Summery of "SpanBERT: Improving Pre-training by Representing and Predicting Spans"

Suvadeep Hajra

November 11, 2020

1 Contribution

The paper has proposed SpanBERT, a pre-training method that is designed to better represent and predict spans of text. SpanBERT extends BERT by (1) masking contiguous random spans, rather than individual random tokens, and (2) training the span boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it.

SpanBERT consistently outperforms BERT and some better-tuned baselines, with substantial gains on span selection tasks such as question answering and coreference resolution.

2 Differences with BERT

SpanBERT differs from BERT in two ways:

1. In SpanBERT, random contiguous spans of tokens are masked instead of random individual tokens.
2. During pre-training, a novel span-boundary objective is minimized.

The span masking process and span-boundary objective are illustrated in Figure 1.

3 Experimental Results

SpanBERT is evaluated on three tasks which require explicit modeling of the token span representation and on GLUE benchmark tasks. The tasks which require explicit token span modeling are: extractive question answering, coreference resolution and relation extraction. In all the benchmarks for the three tasks, SpanBERT has shown significant improvement over BERT baselines.

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$

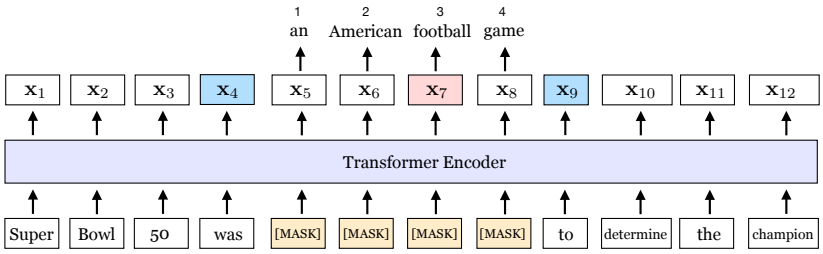

Figure 1: An illustration of SpanBERT training. The span *an American football game* is masked. The span boundary objective (SBO) uses the output representations of the boundary tokens, $\mathbf{x}_4$ and $\mathbf{x}_9$ (in blue), to predict each token in the masked span. The equation shows the MLM and SBO loss terms for predicting the token, *football* (in pink), which as marked by the position embedding $\mathbf{p}_3$, is the *third* token from x_4.

Figure 1: Span masking and span-boundary objective

Tasks on GLUE benchmark do not involve explicit token span modeling. SpanBERT has shown similar or marginally better performance on most of the GLUE tasks compared to the BERT baselines.