Motivation
Sentence completion is critical for construction of language models. High quality cloze-style dataset with multiple sentences and multiple blanks only came into existence in 2018. An accurate blank filling model not only demonstrates capability of advanced language models, but will also help OCR tasks and handwriting recognition that benefit from longer context. We reworked existing models including ELMo and BERT then analyzed performance for various scenarios.

Dataset
The CLOTH dataset [1] is a collection of cloze tests containing 7,131 articles and 99,433 questions. These samples were collected from online resources for English examinations in China. Articles are split using 70:15:15 ratio.

Approach
We first dissected and reproduced the architecture of an adapted BERT model [2], developed by authors of the CLOTH dataset. This model uses a pre-trained BERT encoder, where each word is converted to 3-length token. The decoder is a fully-connected linear layers which turns the 768-length embedding vector into a 30552 long vector, corresponding to the vocabulary space.

The BERT tokenizer was converted to generate character-level encoding for ELMo [3]. Specifically, the character embedding is context aware and maps to a 50 dimensional space. At the encoder level, ELMo outputs a 1024-length embedding vector. At the decoder level, we again used linear layers to turn the embedding into 28k long decoded vectors representing the ELMo vocabulary space. A bidirectional TCN model was created as an attempt to replace LSTM structure used for ELMo. All other components of the architecture are kept the same. This model was trained using the same method as BERT.

Results and Discussion
It comes with no surprise that BERT models are the highest performing because BERT was trained to recover masked tokens. The ELMo model is on-par with previous baseline, while identifying answer through sentence classification exceeded baselines by 5%. Although additional datasets were used, overfitting was still present for all non-BERT models, suggesting that alternative training methods should be considered.

Learning rate scheduling is also important. As shown in Figure 3, jumps in accuracy were observed when learning rate stepped down. Similar effects were also observed for training TCN and BERT models.

Future Works
Overfitting of ELMo and TCN models can be alleviated by introducing more datasets. These two models should be trained on single-sentence completion before fine-tuning on CLOTH dataset. Given the superior performance of BERT, difficult cloze questions can be generated by interpreting BERT output. It is worth investigating human performance on BERT-generated articles.

References