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.

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.

Model 1: BiDAF with Character Level Embedding

The best BiDAF based model is using 2nd character level embedding structure. It reaches F1 score of 62.29 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 2nd model 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.

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.


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 50, which should not impact the model at all.

Model 2: QANet with Pre-trained Char Embed

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.