Applying Ensembling Methods to BERT to Improve Model Performance

Charlie Xu, Solomon Barth, Zoe Solis
{cxu2, sbarth, zoesolis}@stanford.edu

Problem

Machine Comprehension (MC) is a complex task in NLP that aims to understand written language. Question Answering (QA) is one of the major tasks within MC, requiring a model to provide an answer, given a contextual text passage and question. It has a wide variety of applications, including search engines and voice assistants, making it a popular problem for NLP researchers.

According to the SQuAD 2.0 leaderboard, most high-performance models incorporate BERT in some way. All of the current top 18 submissions incorporate BERT in some way. However there is much variation in the choice of ensembling and parameter tuning that can be done on top of BERT that differentiates much of the leaderboard.

Data/Task

Dataset: The Stanford Question Answering Dataset (SQuAD) is a large, diverse database of over 150,000 high-quality Wikipedia passages, reading comprehension questions, and accepted answers compiled by Stanford researchers. Roughly half of all questions are impossible to answer based on the given context. It uses the Exact Match (EM) and harmonic mean (F1) scores as metrics and maintains a leaderboard to see how the highest performing models compare against one another and against human performance.

Context:	"The descendants of Rollo's Vikings and their Frankish wives would replace the Norse religion and Old Norse language with Catholicism (Christianity) and the Gallo-Romance language of the local people"
Question:	What was the Norman religion?
True Answer:	Catholicism

Context:	"Norman mercenaries were first encouraged to come to the South by the Lombards to act against the Byzantines"
Question:	Who did the Normans encourage to come to the South?
True Answer:	<No Answer>

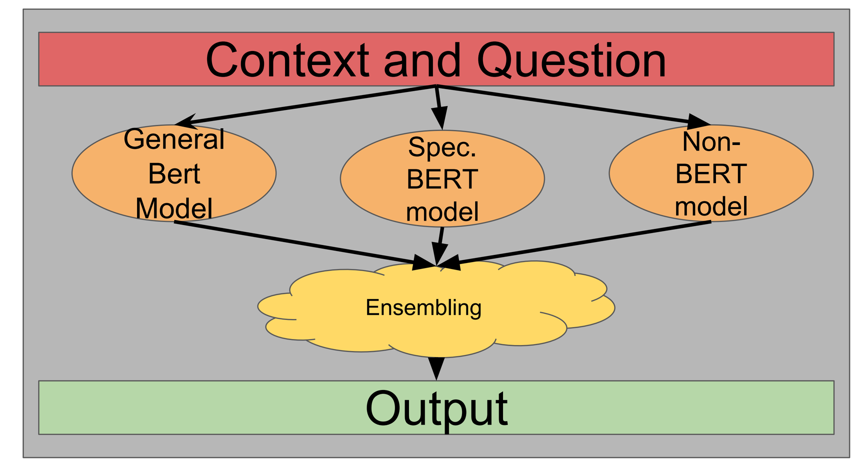
We also used the SQuAD 1.1 dataset in the process of building our models. SQuAD 1.1 contains over 100,000 context paragraphs, questions, and answers, although it differs from SQuAD 2.0, in that all of its questions are possible to answer.

Task: Use the provided context to produce an answer to the given question, or no answer if the question is impossible to answer. With the SQuAD dataset, all answers are selected to be subsets of the context, so the task can be reduced to finding the start and end indices of the predicted answer within the context.

Approach

Model Architectures Used:

- BiDAF
- BERT
- SSTQA (intended)



Overall Approach:

- Implement a variety of QA models
- Adjust dataset to facilitate training of both general and specialized models
- Ensemble models together to obtain superior performance

Dataset Manipulation:

