AML Mock Exam

Lena Jäger, Deborah Jakobi, David Reich

May 2025

1 Relative Positional Encoding Linear Projection

In the original Transformer, the sinusoidal positional encoding at position t and dimension i is given by

$$\mathrm{PE}_{t,2k} = \sin(\omega_k\, t), \quad \mathrm{PE}_{t,2k+1} = \cos(\omega_k\, t), \quad \omega_k = \frac{1}{10000^{2k/d_{\mathrm{model}}}}.$$

some math

Show that for any fixed offset k, there exists a constant 2×2 matrix $M_{\omega,k}$ such that

$$M_{\omega,k} \begin{pmatrix} \sin(\omega\, t) \\ \cos(\omega\, t) \end{pmatrix} = \begin{pmatrix} \sin\big(\omega\,(t+k)\big) \\ \cos\big(\omega\,(t+k)\big) \end{pmatrix} \quad \forall\, t.$$

1. Derive $M_{\omega,k}$ explicitly using trigonometric addition theorems.
2. Show that $M_{\omega,k}$ is a rotation matrix independent of t.
3. Explain briefly why this property allows the Transformer to learn relative position biases via a single learned linear map per offset.

2 Complexity and Path Length

Compare a single Transformer self-attention layer with sequence length T to a single-layer unidirectional RNN in terms of:

(a) Computational complexity per layer as a function of T and d_{model}.

(b) Maximum "path length" (i.e. number of sequential nonlinearities) connecting any two input positions.

Provide Big-$\mathcal{O}$ expressions and briefly interpret why self-attention can better capture long-range dependencies.

3 Masked Language Modeling (MLM)

1. Conceptually, what is masked language modeling (MLM), and what problem does it try to solve?

2. In what types of models is MLM typically used?

4 Retrieval-Augmented Generation

1. Describe a concrete scenario where retrieval augmentation would significantly improve model performance compared to a standard language model without retrieval.

2. Explain why retrieval is beneficial in this scenario. Discuss what limitations it helps to overcome.

3. Briefly describe how retrieval is integrated into the model architecture (e.g., RAG or similar approaches).

5 Graph Neural Networks by Hand

Consider the 4-node undirected graph with adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix},$$

input feature matrix

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \text{and weight matrix} \quad W = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Assume the graph convolution operation

$$H = \sigma(\hat{A}XW), \quad \text{where} \quad \hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}, \quad \tilde{A} = A + I,$$

and $\sigma(z)$ is the ReLU activation function. Compute the following:

1. The normalized adjacency matrix $\hat{A}$, showing all intermediate steps. It is enough to provide fractions.

2. Given **pre-activation** $Z = \hat{A}\,X\,W$:

$$XW = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 4 & 6 \\ 0 & 0 \end{pmatrix}, \quad Z = \hat{A}\,(XW) \approx \begin{pmatrix} 2.121 & 2.828 \\ 3.061 & 4.582 \\ 2.154 & 3.077 \\ 2.154 & 3.077 \end{pmatrix}.$$

The final output H after applying the ReLU nonlinearity.

6 Generative Modeling

Explain the reparametrization trick:

1. Explain the purpose of the reparameterization trick in the context of training generative models (for example Diffusion models). Why is it necessary for gradient-based optimization? Describe the conceptual idea behind the trick and how it allows backpropagation through stochastic nodes.

2. Provide and explain the mathematical formulation of the reparameterization trick for a Gaussian latent variable.

7 State-Space Models CNNs and RNNs

State-space models can be interpreted through the lens of recurrent neural networks (RNNs), particularly when unrolled over time.

1. Explain the difference between the continuous, recurrent and convolutional views of an SSM. What is each view used for? Why?

2. Draw a schematic diagram showing the RNN-like unrolled view of a state-space model, including the hidden states, observations, and transitions over time.

3. Clearly label the latent state variables z_t, observed variables x_t, and indicate the dependencies between them.

4. From the recurrent view, derive the convolutional view of an SSM.