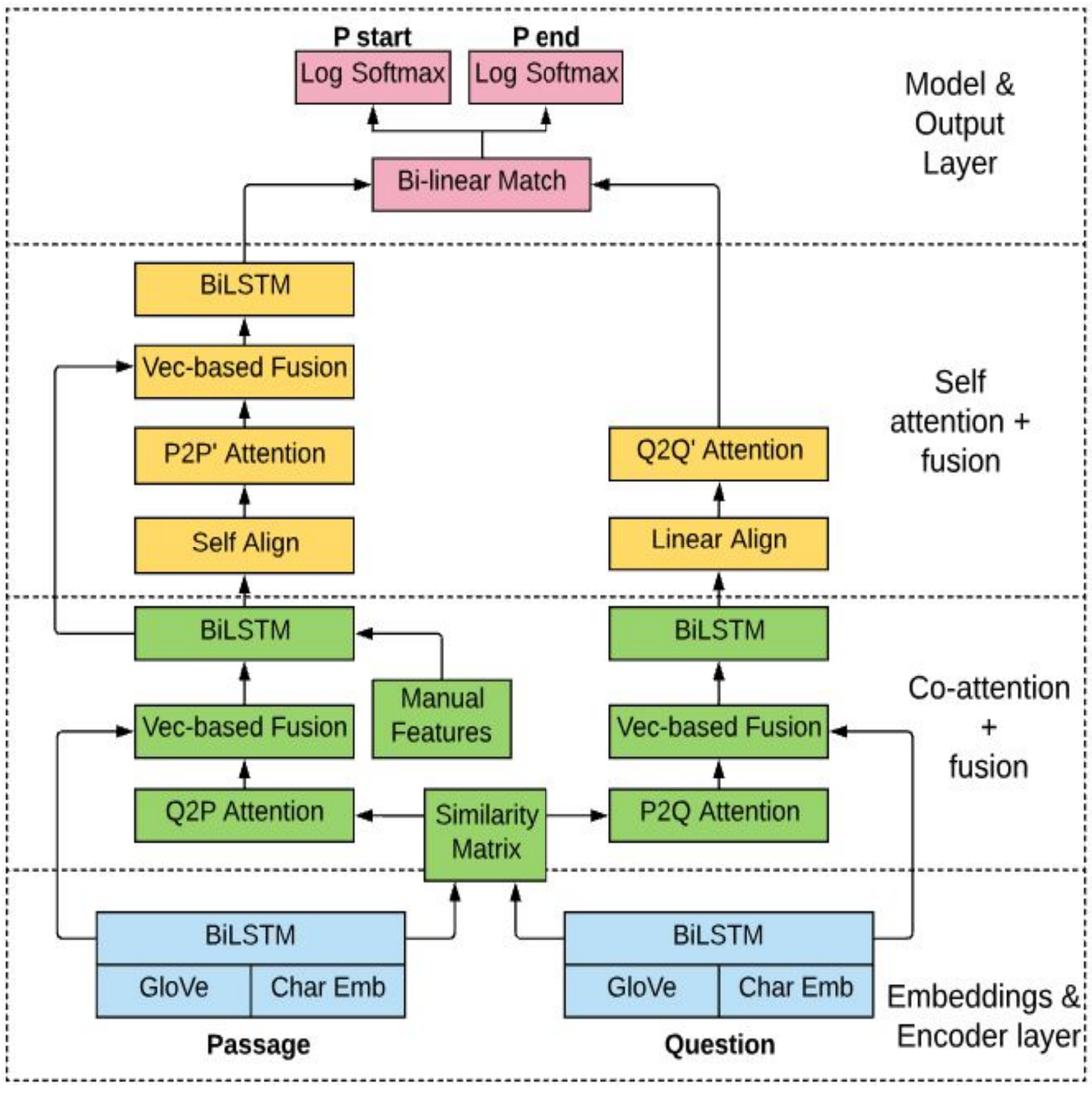
# Hierarchical Attention Fusion Network for Question Answering

Yang Chen and Marius Seritan

## Problem

- **Goal**: Build a Question Answering model on Stanford Question Answering Dataset (SQuAD)
- **Data**: Default SQuAD train, dev, test set provided by TA
- **Evaluation**: F1, EM

## Architecture




## Results

| SLQA Layers [3] | Dev | | Test | |
|---|---|---|---|---|
| | EM | F1 | EM | F1 |
| Embedding (Word + Char) + Output | 47.01 | 48.65 | - | - |
| ... + Paragraph Co-attention + Fusion | 51.59 | 54.09 | - | - |
| ... + Paragraph Self Attention+ RNN | 60.78 | 63.62 | 57.95 | **61.59** |
| ... + Question Self Align + RNN | 61.45 | **64.61** | 57.80 | 61.35 |
| ... + Manual feature | 58.60 | 61.76 | - | - |
| **BiDAF [1]** | **EM** | **F1** | **EM** | **F1** |
| Dropout prob = 0 | 58.03 | 61.4 | - | - |
| Dropout prob = 0.2 | 57.44 | 60.74 | - | - |
| Dropout prob = 0.5 | 52.13 | 55.53 | - | - |
| + Char embedding [2] | 59.15 | 62.43 | - | - |

## Conclusions and Discussions

### Conclusions

- By replacing the attention layer with the hierarchical attention fusion network, we achieved F1 score 64.6 comparing with 61.4 for the BiDAF model. On test set we achieved F1 score 61.6.
- Paragraph self-attention gave the largest boost among all changes we made.
- Adding character embedding helped improve F1 score.

### Discussions

- Our reference paper claimed a higher F1 score (74.43) on SQuAD 2.0 than ours in the experiments. We believe incorporating ELMo and an unknown set of manual features are the two major reasons for the performance disparity.
- Adding question self attention led to a lift of F1 score on dev set, but caused a slight drop on test set. The reason may involve different questions in the dev and test sets.
- Adding manual features caused decreasing F1 on the dev set. We added the frequency of each word as one single feature, which was not effective. If using word entity type, POS tagging, or other features that are more relevant, performance might be boosted. The result also shows that manual features could make large impact (good or bad) directly to the result.

## References

[1] Seo, M., Kembhavi, A., Farhadi, A., etc. (2016). Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.
[2] Kim, Y., Jernite, Y., Sontag, D., etc. (2016). Character-aware neural language models. In 13th AAAI Conference on Artificial Intelligence.
[3] Wang, W., Yan, M., Wu, C. (2018). Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering. arXiv preprint arXiv:1811.11934.