SQuAD Model Comparison

Jiehan Zhu
jzhu01@stanford.edu

Video: https://www.youtube.com/watch?v=I0rOxMMHOX0

Background

Recent development in Deep Learning in NLP has reach better than human performance on multiple NLP tasks. The sequence nature of RNN and LSTM limited the neural network model to parallel calculation, and requires long training time for large dataset. This leads to application of CNN and Attention layer that enable faster training and smaller model with similar or higher preference. The application of BERT has reached new state of art in various tasks, including reading comprehension, and encourage a multitask model structure.

Problem Statement

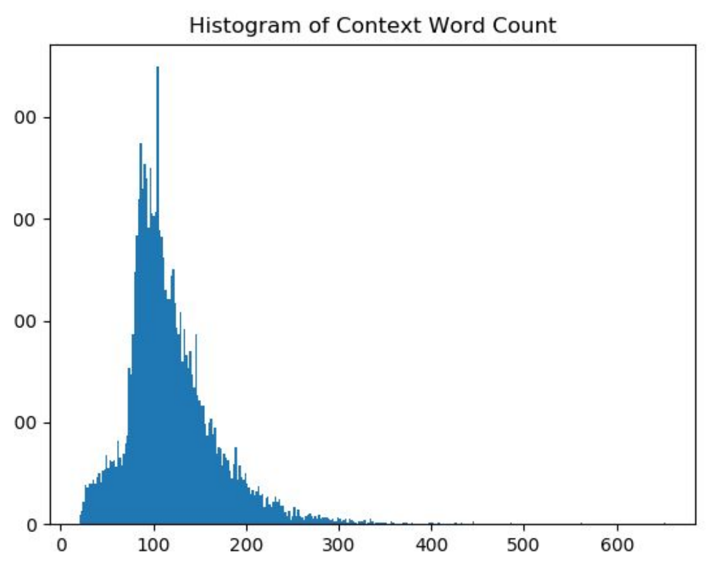
Recent development in Deep Learning in NLP has reach better than human performance on multiple NLP tasks. The sequence nature of RNN and LSTM limited the neural network model to parallel calculation, and requires long training time for large dataset. This leads to application of CNN and Attention layer that enable faster training and smaller model with similar or higher preference. The application of BERT has reached new state of art in various tasks, including reading comprehension, and encourage a multitask model structure.

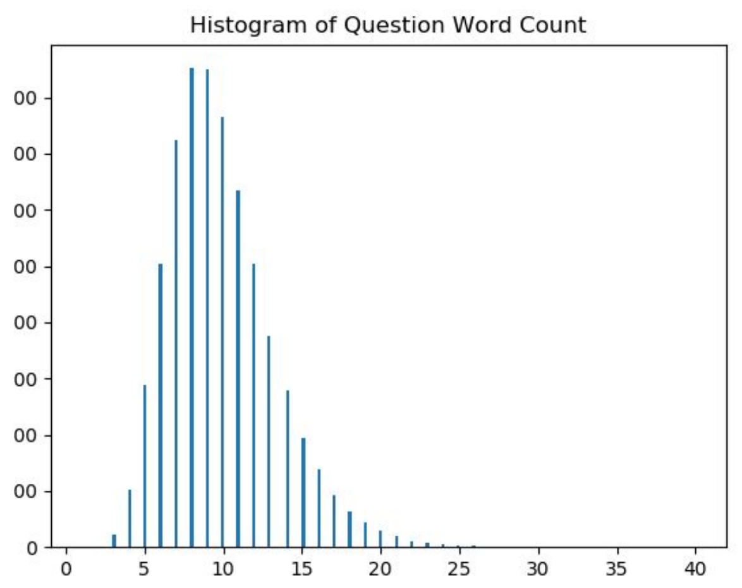
Data Preprocess & Visualization

This paper uses Stanford Question Answering Dataset (SQuAD), which consists questions posed by crowd-workers on a set of Wikipedia articles. About half of questions are answerable, and the answer to answerable every question is a segment of text, or span, from the corresponding reading context.

