Assignment 2: Transformers

Ohad Kiperman, Oriya Sheetrit , Yarden Cohen , Noam Klainer

December 24, 2024

1 Attention Exploration

1.1 Copying in Attention

(a)

The attention weights α are computed using the softmax function:

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

where k_i are the keys, and q is the query. α can be interpreted as a categorical probability distribution because:

- **Non-negativity:** The exponential function ensures that $\exp(k_i^\top q) \geq 0$, so $\alpha_i \geq 0$.
- **Normalization:** The softmax function normalizes the weights such that $\sum_{i=1}^{n} \alpha_i = 1$.

Thus, α represents a categorical distribution over the indices $\{1, 2, \ldots, n\}$, where each α_i corresponds to the probability of selecting the i-th key-value pair.

(b)

If the key values k_j compared to other key values $k_{i \neq j}$ are large (i.e., $k_j \gg k_i$, for $i \in \{1, ..., n\}$ and $i \neq j$), then the dot product between the key and the query will be large. This will cause the softmax function to put most of its probability mass onto this large value.

(c)

The output c now will almost be a copy of the value v_j due to the fact that $\alpha_j \gg \sum_{i \neq j} \alpha_i$, i.e., $c \approx v_j$. This happens because the attention mechanism computes c as a weighted sum of the values v_i, where the weights α_i are determined by the similarity between the query q and each key k_i. When α_j dominates, it effectively ignores contributions from other keys, making the output c highly similar to v_j.

(d)

Intuitively, c is a weighted sum of the value vectors $\{v_1, \ldots, v_n\}$, but when the query q is closely aligned with a specific key k_j, the output c becomes dominated by the corresponding value v_j. This means the attention mechanism effectively focuses on the most relevant information and outputs it as c.

1.2 Average in Attention

(a)

Assume that c is approximated as follows:

$$c \approx 0.5v_a + 0.5v_b$$

This means that $\alpha_a \approx 0.5$ and $\alpha_b \approx 0.5$, which can be achieved when (whenever $i \neq a$ and $i \neq b$):

$$k_a^\top q \approx k_b^\top q \gg k_i^\top q$$

If q is constructed as:

$$q = \beta(k_a + k_b), \quad \text{where } \beta \gg 0$$

then, since the keys are orthogonal to each other, it is easy to see that:

$$k_a^\top q = \beta k_a^\top k_a + \beta k_a^\top k_b = \beta \times 1 + \beta \times 0 = \beta$$

$$k_b^\top q = \beta k_b^\top k_a + \beta k_b^\top k_b = \beta \times 0 + \beta \times 1 = \beta$$

$$k_i^\top q = \beta k_i^\top k_a + \beta k_i^\top k_b = \beta \times 0 + \beta \times 0 = 0, \quad \text{for } i \notin \{a, b\}$$

Thus, when we exponentiate, only $\exp(\beta)$ will matter, because $\exp(0) = 1$ will be insignificant to the probability mass. We get that:

$$\alpha_a = \alpha_b = \frac{\exp(\beta)}{n - 2 + 2\exp(\beta)} \approx \frac{\exp(\beta)}{2\exp(\beta)} \approx \frac{1}{2}, \quad \text{for } \beta \gg 0$$

1.3 Permutation Equivariance in Self-Attention

1. A permutation matrix Π rearranges the rows of the input X. If X is the input, then $X\Pi$ represents the permuted input.

2. For the permuted input $X\Pi$, we compute:

$$Q(X\Pi) = (X\Pi)W_Q = Q(X)\Pi$$

$$K(X\Pi) = (X\Pi)W_K = K(X)\Pi$$

$$V(X\Pi) = (X\Pi)W_V = V(X)\Pi$$

3. The attention scores for the permuted input are:

$$S(X\Pi) = \frac{Q(X\Pi)K(X\Pi)^\top}{\sqrt{d_k}} = \frac{(Q(X)\Pi)(K(X)\Pi)^\top}{\sqrt{d_k}}$$

Using matrix properties, this simplifies to:

$$S(X\Pi) = \frac{Q(X)K(X)^\top}{\sqrt{d_k}} = S(X)$$

4. The softmax operation, applied row-wise to the scores, gives:

$$\text{Softmax}(S(X\Pi)) = \text{Softmax}(S(X))\Pi$$

5. Finally, the weighted sum with $V(X)$ results in:

$$\text{Attention}(X\Pi) = \text{Attention}(X)\Pi$$

Conclusion:

The self-attention mechanism satisfies permutation equivariance:

$$\text{Attention}(X\Pi) = \text{Attention}(X)\Pi$$

This ensures that the output of the attention mechanism respects permutations of the input.

