# Applying Transformer-XL to Q&A

Sam Xu
{samx@Stanford.edu}

## Problem and Motivation

### Goal

- Question and answering is a machine comprehension task in which a passage and a question are provided to a machine, and the machine must provide the answer.
- The recent release of SQuAD 2.0 has expanded the task by including unanswerable questions.

### Motivation

- We attempt to utilize the recent Transformer-XL architecture, which leverage segment-level recurrence mechanism to perform better on both long and short sequences [2].
- We use QANet as our starting point [1], but we also look at other recent models for inspiration.

### Task

In this project, we build a deep learning model to perform machine comprehension on the SQuAD 2.0 dataset. We utilize techniques such as multi-level attention, self-attention, fuse representation, and Transformer-XL hidden states [3][4].

## Embedding Layer

We used the following embeddings as the input of our models.

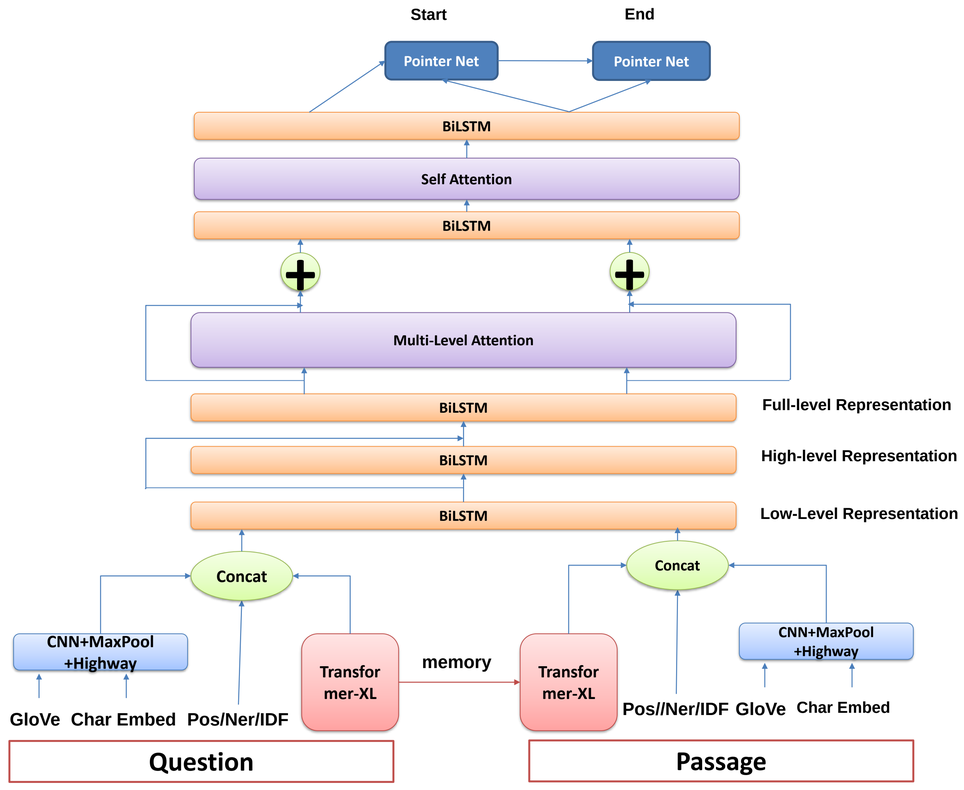
- GLoVE
- Character-level embedding with convolutions, max-pooling, and highway layer activation.
- Parts of Speech (PoS)
- NeR (Nearest Entity recognition)
- Exactly matching tokens
- Lower case match
- Lemma matches
- Term frequency – inverse document frequency (tf-idf)

The additional embeddings are extracted with the aid of spaCy.

Additionally, we utilize the final hidden states of the Transformer-XL in our model. Where Transformer-XL is pretrained on the WikiText Long Term Dependency Language Modeling Dataset.

*Code Source for transformer-XL: https://github.com/huggingface/pytorch-pretrained-BERT*

## Architecture Implementation

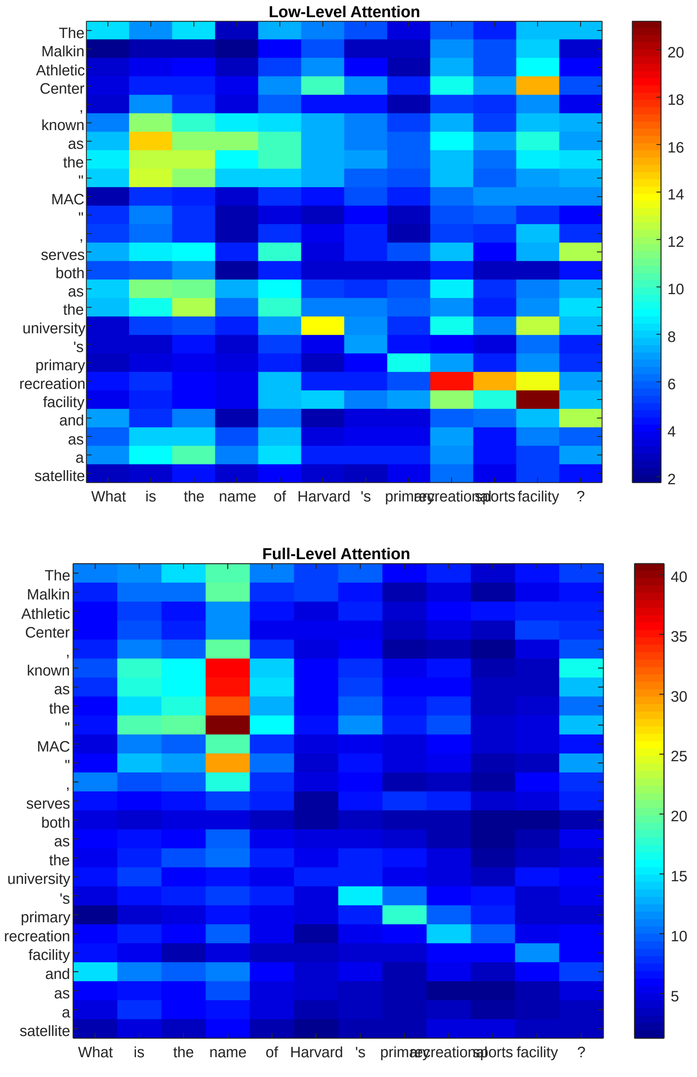



## Results

| Model (dev) | F1 | EM |
|---|---|---|
| Baseline | 55.6 | 58.9 |
| QANet | 64.2 | 67.9 |
| Final Model | 66.3 | 70.2 |

- We used the AdaMax optimizer with a decaying learning rate starting at 0.002.
- We used a dropout rate of 30% across all layers, including the embedding layer
- Whereas QANet completed a single epoch within 20 minutes, the final model took over 3 hours per epoch (trained on a RTX 2080ti).
- This massive reduction in speed is likely caused by the removal of convolutional parallelism and the inclusion of Transformer-Xl
- We achieved a final test set result of EM and F1 of 64.3 and 68.1

## Attention Visualization




## References

[1] . Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, Quoc V. Le . (2018) *QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension*

[2]. Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, Ruslan Salakhut-dinov . (2019) *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context.*

[3] Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, Weizhu Chen: (2017) *FusionNet: Fusing viaFully-Aware Attention with Application to Machine*

[4]. Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, Ming Zhou . (2018) *Read + Verify:Machine Reading Comprehension with Unanswerable Questions*.