Context: *Following* the disbandment of Destiny's Child in June 2005, she released her *second solo* album, B'Day (2006), which contained hits "Déjà Vu", "Irreplaceable", and "Beautiful Liar". Beyoncé also ventured into **acting**, with a Golden Globe-nominated performance in Dreamgirls (2006), and starring roles in The Pink Panther (2006) and Obsessed (2009).
Question: After her second solo album, what other entertainment venture did Beyoncé explore?
Answer: acting

The length of context various in the data set, most of the context have 50 to 250 words, while some of the context have more than 600 words. Due to the memory restriction, the maximum number of word loaded for each context is limited at 400, and this only impact less then one percent of data. And the maximum number of word loaded for each question is limited at 50, which should not impact the model at all.





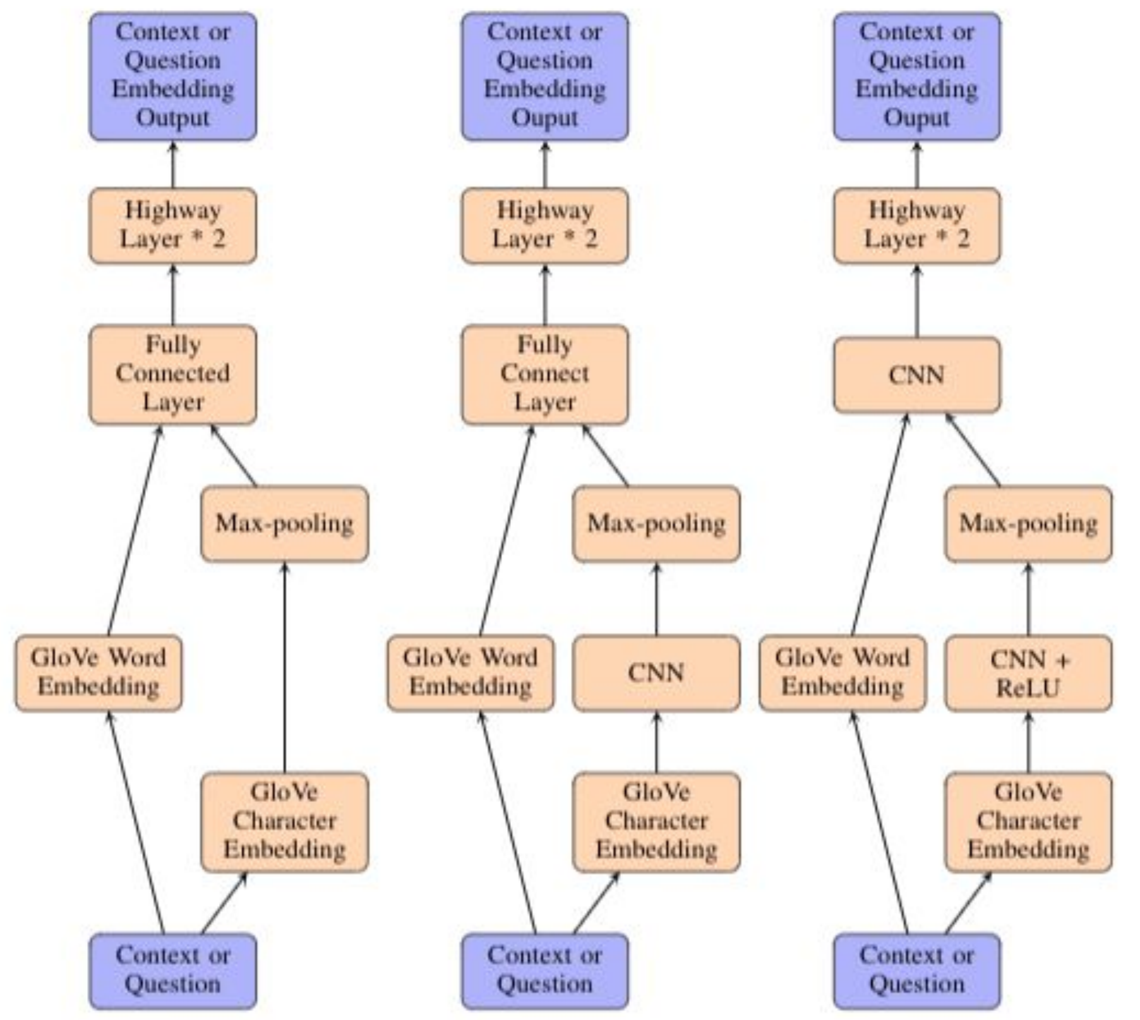
Implementation

Baseline Model: BiDAF without Char Embed

The baseline model has five block, Embedding Layer, Encoder Layer, Attention Layer, Model Layer and Output Layer.