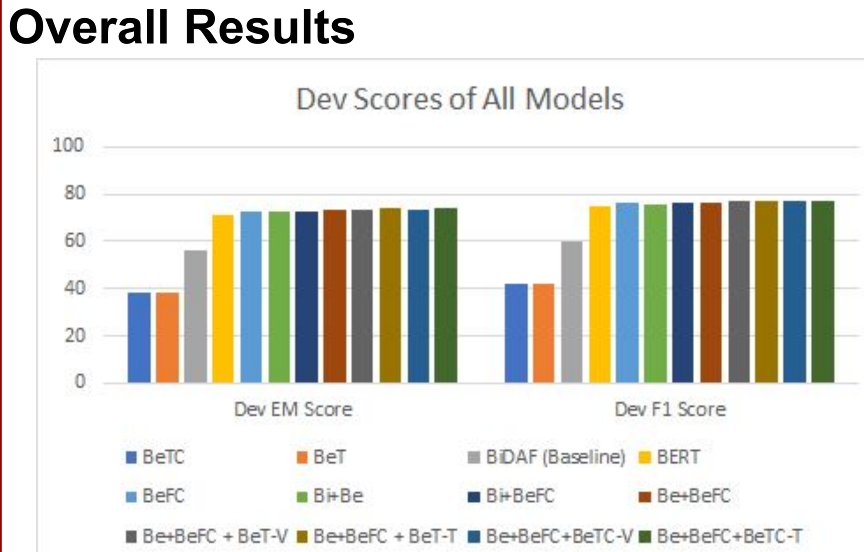
- Trimmed SQuAD 2.0 to create dataset of only possible questions
- Combined SQuAD 1.1 data with both regular 2.0 dataset and trimmed dataset

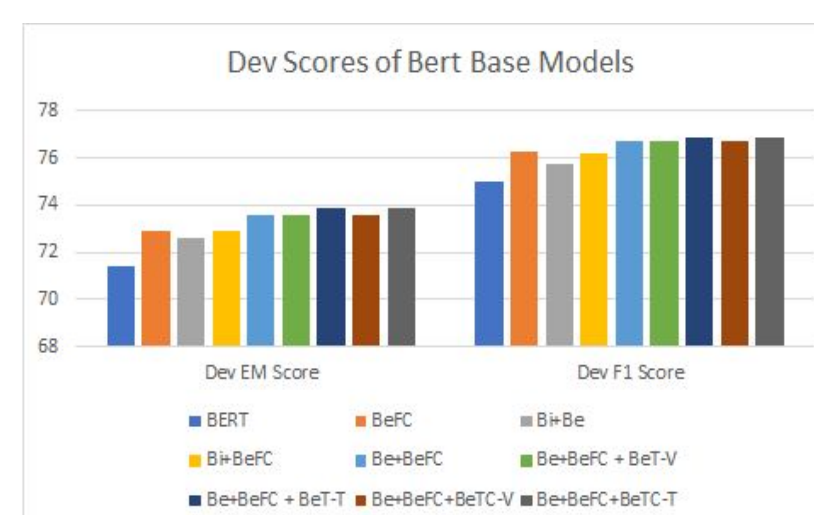
Ensembling:

- Developed three ensembling methods
  - Selecting predictions with the highest joint probability
  - Using a general model as a classifier to identify possible questions and comparing the highest joint probability between general model predictions and specialized model predictions
  - Using a general model as a classifier to identify possible questions and using the specialized model's predictions

Results

Overall Results

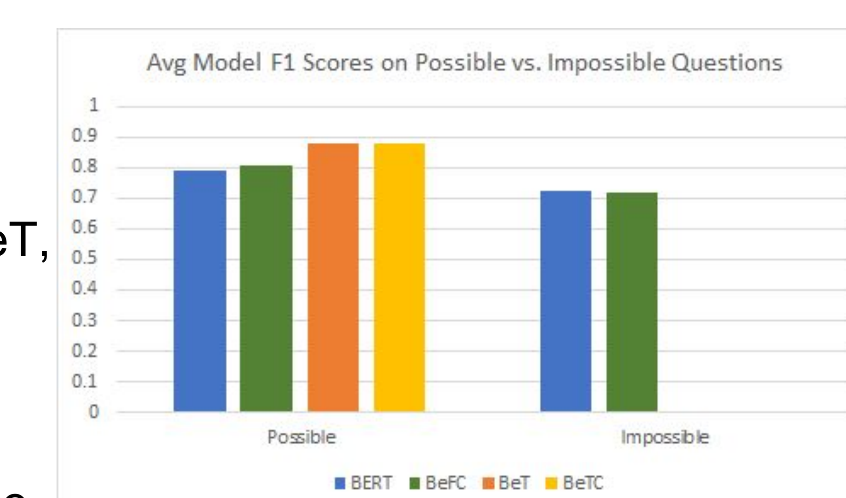




- Ensembling models, except BiDAF sometimes, improves performance
- Use of BeFC and BeTC-T/BeT-T tends to increase overall performance

