Improving QA Performance Using Bert + X

Danny Takeuchi (dtakeuch@stanford.edu), Kevin Tran (ktran23@stanford.edu)

Problem Definition

In recent years, Bidirectional Encoder Representations from Transformers (BERT) implementations have made significant gains compared to previous state-of-the-art QA models. However, the BERT models are still fooled into answering questions about contexts that don't contain the answer. We are leveraging the BERT huggingface pytorch implementation and attempting to improve upon this baseline model, especially with false positives where a question has no answer but we predict one anyway.

Dataset

We use the Squad 2.0 question-answering dataset which is composed of over 100,000 questions on crowdsourced Wikipedia articles. The Squad 2.0 also contains question, context pairs that have no correct answer as opposed to the Squad 1.1 dataset.
We also analyze:
- Exact Match without non-answers
- % of non-answer false positives (FP)
- % of non-answer false negatives (FN)

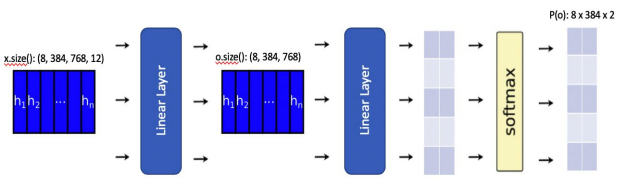
Results

Table 1: BERT PCE Scores

Model	EM	F1	EM(No non-answers)	FP	FN
BERT Baseline + Hyperparam. Tuning	73.288	76.376	73.68	26.70	13.06
BERT + Highway	73.001	76.035	74.45	27.81	12.64
BERT Baseline	72.606	75.843	73.60	27.17	13.06
BERT + WCEL	72.359	75.379	74.25	27.56	13.92
BERT + Last3WHL	66.535	69.402	71.83	32.26	19.48
BERT + Last4WHL + ReLU	65.729	68.613	70.32	28.35	24.98
BERT + BiDAF	65.093	67.697	68.29	29.53	25.32
BERT + Transformers	63.123	65.813	67.91	31.53	23.32
BERT + WHL	61.056	64.322	67.80	37.31	22.20
BERT + WHL + Highway	57.947	61.508	63.45	39.80	22.23
BERT + CNN	51.468	56.352	58.93	44.58	26.89

Models

BERT with Dot Product Attention over Hidden Layers (BERT + WHL)
Concatenates the last x hidden states into a matrix of size (batch-size, input-len, emb-dim, x) in order to generate an overall representation that may be able to capture additional semantic and positional patterns.



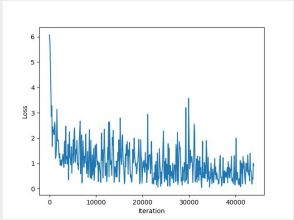
BERT with CNN (BERT + CNN)
Utilizes a 1-dimensional Convolutional Neural Network to pool information from four hidden layers of the BERT model with a kernel size of 2.

BERT + Highway
Runs the final hidden states through a highway network before computing logits. Highway networks contain gating mechanisms that can filter out potentially irrelevant information.

BERT with additional Transformer Layers (BERT + Transformers)
Adds two additional transformer layers on top of the pre-trained BERT.

BERT with BiDAF (BERT + BiDAF)
Segments the final hidden states of BERT into question hidden states and context hidden states and fed these hidden states into a BiDAF model with the Embedding and Highway layers removed.

BERT with Weighted Cross Entropy Loss (BERT + WCEL)
Penalizes the loss for cases in which there is no answer by weighting the start and positions of 0 higher.



Analysis

Due to constraints in space, we decided to analyze the errors of a few models that we expected to perform the best.

BERT + Highway
Context: "The Normans were famed for their martial spirit and eventually for their Christian piety."
Question: "Who was famed for their Christian Spirit?"
Correct Answer: No Answer
Bert Baseline Answer: No Answer
BERT + Highway Answer: "The Normans"
Issue: The gating mechanism is unable to realize “spirit” and “Christian” have no relation to each other.

BERT + BiDAF
Context: "They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia."
Question: "Who was the Norse leader?"
Correct Answer: "Rollo"
Bert Baseline Answer: "Rollo"
BERT + BiDAF Answer: "King Charles III"
Issue: The BERT + BiDAF model seems to be overcomplicating a simple question by inferring some additional knowledge. Since the Norse swore fealty to King Charles III, he is now the new leader.

Conclusion

Despite trying many different approaches, our best performing model was BERT Baseline + Hyperparameter Tuning. Although several of our approaches, such as weighting the last three hidden layers to generate an overall representation layer, made intuitive sense our ablation studies showed us that introducing additional complexities may not always improve model performance.

References

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." arXiv preprint arXiv:1611.01603 (2016).