longer answers over shorter answers. This is likely a feature learned from training on the old-domain, which may have featured longer ground truth answers. Going forward, reducing this error may require the use of a length penalty through a custom loss function that penalizes over-length non-exact matches when the model is being fine-tuned.

Conclusions

We developed BoBA, a MoE that uses random swapping and synonym replacement augmentation along with fine-tuned unfrozen experts and a DistilBERT gating function, as a means of improving the domain generaliztion of DistilBERT on QA tasks. We gain a 5.17 point increase in F1 score and 6.55 point increase in EM score. On an unseen test-set our model reached an F1 of 59.03 and an EM of 40.69 indicating strong generalization to the new domain.

Over the course of our experiments we found that data augmentation requires careful fine-tuning on a case-by-case basis due to the significant distributional differences between domains. Furthermore, hyper-parameters (largely learning rate and batch size) appeared to have a large impact on the generalization of the model. While frozen experts appeared to not generalize as well as unfrozen experts, exploring models that had different numbers of frozen transformer blocks may prove insightful going forward. Finally, we were unable to explore other augmentation practices such as back translation or random insertion/deletion, which may further boost the performance of our model. We leave these as avenues to explore in the future.

rdesai2@stanford.edu

CS 224N: Natural Language Processing with Deep Learning

ishira00@stanford.edu