Contextual word representations: BERT

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding
Core model structure

Transformer blocks

Model structure  MLM  MLM loss  Next Sentence Prediction  Fine-tuning  Tokenization  Model releases  Limitations
Masked Language Modeling (MLM)

Transformer blocks

masking: none
Masked Language Modeling (MLM)

Masking: [MASK]

Transformer blocks

CLSim, a_in, b_in, c_in, S_in

CLSout, a_out, b_out, c_out, S_out
Masked Language Modeling (MLM)

Transformer blocks

masking: random word
MLM loss function

For Transformer parameters $H_\theta$ and sequence $\mathbf{x} = [x_1, \ldots, x_T]$ with masked version $\hat{\mathbf{x}}$:

$$\max_\theta \sum_{t=1}^T m_t \log \frac{\exp (e(x_t)^T H_\theta(\hat{\mathbf{x}})_t)}{\sum_{x' \in \mathcal{V}} \exp (e(x')^T H_\theta(\hat{\mathbf{x}})_t)}$$

where $\mathcal{V}$ is the vocabulary, $x_t$ is the actual token at step $t$, $m_t = 1$ if token $t$ was masked, else 0, and $e(x)$ is the embedding for $x$. 
Binary next sentence prediction pretraining

Positive: Actual sentence sequences
- [CLS] the man went to [MASK] store [SEP]
- he bought a gallon [MASK] milk [SEP]
- Label: IsNext

Negative: Randomly chosen second sentence
- [CLS] the man went to [MASK] store [SEP]
- penguin [MASK] are flight ##less birds [SEP]
- Label: NotNext
Transfer learning and fine-tuning

Transformer blocks
Tokenization and the BERT embedding space

[1]: `from transformers import BertTokenizer`

[2]: `tokenizer = BertTokenizer.from_pretrained('bert-base-cased')`

[3]: `tokenizer.tokenize("This isn't too surprising.")`

[3]: `['This', 'isn', '', 't', 'too', 'surprising', '']`

[4]: `tokenizer.tokenize("Encode me!")`

[4]: `['En', '##code', 'me', '!']`

[5]: `tokenizer.tokenize("Snuffleupagus?")`

[5]: `['S', '##nu', '##ffle', '##up', '##agu', '##s', '?']`

[6]: `tokenizer.vocab_size`

[6]: `28996`
Initial BERT model releases

Base

- Transformer layers: 12
- Hidden representations: 768 dimensions
- Attention heads: 12
- Total parameters: 110M

Large

- Transformer layers: 24
- Hidden representations: 1024 dimensions
- Attention heads: 16
- Total parameters: 340M

Limited to sequences of 512 tokens due to dimensionality of the positional embeddings.

Many new releases at the project site and on Hugging Face.
Known limitations with BERT

1. Devlin et al. (2019:§5): admirably detailed but still partial ablation studies and optimization studies.

2. Devlin et al. (2019): “The first [downside] is that we are creating a mismatch between pre-training and fine-tuning, since the [MASK] token is never seen during fine-tuning.”

3. Devlin et al. (2019): “The second downside of using an MLM is that only 15% of tokens are predicted in each batch”

4. Yang et al. (2019): “BERT assumes the predicted tokens are independent of each other given the unmasked tokens, which is oversimplified as high-order, long-range dependency is prevalent in natural language”