Exercises

- Explain the difference between self-attention and cross-attention in the context of transformers. How does each mechanism contribute to an encoder-decoder architecture, and why is cross-attention necessary in models?
- In transformer models, layer normalization is applied to stabilize training. Explain the mechanics of layer normalization in the transformer context, and discuss why it might be less effective or problematic if applied after, rather than before, the residual connection.
- What is the difference between encoder-only, decoder-only, and encoder-decoder transformers?
- Discuss the primary reasons why transformers have largely replaced RNNs in natural language processing tasks.
- Why was it necessary to have contextual embedding models over static embedding models?
- What are the main differences between a BERT model and word2vec model in terms of input, output, tokenization and downstream tasks?