Explore Different Transfer Learning Methods for More Robust QA Models

Yu Shen Lu[1], Dingyi Pan[2]

[1]Computer Science Department, [2]Symbolic Systems Program, Stanford University

Motivation

Goals

- To build a robust QA system that can be generalized to out-of-domain dataset
- To analyze Feature Distortion Theory in NLP QA task

Background

- Previous studies [1] in computer vision shows complete fine-tuning(FT) distorts the pretrained features, as it tries to fit them onto a randomly initialized head. Methods, such as linear probing (LP), that only tune the randomly initialized head, before complete fine-tuning leads to better performance in out-of-domain tasks. LP+Fine-tuning(LPFT) is shown to be effective in image classification.
- Various methods such as data-augmentation and mixture of experts are shown to be effective in few-shot learning, can they help our model adapt to out-of-domain data?

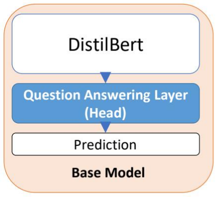
Questions

- Can the feature distortion theory be generalized to NLP domain for the QA task?
- What techniques help create a robust few-shot QA model?

Experiment Setup

Model Structure

- The base model consists of pretrained DistilBert and a randomly initialized head.
- Model predicts the answer to the question given the context.



Training Strategies

- Bayesian Optimization
- Partial Tuning

Datasets

	In-domain			Out-of-domain		
	SQuAD	NewsQA	Natural Questions	DuoRC	RACE	Relation Extraction
Train	50,000	50,000	50,000	127	127	127
Eval	10,507	4,212	12,836	126	128	128

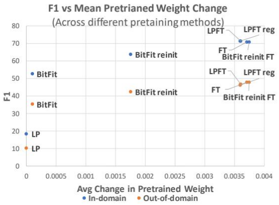
Training Strategies Exploration (In-Domain)

Feature Distortion Theory

During fine-tuning, only the gradients in the in-domain training subspace are updated, but those that are orthogonal to the training subspace remain unchanged [1]. This means we want to do partial fine-tuning before full fine-tuning.

Partial Fine-Tuning Methods

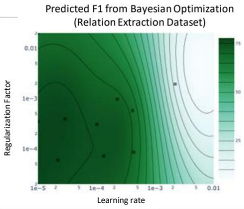
- Linear Probing (LP): Tuning the random head
- BitFit [2]: Tuning bias terms on top of LP
- Reinit: Reinitialize last pretrained layer
- Fine-Tuning (FT): Tuning all parameters



Fine-Tuning (FT) is required for high F1.

Hyper-parameter Search

Bayesian Optimization is used to search for the optimal learning rate and weight decay value.



Findings

We found Feature Distortion Theory does not apply to NLP QA task here. High F1 requires changing the pretrained features to some extent in this task. Maybe it is because the pretrained weights are for text generation and not QA tasks.

Data Augmentation

Strategies

- Token-level: We implemented four data augmentation methods using EDA [3], including Random Deletion (RD), Random Insertion (RI), Random Swap (RS), Synonym Replacement (SR), which results in x4 more data.
- Sentence-level: Random Insertion (a random sentence from the in-domain context), resulting in x5 more data.

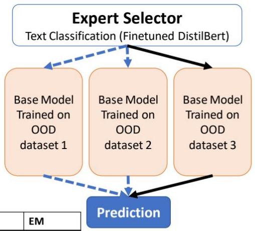
Findings

We tested the BitFit + Reinit + FT model on the original and the augmented datasets. The results suggest that data augmentation slightly improves the performance.

	Training		Validation	
	F1	EM	F1	EM
Baseline	46.60	32.20	48.75	34.82
RD	47.83	33.51	49.89	**35.86**
RI	48.14	32.98	50.26	34.82
RS	**48.48**	**33.77**	50.00	35.08
SR	47.43	33.51	**50.36**	35.60
Sent_RI	48.09	33.51	49.13	33.77

Mixture of Experts (Out-of-Domain)

Our final model is inspired by Mixture of Experts (MoE). Since we need to handle queries from three different datasets, we train three models, each fine-tuned on one out-of-domain dataset. Expert selector is trained to pick which dataset the input is from and query the correct model.


Expert Selector
Text Classification (Finetuned DistilBert)
Base Model Trained on OOD dataset 1
Base Model Trained on OOD dataset 2
Base Model Trained on OOD dataset 3
Prediction

Model	Data Augmentation	F1	EM
Baseline	True	49.27	35.08
Baseline	False	48.75	34.82
MoE	True	**50.79**	**36.65**
MoE	False	49.62	34.82

Conclusions

- Feature distortion theory does not apply in our case. This may be because the pretrained model is for a different task, and the model is lightweight. Therefore, we observe that some distortion is required to achieve good performance. Partial training doesn't yield a good enough initialization, and the model performance still largely relies on the later fine-tuning process.
- Since there are very few out-of-domain training data, randomly changing the context not only increases the number of data but also adds noise to it, which leads to a slightly better performance.
- Mixture of Expert picks the best model to answer each question, but due to limitation on the performance of each expert model, it only improves the overall performance slightly.

References

[1] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning distorts pretrained features and underperforms out-of-distribution. In *International Conference on Learning Representations*, 2

[2] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2021.

[3] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196, 2019.