- Embedding Layer: The Embedding Layer project the word embedding from GloVe embedding with a FC layer, and two Highway layers.
- Encoder Layer: The encoder layer uses a bidirectional LSTM without sharing parameters between left-to-right and right-to-left LSTM.
- Attention Layer: The two Attention layer allow in this part allow the attentions flow both way, Context-to-Question (C2Q) Attention and Question-to-Context(Q2C) Attention.
- Model Layer: The model layer have same structure as encoder layer, a bidirectional LSTM without sharing weights.
- Output Layer: The output layer applies a bidirectional LSTM to the modeling layer outputs, and Fully Connected layers with softmax to product distribution probability of answer start and end positions.

Model 1: BiDAF with Character Level Embedding



Bidirectional Attention Based Model Embedding Layer Flow Charts

The best BiDAF based model is using 2nd character level embedding structure. It reaches F1 score of 62.26 on development dataset with 20 epochs training.

Comparing with the 1st BiDAF with character level embedding models, the 2nd model has additional CNN layer on character level embedding.

Comparing with 3rd model, the 2ndmodel uses Fully Connected layer instead of CNN layer on merged embedding output of word and character level embedding, which have more parameters.

Model 2: QANet with Pre-trained Char Embed

Built on Transformer, QANet used only CNN and self-attention layer in order to train parallel on GPU and obtain much faster training process.

The QANet based model used in this project are slightly different from the implementation in original paper. The original QANet used random initialization 200 dimension character level embedding, while this model uses pretrained 300 dimension GloVe character level embedding.

Model 3: Bert

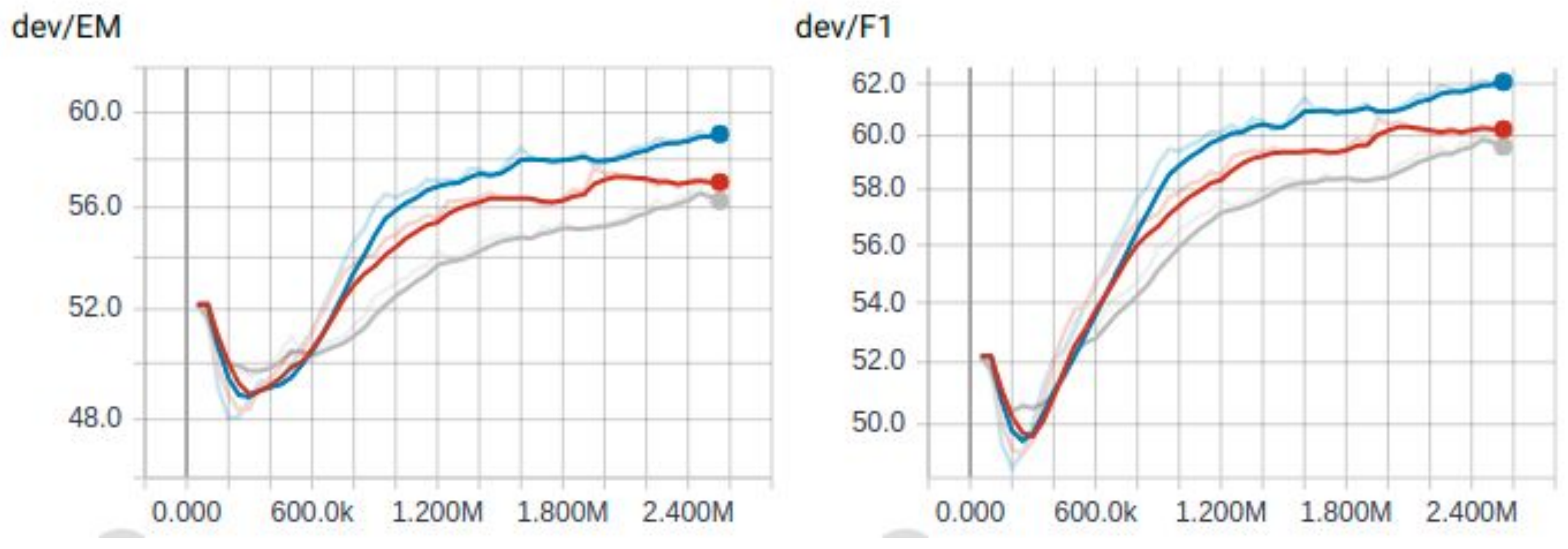
BERT base model is a fine tuning implementation.

