Exercises

- What are the query (Q), key (K), and value (V) matrices in self-attention, and how are they computed from input embeddings?
- Why do we use three separate linear transformations instead of directly computing attention on the raw embeddings?
- Why is the dot product used to measure similarity between queries and keys? Could other similarity measures (e.g., cosine similarity) be used?
- Why do we divide the dot product sqrt{d}?
- What would happen to the gradients or attention weights if this scaling factor were removed?
- Since self-attention has no inherent notion of word order, how does the Transformer model encode positional information?