Deep Learning Winter 2025–2026

Homework 3 — Sequence Models

Naaman Kopty naamankoptyta@gmail.com

Due: 07.01.2026

Questions 1–2: (70 Points)

Programming assignment — See the attached Jupyter notebook on Moodle.

Question 3: Attention Exploration (20 Points)

Self-attention is a central component of the Transformer architecture. In this question, we explore several mathematical properties of the scaled dot-product self-attention mechanism and motivate its behavior.

Recall that attention can be viewed as an operation on a query vector $q \in \mathbb{R}^d$, a set of key vectors $\{k_1, \ldots, k_n\}$, $k_i \in \mathbb{R}^d$, and a corresponding set of value vectors $\{v_1, \ldots, v_n\}$, $v_i \in \mathbb{R}^d$, defined as:

$$c = \sum_{i=1}^{n} v_i \alpha_i, \tag{1}$$

$$\alpha_i = \frac{\exp(k_i^\top q)}{\sum_{j=1}^{n} \exp(k_j^\top q)}, \tag{2}$$

where $\alpha = \{\alpha_1, \ldots, \alpha_n\}$ are referred to as the *attention weights*. Observe that the output $c \in \mathbb{R}^d$ is a weighted average of the value vectors.

(a) **Copying in attention.** One advantage of attention mechanisms is their ability to "copy" information from specific value vectors into the output.

(i) Describe a condition under which the categorical distribution defined by α concentrates almost all of its mass on a single index j (i.e., $\alpha_j \gg \sum_{i \neq j} \alpha_i$). What must be true about the query q and the keys $\{k_1, \ldots, k_n\}$?

(ii) Under the condition described above, describe the resulting output vector c.

(b) **An average of two.** Instead of focusing on a single value vector, attention may need to incorporate information from multiple sources.

Assume that the key vectors are orthogonal and have unit norm, i.e.,

$$k_i^\top k_j = 0 \quad \text{for } i \neq j, \qquad \|k_i\| = 1.$$

Consider two value vectors v_a and v_b with corresponding keys k_a and k_b.

Find an expression for a query vector q such that the attention output satisfies

$$c \approx \tfrac{1}{2}(v_a + v_b),$$

and briefly justify why your choice of q produces this behavior.

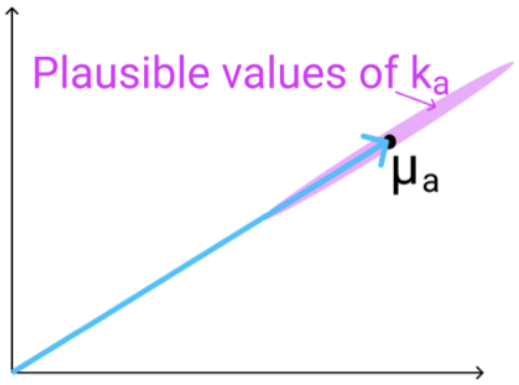

Figure 1: Illustration of a mean key vector μ_a (shown here in 2D for clarity) and representative plausible realizations of the corresponding key vector k_a. The key vectors remain approximately aligned with μ_a but may differ in magnitude.

(c) **Sensitivity of single-headed attention.** In practice, attention mechanisms operate on key vectors that may be subject to small perturbations.

Assume that the keys $\{k_1, \ldots, k_n\}$ are random samples drawn from distributions

$$k_i \sim \mathcal{N}(\mu_i, \Sigma_i),$$

where the means $\mu_i \in \mathbb{R}^d$ are known, mutually orthogonal, and satisfy $\|\mu_i\| = 1$. Assume further that the covariance matrices are isotropic and small,

$$\Sigma_i = \alpha I \quad \text{for } \alpha \to 0,$$

so that each k_i remains approximately aligned with its mean direction μ_i, but may vary in magnitude (see Figure 1).

Explain how one can choose a query vector q in terms of the means $\{\mu_i\}$ such that the attention output approximately averages two desired value vectors, as in part (b). Your answer should be qualitative but grounded in the geometry of the dot-product attention mechanism, and should focus on directional alignment and relative magnitudes rather than probabilistic variance analysis.

Question 4: Position Embeddings Exploration (10 Points)

Position embeddings are an important component of the Transformer architecture, allowing the model to differentiate between tokens based on their position in the sequence. In this question, we explore why positional information is needed and how sinusoidal position embeddings address this issue.

Consider a simplified Transformer block applied to an input sequence embedding matrix $X \in \mathbb{R}^{T\times d}$, where T is the sequence length and d is the embedding dimension. The self-attention layer is defined as:

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V,$$

$$H = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$, and $H \in \mathbb{R}^{T\times d}$. Next, the feed-forward layer applies:

$$Z = \text{ReLU}(HW_1 + \mathbf{1}b_1)\, W_2 + \mathbf{1}b_2,$$

where $W_1, W_2 \in \mathbb{R}^{d\times d}$, $b_1, b_2 \in \mathbb{R}^{1\times d}$, and $\mathbf{1} \in \mathbb{R}^{T\times 1}$ is an all-ones vector.

(a) **Permutation equivariance without positional information.** Let $P \in \mathbb{R}^{T\times T}$ be a permutation matrix, and define the permuted input

$$X_{\text{perm}} = PX.$$

Show that the corresponding output satisfies

$$Z_{\text{perm}} = PZ.$$

You may use the following facts (without proof): for any permutation matrix P and any matrix A,

$$\text{softmax}(PAP^\top) = P\,\text{softmax}(A)\,P^\top, \qquad \text{ReLU}(PA) = P\,\text{ReLU}(A).$$

Finally, explain in one or two sentences why this property is problematic for processing natural language.

(b) **Sinusoidal position embeddings.** A common approach is to add deterministic position embeddings $\Phi \in \mathbb{R}^{T\times d}$ to the input embeddings:

$$X_{\text{pos}} = X + \Phi.$$

For $t \in \{0, 1, \ldots, T-1\}$ and $i \in \{0, 1, \ldots, \frac{d}{2} - 1\}$, the sinusoidal embeddings are defined by:

$$\Phi_{t,2i} = \sin\left(\frac{t}{10000^{2i/d}}\right), \qquad \Phi_{t,2i+1} = \cos\left(\frac{t}{10000^{2i/d}}\right).$$

(i) Do sinusoidal position embeddings resolve the issue identified in part (a)? Explain briefly.

(ii) Can it happen that $\Phi_t = \Phi_{t'}$ (i.e., the position-embedding vectors for two different positions $t \neq t'$ are identical)? Answer yes or no, and justify your answer. If you claim yes, provide an explicit example; if you claim no, provide a clear argument.

Good Luck!