Deep Learning Winter 2025–2026
Homework 3 — Sequence Models
Solutions

Mahmnoud Abade, 206773756 —— Firas Dwere, 214225021

January 15, 2026

Question 3: Attention Exploration (20 Points)

Part (a): Copying in attention

(i) Condition for concentration

For the categorical distribution defined by α to concentrate almost all of its mass on a single index j (i.e., $\alpha_j \gg \sum_{i \neq j} \alpha_i$), the following condition must hold:

Condition: The query vector q must be highly aligned with key k_j and poorly aligned with all other keys $\{k_i\}_{i \neq j}$.

Mathematically, this requires:

$$k_j^\top q \gg k_i^\top q \quad \text{for all } i \neq j \tag{1}$$

Since the attention weights are given by:

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

When $k_j^\top q \gg k_i^\top q$ for all $i \neq j$, the exponential amplifies this difference, causing $\exp(k_j^\top q)$ to dominate the denominator. This makes $\alpha_j \approx 1$ and $\alpha_i \approx 0$ for $i \neq j$.

In geometric terms, q should point in nearly the same direction as k_j while being orthogonal or pointing away from the other keys.

(ii) Resulting output vector

Under the condition described above, where $\alpha_j \approx 1$ and $\alpha_i \approx 0$ for all $i \neq j$, the output vector becomes:

$$c = \sum_{i=1}^{n} v_i \alpha_i \tag{2}$$

$$\approx v_j \cdot 1 + \sum_{i \neq j} v_i \cdot 0 \tag{3}$$

$$\approx v_j \tag{4}$$

Conclusion: The output vector c approximately equals the value vector v_j. This is the "copying" behavior — the attention mechanism extracts and copies the specific value vector corresponding to the key most aligned with the query.

Part (b): An average of two

Given that key vectors are orthogonal and have unit norm ($k_i^\top k_j = 0$ for $i \neq j$, $\|k_i\| = 1$), we want to find a query vector q such that $c \approx \frac{1}{2}(v_a + v_b)$.

Solution: Choose the query vector as:

$$q = \frac{1}{\sqrt{2}}(k_a + k_b) \tag{5}$$

Justification:

For this choice of q, we compute the dot products:

$$k_a^\top q = k_a^\top \cdot \frac{1}{\sqrt{2}}(k_a + k_b) = \frac{1}{\sqrt{2}}(k_a^\top k_a + k_a^\top k_b) = \frac{1}{\sqrt{2}}(1 + 0) = \frac{1}{\sqrt{2}} \tag{6}$$

$$k_b^\top q = k_b^\top \cdot \frac{1}{\sqrt{2}}(k_a + k_b) = \frac{1}{\sqrt{2}}(k_b^\top k_a + k_b^\top k_b) = \frac{1}{\sqrt{2}}(0 + 1) = \frac{1}{\sqrt{2}} \tag{7}$$

$$k_i^\top q = \frac{1}{\sqrt{2}}(k_i^\top k_a + k_i^\top k_b) = \frac{1}{\sqrt{2}}(0 + 0) = 0 \quad \text{for } i \neq a, b \tag{8}$$

Therefore, the attention weights become:

$$\alpha_a = \frac{\exp(1/\sqrt{2})}{\exp(1/\sqrt{2}) + \exp(1/\sqrt{2}) + \sum_{i \neq a,b} \exp(0)} = \frac{\exp(1/\sqrt{2})}{2\exp(1/\sqrt{2}) + (n-2)} \tag{9}$$

$$\alpha_b = \frac{\exp(1/\sqrt{2})}{2\exp(1/\sqrt{2}) + (n-2)} \tag{10}$$

For large values of $\exp(1/\sqrt{2}) \approx 2.03$, we have $\alpha_a \approx \alpha_b \approx \frac{1}{2}$ and $\alpha_i \approx 0$ for $i \neq a, b$. Thus:

$$c = \sum_{i=1}^{n} v_i \alpha_i \approx v_a \cdot \frac{1}{2} + v_b \cdot \frac{1}{2} = \frac{1}{2}(v_a + v_b)$$

Part (c): Sensitivity of single-headed attention

Given that keys are sampled from $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ where $\Sigma_i = \alpha I$ for $\alpha \to 0$, and the means μ_i are mutually orthogonal with $\|\mu_i\| = 1$, we want to choose q to average two desired value vectors.

Solution:

To approximately average value vectors v_a and v_b, choose the query vector as:

$$q = \beta(\mu_a + \mu_b) \tag{11}$$

where $\beta > 0$ is a scaling factor (e.g., $\beta = 1/\sqrt{2}$ for normalization).

Geometric explanation:

Since $\alpha \to 0$, each key k_i remains tightly concentrated around its mean μ_i. The key insight is that the dot product $k_i^\top q$ is determined primarily by the alignment between the *directions* of k_i and q.

For our choice of $q = \beta(\mu_a + \mu_b)$:

- The expected dot products are $\mathbb{E}[k_a^\top q] = \beta\mu_a^\top(\mu_a + \mu_b) = \beta$ and $\mathbb{E}[k_b^\top q] = \beta$, since $\mu_a^\top \mu_b = 0$ and $\|\mu_a\| = 1$.
- For $i \neq a, b$: $\mathbb{E}[k_i^\top q] = \beta\mu_i^\top(\mu_a + \mu_b) = 0$ due to orthogonality.
- Since the variance is small ($\alpha \to 0$), actual values $k_i^\top q$ concentrate tightly around their expectations.

