SQuAD 2.0 Project

Liyang Sun, Konrad Morzkowski, Phillip Hoovestol

Problem and Data

SQuAD2.0
The Stanford Question Answering Dataset

- Problem of general question answering
- Current top solutions based on transformer architectures (BERT)
- SQuAD 2.0 dataset

- Difference with old SQuAD: unanswerable questions
- Train/Dev/Test split: 129 941/5951/5915

Approach

Char-BiDAF

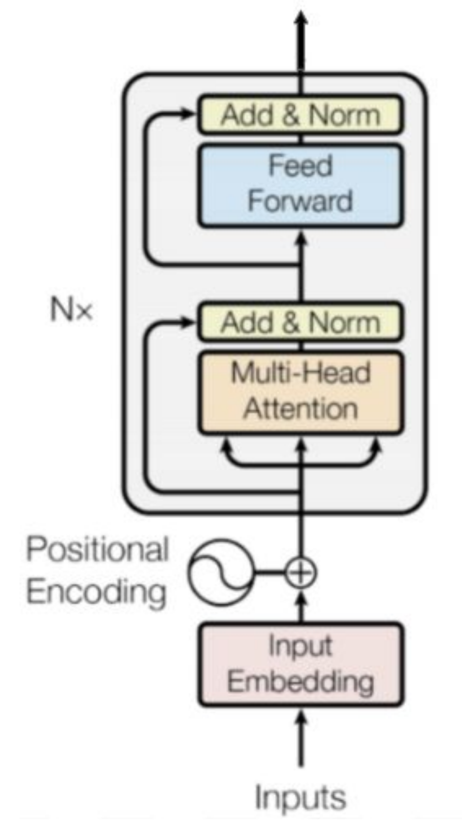
- Adding character based embeddings.
- Learned via 1-D convolutional network.

AoA

- Question to document and document to question attention vectors computed.
- Used to scaled BERT's output

Bert

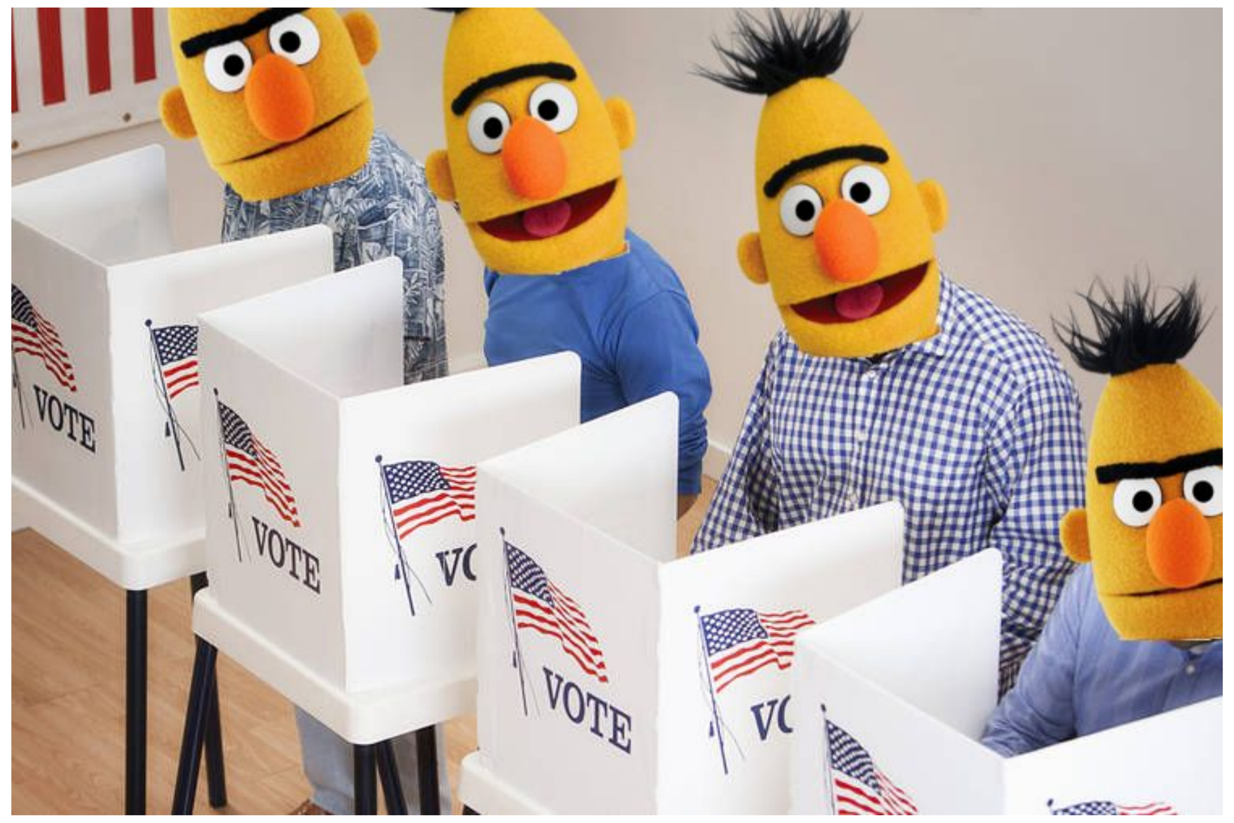
- Based on the transformer encoder
- Pre-trained on masked language modeling and next sentence prediction tasks.
- Variable threshold for none answer predictions



The Transformer Encoder.

Ensemble

- Combining different individual models' outputs
- Plurality vote method
- Models' ranking as tie-breaker
- Specific selection rules, by taking into account models' diversity
- Predict no-answer anytime BERT with threshold +3.0 predicted no-answer.



Behind the Scenes of our Plurality Vote Ensemble

Results

Model	Dev F1	Dev EM
Baseline	61.44	57.92
BiDAF (Character embeddings)	62.24	59.05
BERT (threshold -1.0)	76.12	72.99
BERT (threshold -3.0)	75.85	73.03
BERT (threshold +3.0)	72.33	68.87
Naive Ensemble	76.47	73.77
Ensemble with ranking	76.61	73.87
Ensemble with ranking, models selection, and null rule	77.07	74.19

Model	Test F1	Test EM
Ensemble with ranking, models selection, and null rule	77.40	74.49
BERT (threshold -1.0)	76.69	73.61

Analysis

AoA
- Performed worse than the baseline
- Scaling of BERT's inputs unnecessary (more sophisticated attention)

BERT
- Null score difference threshold: substantial role in performance

Ensemble
- Ranking tie-breaker and no-answer technique improves the model
- Better performance with diverse individual models, even if those models individually perform worse.

Model	Type 1 error	Type 2	Type 3
BERT (threshold -3.0)	647	637	416
BERT (threshold -1.0)	451	782	477
BERT (threshold +3.0)	240	1025	513

Conclusion

Takeaways

- Solid grasp on current state-of-the-art question answering models.
- General understanding of neural networks implementation and evaluation.
- Rewarding to be able to identify error categories and improve upon them by simple modeling changes.

Future Work

- Training our models longer to enhance performance
- Synthetic self-training to generate relevant examples for training.

References

Devlin J., Chang M.-W., Lee K., \& Toutanova K. (2018) Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. \textit{arXiv preprint arXiv:1810.04805}

Cui Y., Chen Z., Wei S., Wang S., Liu T. \& Hu G. (2017) Attention-over-Attention Neural Networks for Reading Comprehension. \textit{arXiv preprint arXiv:1607.04423v4}

Dietterich T. G. (2000) Ensemble Methods in Machine Learning.