Using the Transformer

BERT – Shortcommings / Critique

Learning goals

- Problem with the [MASK] token
- Inter-token dependencies

PRETRAIN-FINETUNE DISCREPANCY

- BERT *artificially* introduces `[MASK]` tokens during pre-training
- `[MASK]`-token does not occur during fine-tuning
  $\rightarrow$ Lacks the ability to model joint probabilities
  $\rightarrow$ Assumes independence of predicted tokens (given the context)

INDEPENDENCE ASSUMPTION

`[MASK]-ing procedure:`

- "Given a sentence, predict `[MASK]`ed tokens"
- All `[MASK]`ed tokens are predicted based on the un-`[MASK]`ed tokens
- *Implicit assumption:* Independence of `[MASK]`ed tokens

$$\mathcal{J}_{\text{BERT}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{is a city}),$$

$$\mathcal{J}_{\text{XLNet}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{New}, \text{is a city})$$

Prediction of [New, York] given the factorization order [is, a, city, New, York]

Source: Yang et al. (2019)

MAXIMUM SEQUENCE LENGTH

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

not cool

cool

Source: Vaswani et al. (2017)

Limitation:

- BERT can only consume sequences of up to 512 tokens
- Two sentences for NSP are sampled such that

$$length_{sentenceA} + length_{sentenceB} \leq 512$$

- Reason: Computational complexity of Transformer scales quadratically with the sequence length
  $\rightarrow$ Longer sequences are disproportionally expensive

BIAS