General vs. Specialized Model Performance

- General models, like Bert and BeFC, perform well on possible and impossible questions
- Specialized models, like BeTC and BeT, perform terribly on impossible questions, but extremely well on possible questions
- Use of classifier-based ensemble method should lead to better performance



Test Set Performance

- Given their high F1 and EM scores, we used our Be+BeFC+BeTC-T model and Be+BeFC model on our test set, yielding these results.

Table 4: Model Test Set Results

Model	Test EM Score	Test F1 Score
Be+BeFC	73.576	76.721
Be+BeFC+BeTC-T	73.221	76.454

Analysis

Comparison of Possible vs. Impossible Questions F1 Scores

Model	Possible	Impossible
BERT	0.7923	0.7229
BeFC	0.8083	0.7206
BeT	0.8777	0.0013
BeTC	0.8776	0.0013
Bi+Be	0.8002	0.7175
Bi+BeFC	0.8176	0.7175
Be+BeFC	0.8064	0.7383
Be+BeFC+BeT-V	0.8064	0.7383
Be+BeFC+BeT-T	0.7866	0.7592
Be+BeFC+BeTC-V	0.8064	0.7383
Be+BeFC+BeTC-T	0.7866	0.7592

- Ensembling generally improved performance in both categories
- Addition of trimmed models improved ability to answer impossible questions, reduced ability to answer possible questions

Comparison of Question Type F1 Scores

Model	How	Other	What	When	Where	Which	Who	Why
BERT	0.7396	0.6158	0.7613	0.7574	0.6791	0.7630	0.7868	0.6915
BeFC	0.7348	0.6002	0.7695	0.7635	0.6887	0.7874	0.7880	0.7027
BeT	0.4054	0.3275	0.4121	0.4714	0.4369	0.5384	0.4186	0.3590
BeTC	0.4109	0.3130	0.4100	0.4786	0.4392	0.5358	0.4211	0.3620
Bi+Be	0.7398	0.6110	0.7606	0.7889	0.6799	0.7587	0.7804	0.6915
Bi+BeFC	0.7378	0.6002	0.7707	0.7978	0.6848	0.7808	0.7848	0.7027
Be+BeFC	0.7569	0.5946	0.7744	0.8104	0.6942	0.7801	0.7870	0.7022
Be+BeFC+BeT-V	0.7569	0.5946	0.7744	0.8104	0.6942	0.7801	0.7870	0.7022
Be+BeFC+BeT-T	0.7623	0.6328	0.7753	0.8106	0.6870	0.7656	0.7915	0.7164
Be+BeFC+BeTC-V	0.7569	0.5946	0.7744	0.8104	0.6942	0.7801	0.7870	0.7022
Be+BeFC+BeTC-T	0.7623	0.6328	0.7753	0.8106	0.6870	0.7656	0.7915	0.7164

- BeFC performed better than BERT in all categories, especially "Which" questions
- Ensembling models increased performance across the board, especially with "When" and "How" questions

Conclusions

Overall Findings

- Ensembling boosts performance by leveraging the relative strengths of different models
- Manipulating training dataset led to significant differentiation of results between models, even those of same model architecture
- Use of a generally trained model as a classifier to determine when to use a specialized model can lead to significant increases in performance

Future Investigations

- Explore generating models specialized to predict impossible questions
- Train explicit NN classifier to classify questions as possible or impossible
- Implement other models and ensemble methods to make more ensemble models

Acknowledgements

Thank you to Sahil Aagrawal, Vivekkumar Patel, and Dr. Chris Manning for their guidance and support!

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
Mrinmaya Sachan and Eric P. Xing. Self-training for jointly learning to ask and answer questions. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1:629–640, 2018.
Shuaipeng Liu, Shuo Liu, and Lei Ren. Trust or suspect? an empirical ensemble framework for fake news classification. WSDM Cup 2019 Fake News Classification Challenge, 2019.