Thus, $k_a^\top q \approx k_b^\top q \approx \beta \gg k_i^\top q \approx 0$ for $i \neq a, b$.

This produces attention weights $\alpha_a \approx \alpha_b \approx 1/2$ and $\alpha_i \approx 0$ for $i \neq a, b$, yielding:

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

The robustness comes from aligning q with the *mean directions* μ_a and μ_b rather than individual samples, exploiting the geometric structure of the key space.

Question 4: Position Embeddings Exploration (10 Points)

Part (a): Permutation equivariance without positional information

To show: $Z_{\text{perm}} = PZ$ where $X_{\text{perm}} = PX$.

Proof:

Starting with the permuted input $X_{\text{perm}} = PX$:

$$Q_{\text{perm}} = X_{\text{perm}} W_Q = (PX)W_Q = P(XW_Q) = PQ \tag{12}$$
$$K_{\text{perm}} = X_{\text{perm}} W_K = PK \tag{13}$$
$$V_{\text{perm}} = X_{\text{perm}} W_V = PV \tag{14}$$

For the attention layer:

$$H_{\text{perm}} = \text{softmax}\left(\frac{Q_{\text{perm}} K_{\text{perm}}^\top}{\sqrt{d}}\right) V_{\text{perm}} \tag{15}$$
$$= \text{softmax}\left(\frac{(PQ)(PK)^\top}{\sqrt{d}}\right)(PV) \tag{16}$$
$$= \text{softmax}\left(\frac{PQK^\top P^\top}{\sqrt{d}}\right)(PV) \tag{17}$$
$$= \text{softmax}\left(P\frac{QK^\top}{\sqrt{d}}P^\top\right)(PV) \tag{18}$$
$$= P\text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)P^\top(PV) \quad \text{(using given property)} \tag{19}$$
$$= P\text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (\text{since } P^\top P = I) \tag{20}$$
$$= PH \tag{21}$$

For the feed-forward layer:

$$Z_{\text{perm}} = \text{ReLU}(H_{\text{perm}} W_1 + \mathbb{1}b_1)W_2 + \mathbb{1}b_2 \tag{22}$$
$$= \text{ReLU}((PH)W_1 + \mathbb{1}b_1)W_2 + \mathbb{1}b_2 \tag{23}$$
$$= \text{ReLU}(P(HW_1) + \mathbb{1}b_1)W_2 + \mathbb{1}b_2 \tag{24}$$
$$= \text{ReLU}(P(HW_1 + \mathbb{1}b_1))W_2 + \mathbb{1}b_2 \quad (\text{since } P\mathbb{1} = \mathbb{1}) \tag{25}$$
$$= P\text{ReLU}(HW_1 + \mathbb{1}b_1)W_2 + \mathbb{1}b_2 \quad \text{(using given property)} \tag{26}$$
$$= P(\text{ReLU}(HW_1 + \mathbb{1}b_1)W_2) + \mathbb{1}b_2 \tag{27}$$
$$= P(\text{ReLU}(HW_1 + \mathbb{1}b_1)W_2 + \mathbb{1}b_2) \quad (\text{since } P\mathbb{1} = \mathbb{1}) \tag{28}$$
$$= PZ \tag{29}$$

Why this is problematic: This property means that the Transformer output is invariant to the order of input tokens — permuting the input simply permutes the output in the

same way. For natural language processing, word order is crucial for meaning (e.g., "dog bites man" vs. "man bites dog"). Without positional information, the model cannot distinguish between different orderings of the same words, making it impossible to understand syntax, grammar, or context-dependent semantics.

Part (b): Sinusoidal position embeddings

(i) Resolution of permutation equivariance

Yes, sinusoidal position embeddings resolve the issue.

When we use $X_{\text{pos}} = X + \Phi$, the position embeddings Φ are *not* permuted when we permute the input tokens. Specifically, if we permute the token embeddings to get $X_{\text{perm}} = PX$, the corresponding position-aware input becomes:

$$X_{\text{perm}}^{\text{pos}} = PX + \Phi \neq P(X + \Phi) = PX_{\text{pos}}$$

The position embeddings Φ remain tied to the *absolute positions* in the sequence, not to the token identities. Therefore, permuting tokens changes their relationship to the fixed positional structure, breaking the permutation equivariance. The model can now distinguish between different orderings of the same tokens.

(ii) Uniqueness of position embeddings

No, it cannot happen that $\Phi_t = \Phi_{t'}$ for $t \neq t'$.

Argument:

The position embedding vector $\Phi_t \in \mathbb{R}^d$ has components:

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

for $i \in \{0, 1, \ldots, d/2 - 1\}$.

Consider the first two dimensions ($i = 0$):

$$\Phi_{t,0} = \sin(t), \quad \Phi_{t,1} = \cos(t)$$

If $\Phi_t = \Phi_{t'}$, then in particular $\Phi_{t,0} = \Phi_{t',0}$ and $\Phi_{t,1} = \Phi_{t',1}$, which means:

$$\sin(t) = \sin(t') \quad \text{and} \quad \cos(t) = \cos(t')$$

This system of equations implies $t = t' + 2\pi k$ for some integer k.

For typical sequence lengths in practice (e.g., $T < 10000$), we have $0 \leq t, t' < T < 2\pi \approx 6.28$. Therefore, if $t \neq t'$ and both are in $[0, T-1]$, there is no integer $k \neq 0$ such that $t = t' + 2\pi k$.

Thus, the position embeddings are unique for all positions within any practical sequence length.