2 Multi-Headed Attention

2.1 Drawbacks of Single-Headed Attention

(a)

To achieve $c \approx \frac{1}{2}(v_a + v_b)$, we construct the query vector q as:

$$q = \beta(\mu_a + \mu_b), \quad \text{where } \beta \gg 0$$

Reasoning:

- **Dot product with k_a:** Since $k_a \sim \mathcal{N}(\mu_a, \alpha I)$ and α is vanishingly small:

$$k_a^\top q = (\mu_a + \epsilon_a)^\top \beta(\mu_a + \mu_b) \approx \beta(\mu_a^\top \mu_a + \mu_a^\top \mu_b) = \beta(1 + 0) = \beta$$

- **Dot product with k_b:** Similarly, $k_b \sim \mathcal{N}(\mu_b, \alpha I)$:

$$k_b^\top q \approx \beta(\mu_b^\top \mu_a + \mu_b^\top \mu_b) = \beta(0 + 1) = \beta$$

- **Dot product with other k_i:** For $i \notin \{a, b\}$, μ_i is orthogonal to both μ_a and μ_b:

$$k_i^\top q \approx \beta(\mu_i^\top \mu_a + \mu_i^\top \mu_b) = \beta(0 + 0) = 0$$

Softmax and Attention Weights:
After applying the softmax function, the attention weights become:

$$\alpha_a = \alpha_b = \frac{\exp(\beta/\sqrt{d_k})}{\exp(\beta/\sqrt{d_k}) + \exp(\beta/\sqrt{d_k}) + \sum_{i \notin \{a,b\}} \exp(0)}$$

Since $\beta \gg 0$, the terms $\exp(\beta/\sqrt{d_k})$ dominate, and we get:

$$\alpha_a \approx \alpha_b \approx \frac{1}{2}, \quad \alpha_i \approx 0 \text{ for } i \notin \{a, b\}$$

Attention Output:
The final attention output is:

$$c = \sum_{i=1}^{n} \alpha_i v_i \approx \alpha_a v_a + \alpha_b v_b = \frac{1}{2} v_a + \frac{1}{2} v_b$$

Conclusion:
By setting $q = \beta(\mu_a + \mu_b)$ with $\beta \gg 0$, the query aligns equally with k_a and k_b, leading to $c \approx \frac{1}{2}(v_a + v_b)$. This construction works because q is orthogonal to all other μ_i, ensuring negligible contributions from the other keys.

(b)

When sampling the set of keys $\{k_1, k_2, \ldots, k_n\}$ multiple times, the behavior of c becomes unstable due to the large variance in the magnitude of k_a. Since the covariance matrix $\Sigma_a = \beta I + \frac{1}{2}(\mu_a \mu_a^\top)$ introduces significant variability in k_a's length, the dot product $k_a^\top q$ fluctuates significantly across samples. This causes the attention weight α_a to vary, resulting in an unstable output c.

Qualitative Behavior of c:
For different samples, c will vary depending on k_a's magnitude. When k_a is larger in a sample, α_a increases, causing v_a to dominate c. When k_a is smaller, α_a decreases, and c places less emphasis on v_a.

Comparison to Part (i):
In part (i), $k_a \sim \mathcal{N}(\mu_a, \alpha I)$, so k_a's magnitude was consistent across samples. This led to stable attention weights α_a and predictable outputs c. Here, the larger variance in k_a's magnitude introduces instability in α_a, making c vary significantly across samples.

Variance of c:
The variance of c increases due to the variability in k_a. As k_a's magnitude changes, the attention mechanism assigns fluctuating weights to v_a, leading to unpredictable changes in c's composition.

2.2 Benefits of Multi-Headed Attention

(a)

To achieve $c \approx \frac{1}{2}(v_a + v_b)$, we design two query vectors q_1 and q_2 as follows:

$$q_1 = \beta(\mu_a + \mu_b), \quad q_2 = \beta(\mu_a - \mu_b), \quad \text{where } \beta \gg 0.$$

Reasoning:

- **For q_1:** The query q_1 aligns equally with μ_a and μ_b, ensuring the attention weights for v_a and v_b are approximately equal ($\alpha_a \approx \alpha_b \approx \frac{1}{2}$).
- **For q_2:** The query q_2 aligns oppositely with μ_a and μ_b, providing diversity while still focusing on the same keys.

Multi-Headed Attention Output:
The outputs of the two heads are:

$$c_1 \approx \frac{1}{2}(v_a + v_b), \quad c_2 \approx \frac{1}{2}(v_a + v_b).$$

The final output of the multi-headed attention is the average:

$$c = \frac{1}{2}(c_1 + c_2) \approx \frac{1}{2}(v_a + v_b).$$

Conclusion:
Using q_1 and q_2, the multi-headed attention mechanism achieves the desired output $c \approx \frac{1}{2}(v_a + v_b)$ while providing diversity in its queries.

(b)

Qualitative Behavior of c:
The output c varies across different samples of the key vectors $\{k_1, k_2, \ldots, k_n\}$, primarily due to the large variance in the magnitude of k_a along its mean direction μ_a. Since k_a contributes significantly to both c_1 and c_2, its fluctuations in magnitude result in variability in these outputs.

Variance in c_1 and c_2:

- For c_1, computed using $q_1 = \beta(\mu_a + \mu_b)$, the large variance in k_a's magnitude causes α_a (the attention weight for v_a) to vary significantly. This makes c_1 unstable across different samples.
- For c_2, computed using $q_2 = \beta(\mu_a - \mu_b)$, the same variability in k_a affects c_2, but its contributions from v_a and v_b differ due to the opposite alignment of q_2 with μ_b.

Final Output c:
The final output $c = \frac{1}{2}(c_1 + c_2)$ inherits the variability from c_1 and c_2. However, since c is an average, the fluctuations in c_1 and c_2 may partially cancel out, reducing the overall variance in c. This demonstrates the robustness of multi-headed attention in handling key variability compared to single-headed attention.