Assignment 2: BERT

1 Abstract

ImageNet [1] moment.

2 Components

Explain BERT components.

2.1 Byte Pair Encoding (BPE)

Explain BPE.
Diagram

2.2 Attention

Explain Attention [2].
Diagram
Explain self-attention.

2.3 Transformer

Explain Transformer [3].
Diagram

3 BERT

Explain BERT [4].
Diagram
Explain scale: size, composition, features, memory size, number of parameters and transformers, inputs and outputs, etc.

3.1 Corpora

Explain Wikipedia [5] and BookCorpus [6].

3.2 Training

Explain training process.

3.2.1 Masked Language Model (MLM)

Explain MLM.
Diagram

3.2.2 Next Sentence Prediction (NSP)

Explain NSP.
Diagram

3.3 Contextualization

Explain contextualization.
Diagram of same words with different tokens.

3.4 Transfer Learning

Explain transfer learning.

3.4.1 Fine-Tuning

Explain how fine-tuning BERT works.

3.4.2 Downstream Tasks

Explain most common downstream tasks.

4 Domain-Specific BERT

Explain domain-specific models.

4.1 BioBERT

Explain BioBERT [7].

4.2 SciBERT

Explain SciBERT [8].

5 Related Work

Explain Related Work.

5.1 ELMo

Explain ELMo [9].

5.2 ULMFiT

Explain ULMFiT [10].

5.3 GPT-2

Explain GPT-2 [11].

6 Implications

Explain implications.
Explain SOTA.

7 References

[1] J. Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. URL: https://ieeexplore.ieee.org/document/5206848.

[2] Andrea Galassi, Marco Lippi, and Paolo Torroni. "Attention in Natural Language Processing". In: (2019). DOI: 10.1109/TNNLS.2020.3019893. arXiv: 1902.02181. URL: https://arxiv.org/pdf/1902.02181.pdf.

[3] Ashish Vaswani et al. "Attention Is All You Need". In: (2017). arXiv: 1706.03762. URL: https://arxiv.org/abs/1706.03762.

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: (2018). arXiv: 1810.04805. URL: https://arxiv.org/pdf/1810.04805.pdf.

[5] Wikipedia. *Wikipedia*. Jan. 2001. URL: https://en.wikipedia.org/wiki/Main_Page.

[6] Yukun Zhu et al. "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books". In: (2015). arXiv: 1506.06724. URL: https://arxiv.org/pdf/1506.06724.pdf.

[7] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: (2019). DOI: 10.1093/bioinformatics/btz682. eprint: arXiv:1901.08746. URL: https://arxiv.org/pdf/1901.08746.pdf.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". In: (2019). eprint: arXiv:1903.10676. URL: https://arxiv.org/pdf/1903.10676.pdf.

[9] Matthew E. Peters et al. "Deep contextualized word representations". In: (2018). arXiv: 1802.05365. URL: https://arxiv.org/pdf/1802.05365.pdf.

[10] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: (2018). arXiv: 1801.06146. URL:

https://arxiv.org/pdf/1801.06146.pdf.

[11] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).