WHAT COMPLIMENTS CO-ATTENTION

Problem:

We aim to reproduce a Co-Attention layer on the Stanford Question Answering dataset (SQuAD) baseline and investigate its relationship with other common SQuAD techniques. While Co-Attention has been used to significantly improve state of the art F1 scores, we wanted to compare how several common SQuAD techniques complement a model with Co-Attention.

Background:

- Hierarchical Question-Image Co-Attention Networks for Question Answering
  - Lu et al., 2017
  - Introduces idea of Co-Attention Network
  - Proposed for relationship between verbal question and visual image
  - Combines attention matrices between a vector of image features and a vector corresponding to the words in the questions
  - Important not just "where to look" but also "what words to listen to"

- Dynamic Co-Attention Networks for Question Answering
  - Xiong et al., 2018
  - Co-Attention Network
  - Utilizes Lu’s co-attention framework
  - Combines a co-attention matrix between the words in the document and the words in the query
  - Dynamic decoder
  - Makes an initial prediction for the start and end predictions
  - And then bases each subsequent prediction on the last prediction
  - Keeps track of all of the previous predictions using an LSTM
  - Motivation: Local mistakes can be overcome by the iterative power of the dynamic decoder

- Highway Matrix Networks
  - Combines a number of different computations at each level
  - Specifically, a number called the size of the matrix pool
  - Identical structure but with differently trained weights
  - Only retains the maximum result achieved from each computation
  - Motivation: There are multiple kinds of questions/documents that require multiple approaches in question answering, which can be calculated in parallel using a highway neural network.

Method:

We couple our baseline with a Co-Attention layer, illustrated in following image by Xiong et al. It contains two-way attention between question and context, and also includes a second-level attention computation which attends over the attention output representation. Here, $a^q$ and $a^c$ represent the normalized attention weights. We test how this layer is complemented by other SQuAD methods such as Dynamic Decoder and Character Embeddings.

Experiments:

We run 8 experiments to test how different techniques improve our baseline model vs. our baseline model with Co-Attention. The techniques we test are any character embeddings and the dynamic decoder.

Analysis:

- Using co-attention network caused major improvements
- Using character embeddings resulted in substantial improvements
- Negative effect of co-attention network greater than positive effect of character embedding
- Major improvements all co-attention-based models by using dynamic decoder

Conclusion:

Although Xiong et al. demonstrated notable improvements using co-attention networks and a 4-attention-based decoder, it seems that most methods implemented under limited resources and training time, conclusion is not a worthy endeavor after all.

References: