Question Answering on SQuAD 2.0 using Transformer-XL

Ianis Bougdal-Lambert(1), Julien Hedou(2)

(1) Department of Aeronautics and Astronautics, Stanford University (2) Department of Biomedical Informatics, Stanford University

Introduction

SQuAD 2.0 tackles the challenge of text comprehension through the task of question answering. Our study aims to use the novel Transformer-XL architecture on this dataset and to compare the results to the current baseline performing results, such as the Bi-directional attention Network (BiDAF). The Transformer-XL model was originally designed to capture longer-term dependencies in the network.

The Transformer-XL architecture has been assessed by monitoring different performance metrics: F1 score, EM score, and models training time using a NV6 Microsoft Azure virtual machine with 6 vcpus and 56 GB memory.
For all the training and testing phases, we used the original available SQuAD 2.0 dataset. This dataset contains 50% context/query pairs that do not have any answer.

Baseline

In order to compare our results with the Transformer-XL architecture, we also worked on variations of the BiDAF model. (See *Fig1.*) First we used a default model that only included word embeddings, which we improved by incorporating character embeddings.

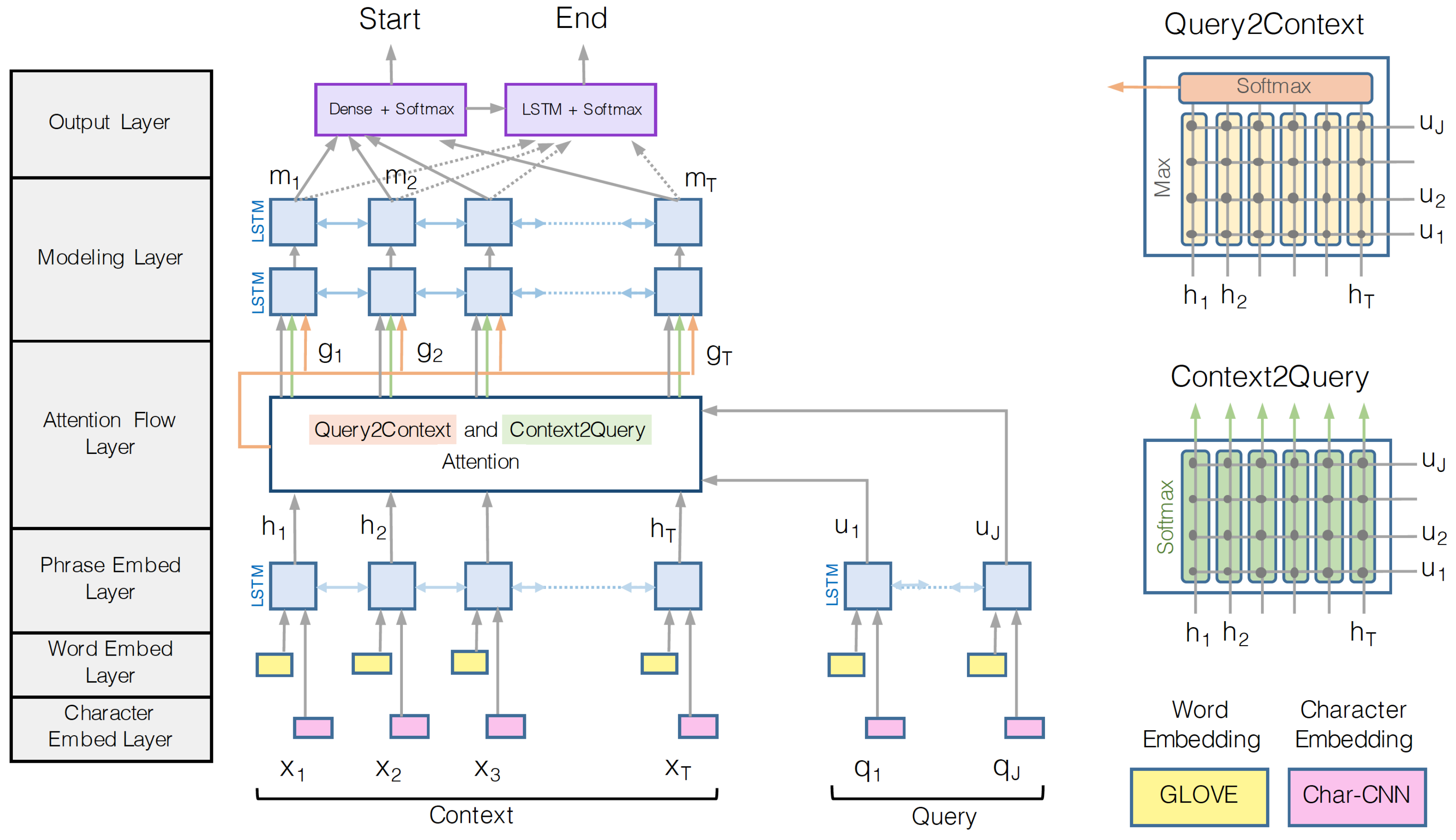


Fig.1 BiDAF architecture

Methods

Transformer-XL is a self-attention network that aims at improving on the vanilla Transformer by taking into account a much larger context.

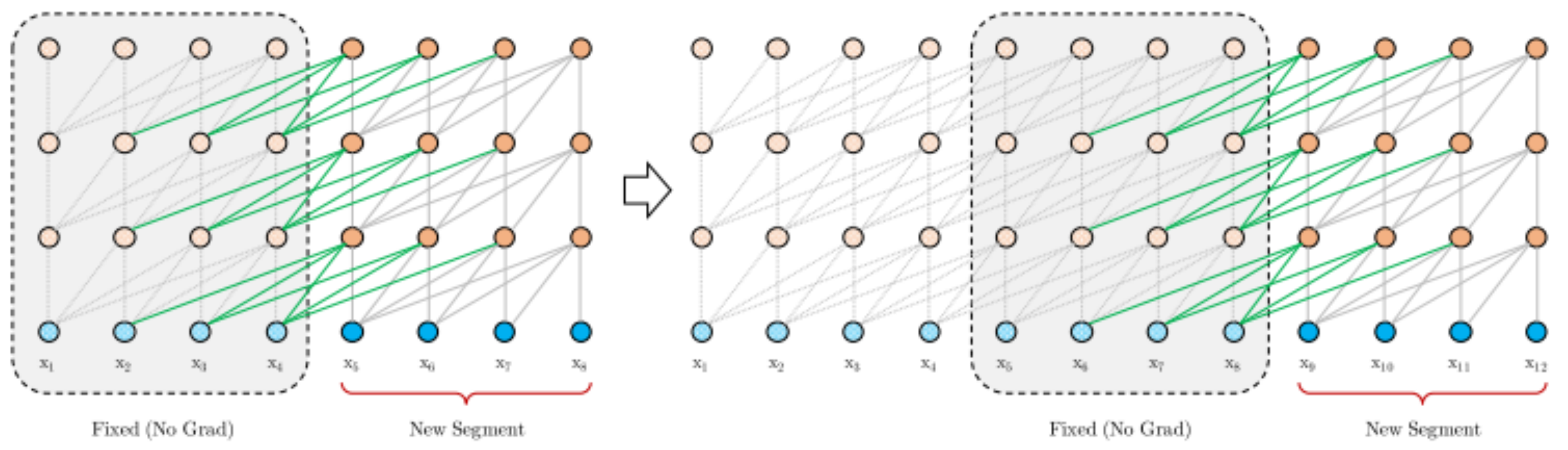
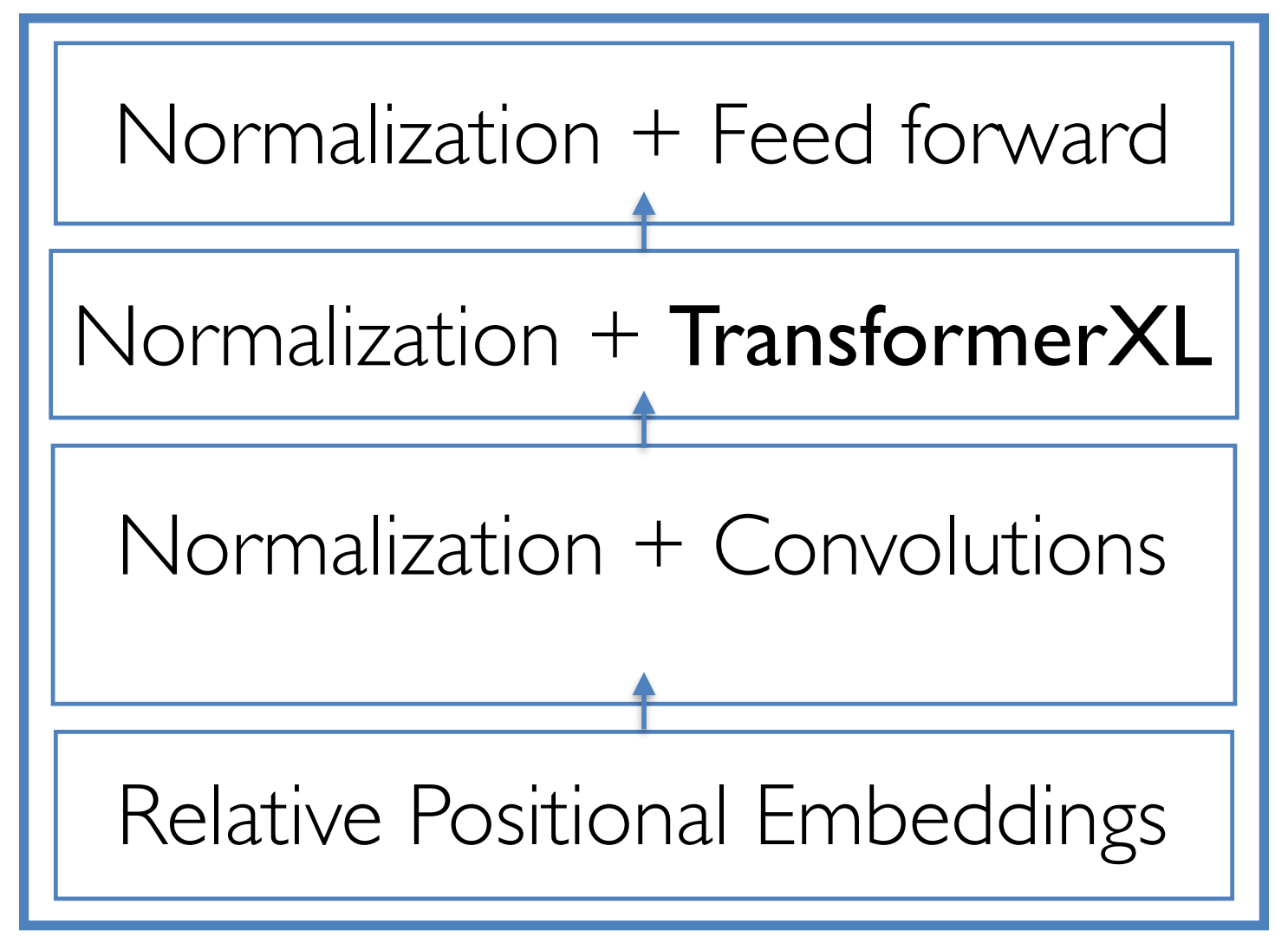


Fig.2 Usage of previous segment weights in to capture longer-term dependencies

Instead of treating each segment of input separately, connections are added between the hidden units of the previous and the current segment and the weights of the previous hidden units are frozen. This allows information to be passed along the segments, effectively introducing a recurrence mechanism at the segment-level.


Normalization + Feed forward

Normalization + **TransformerXL**

Normalization + Convolutions

Relative Positional Embeddings

Fig.3 New encoder block

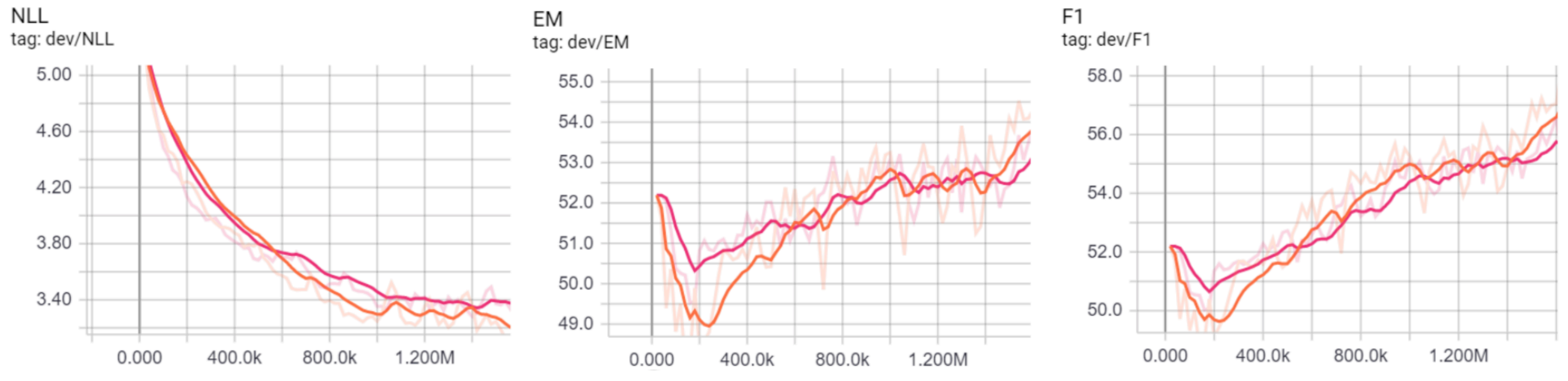
We adapted the QANet architecture to include the Transformer-XL layers in the encoder blocks. Each block consists of a stack of several depth-wise convolutional layers, followed by a self-attention layer and a feed-forward layer. Here we modify the original vanilla transformer and replace it by a TransformerXL attention.

Metrics Performances

Incorporating character embeddings considerably improves on the baseline score. The final scores on the test set are presented in the table. We see that our models achieve good scores on the SQuAD dataset, although they do not improve on their respective baseline.

Model	EM score	F1 Score	Time per epoch
Baseline	55.99	59.29	~40 mins
Baseline + char embeddings	60.3	64.19	~40 mins
TransformerXL	52.7	55.6	~50 mins
TransformerXL + char embeddings	51.11	55.12	~50 mins

Transformer-XL Results



(a) NLL on dev set (b) EM score on dev set (c) F1 score on dev set

Fig.4 Performance metrics on the development set

In orange is the basic model and in red is the model that incorporates character embeddings. The models converge after around 1.5M iterations.

Discussion

Surprisingly, we note that models based on Transformer-XL do not reduce training time as we expected it. One reason could be that the reduced training time was achieved on a model consisting only of a Transformer-XL attention. Here, we also add several layers of convolution that make the model more complex and likely slow it down. Further improvement of our models and hyperparameter tuning could allow us to surpass scores obtained by state-of-the-art non-PCE models such as QANet.

However, our models reveal two limitations. First, we could not use the full range of the Transformer-XL laters due to memory issues. Additionally, the relatively small number of words in the input may explain why the Transformer-XL does not yield strong results when compared to a more simple vanilla architecture.

References

1. Seo, Minjoon / Kembhavi, Aniruddha / Farhadi, Ali / Hajishirzi, Hannaneh(2016): *Bidirectional attention flow for machine comprehension*.
2. Yu, Adams Wei / Dohan, David / Luong, Minh Thang / Zhao, Rui / Chen, Kai / Norouzi, Mohammad / Le, Quoc V(2018): *Qanet: Combining local convolution with global self-attention for reading comprehension*.
3. Dai, Zihang / Yang, Zhilin / Yang, Yiming / Cohen, William W / Carbonell, Jaime / Le, Quoc V / Salakhutdinov, Ruslan(2018): *Transformer-XL: Language Modeling with Longer-Term Dependency*
4. [https://github.com/kimiyoung/transformer xl] *https://github.com/kimiyoung/transformer-xl*.