WHAT COMPLIMENTS COATTENTIO

Problem:

We aim to reproduce a Coattention layer on the Stanford Question Answering dataset (SQuAD) baselin and investigate its relationship with other common SQuAD techniques. While Coattention has been pro et al.) to significantly improve state of the art F1 scores, we wanted to compare how several common S techniques compliment a model with Coattention.

Background:

Papers:

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**
    - **Computes an attention matrix between a vector of image features and a vector corresponding to the words in the questions**
- **Important not just "where to look" but also "which words to listen to"**

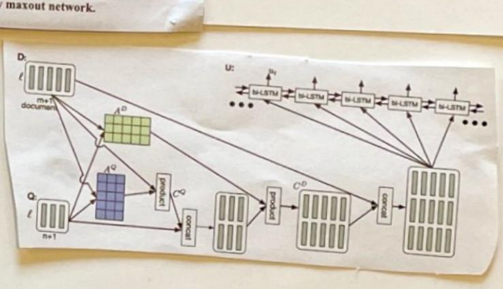
Dynamic Coattention Networks for Question Answering
Xiong et al, 2018

- **Coattention Network**
    - **Utilizes Lu's co-attention framework**
    - **Computes 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 prediction**
    - **And then bases each subsequent prediction on the last prediction**
    - **Keeps track of all of the previous predictions using an LSTM**
    - **Motivation: Local minima can be overcome by the iterative power of the dynamic decoder.**
- **Highway Maxout Network**
    - **Computes a number of different computations at each level**
    - **Specifically, a number called the size of the maxout pool**
    - **Identical structures 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 to question-answering, which can be calculated in parallel using a highway maxout network.**

Method:

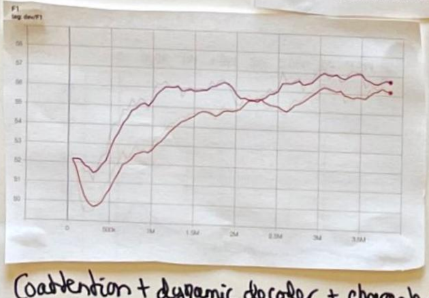
Methods:

We couple our baseline with a Coattention 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^D and A^Q represent the normalized attention weights. We test how this layer is complimented by other SQuAD methods such as Dynamic Decoder and Character Embeddings.

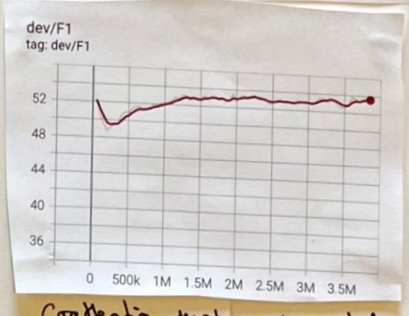


Experiments:

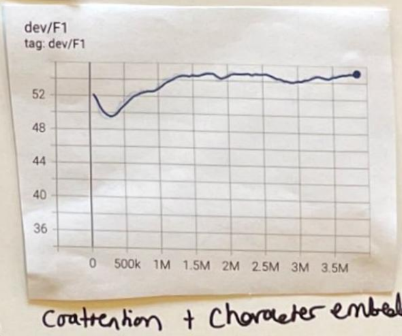
We run 8 experiments to test how different techniques improve our baseline model vs our baseline mod Coattention. The techniques we test are character embeddings and the dynamic decoder.



Coattention + dynamic decoder + character embeddings
-orange: 1 iter g decoder
-red : 2 iters g decoder



Coattention
1 iter g decoder - dynamic decoder



Coattention + Character embeddings

Analysis

- **Using coattention network caused major deprovements**
- **Using character embeddings resulted in substantial improvements**
- **Negative effect of coattention network greater than positive effect of character embeddings**
- **Minor improvements all coattention-based models by using dynamic decoder**

Conclusion

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

References

References:
Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh. *Hierarchical Question-Image Co-Attention for Visual Question Answering*. (19 Jan 2017). https://arxiv.org/pdf/1606.00061.pdf
Caiming Xiong, Victor Zhong, Richard Socher. *Dynamic Coattentive Networks for Question Answering.* (6 Mar 2018). https://arxiv.org/pdf/1611.01604.pdf