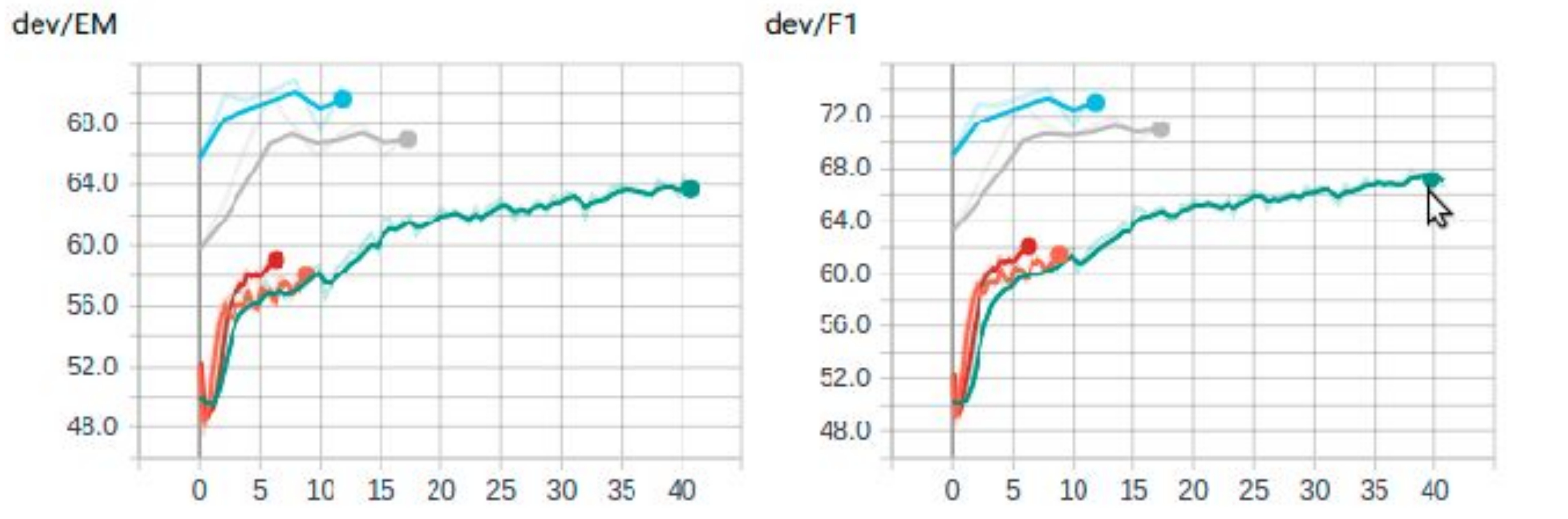
In addition, BERT with additional layer model, a model with additional layers specific for SQuAD has been trained and compares. A Fully Connected layer, ReLU activation, and dropout layers have been added before question answering output linear layer that calculating start and end position probability distribution.

Comparison of Results

Largest pretrained contextual embedding model, a BERT based model with additional layers for SQuAD, provides best performance among the models that have been tested., F1 score of 73.77 and 72.95 on development and test sets. Median size model QANet provides second best and BiDAF based have lowest performances. But BERT take very long time to train for each epoch, and reach performance cap within 3 epoch. On the other hand, Non-PCE models have lower initial performance, but can improve with more train epochs.

Using pretrained character level embedding, regularization improves model performance slightly, but not a significant amount. Model size and the depth are





Model	Epochs	Train Time	Dev EM	Dev F1
Baseline	30	9	58.58	61.88
Baseline with L2	30	7.5	57.47	60.78
BiDAF with 1st char embed	20	4.5	57.64	60.63
BiDAF with 2nd char embed	20	6	59.23	62.26
BiDAF with 3rd char embe	20	5	57.03	60.27
QANet with GloVe char embed	30	31	64.24	67.84
Bert Base	3	17	69.84	73.09
Bert with add. layers	3	11	70.32	73.77