A Mixture 0f Experts For Out Of Domain Expertise

Mia Sarojini Kynadi
Stanford Center For Professional Development
Mentor TA: Kamil Ali

Introduction

Out-of-Domain Question Answering

How to get a model to generalise well beyond its training distribution?

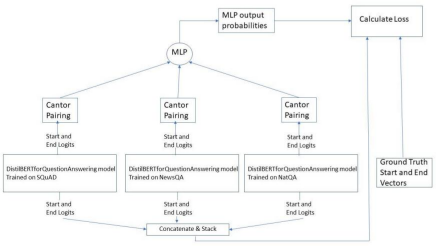
1.Use separate models or Experts, trained on different training datasets
2.Combine them with a gating mechanism

Key Findings

- Cantor Pairing to to generate single valued input to classifier (gating function) from DistilBERT output
- Loss function based on mixture of gaussians assumption

$$E^c = -log \sum_i p_i^c e^{-\frac{1}{2}||a^c - o_i^c||^2}$$

Methods

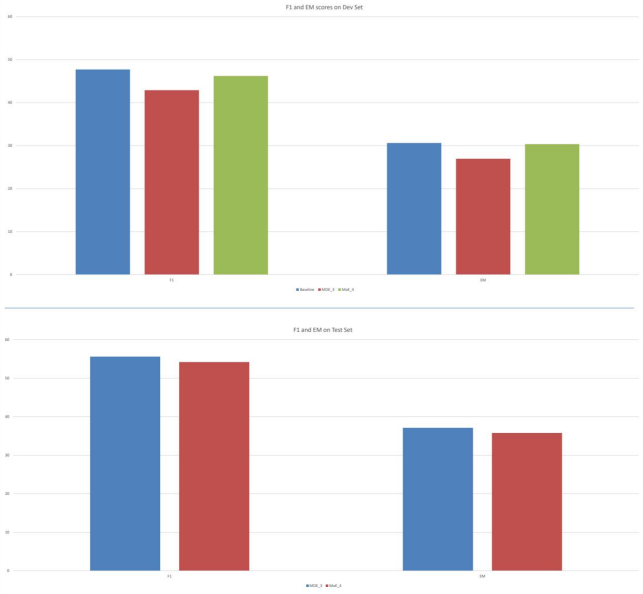


- 3 DistilBERTforQuestionAnswering models trained separately on SQuAD, NewsQA and NatQA datasets
- Outputs of models converted to single values with cantor pairing
- Probability distribution over experts generated by MLP

Analyses

- The MoE model with 4 experts does better than the MoE model with 3 experts during training because it benefits from the fully trained and finetuned 4th expert. However, I think this causes the MLP to preferentially choose the 4th expert very often and hence the ability to generalise outside the training set is reduced.
- The MoE model with 3 experts very often tends to predict an answer with a longer context window than the provided answer (ground truth label), causing the EM score to drop, but not affecting the F1 score as much.
- Further finetuning of the MLP parameters, the loss function, the appropriateness of cantor pairing need to be investigated to improve out-of-domain performance.

Results



Conclusions

- Mixture of Experts model approach used here does not improve upon the baseline performance of DistilBERTforQuestionAnswering model trained on in-domain datasets and finetuned on small out-of-domain training set
- Loss function based on mixture of gaussians assumptions performs better than cross entropy loss based on performance during training

Future Work

Explore unsupervised clustering of training data to find patterns in dataset and use that to train separate models

References

- "Adaptive Mixture of Local Experts", Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, Geoffrey E. Hinton
  Neural Computation, 1991
- "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, 2019.

Acknowledgments

- Mentor TA Kamil Ali
- Professor Manning and all TAs