Machine Learning Laboratory

Session 8 – Transformer

Theoretical Preparation

Transformers have revolutionized natural language processing (NLP) and beyond. Their importance lies in their unique architecture that utilizes self-attention mechanisms, enabling them to process entire sequences of data simultaneously. This feature makes them exceptionally good at capturing complex relationships and dependencies in data, unlike previous models that processed data sequentially.

- **Preparation TODO**
  - Ashish Vaswani et al.: Attention Is All You Need. `https://arxiv.org/pdf/1706.03762`.
  - Alammar, J (2018). The Illustrated Transformer [Blog post]. `https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-`
  - Alammar, J (2018). The Illustrated Transformer [Blog post]. `https://jalammar.github.io/illustrated-transformer/`.
  - Sebastian Raschka. Understanding and Coding the Self-Attention Mechanism of Large Language Models From Scratch [Blog Post]. `https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html`.
  - Amirhossein Kazemnejad. Transformer Architecture: The Positional Encoding [Blog Post]. `https://kazemnejad.com/blog/transformer_architecture_positional_encoding/`.
  - Gary's Notebook. How Exactly Does Masking in Transformer Work [Blog Post]. `https://www.garysnotebook.com/20210128_1`.
- **Optional**
  - Andrej Karpathy. Let's build GPT: from scratch, in code, spelled out [Video]. `https://www.youtube.com/watch?v=kCc8FmEb1nY`.