Sheng Li
February 19, 2021

CS224n Assignment 5

1 Attention Exploration

(a) Since $c = \sum_{i=1}^{n} v_i \alpha_i$, if $\alpha_i \approx 0$ for $i \neq j$ and $\alpha_j \approx 1$, then c will be approximately equal to α_j.
To achieve this for attention weights α_i's, $k_j^\top q \gg k_i^\top q$ must be satisfied for $i \neq j$ (if $k_i^\top q < 0 < k_j^\top q$, the condition could be easily satisfied).

(b) Letting $c \approx \frac{1}{2}(v_a + v_b)$ requires $\alpha_a \approx \alpha_b \approx 1/2$ and $\alpha_i \approx 0$ for $i \notin \{a, b\}$. This further requires $k_a^\top q \approx k_b^\top q \gg k_i^\top q$.
If we set $q = w(k_a + k_b)$ with scalar $w \gg 0$, then $k_a^\top q = k_b^\top q = w$ and $k_i^\top q = 0$. This satisfies the aforementioned requirement.

(c) i. Since the covariance coefficient α is vanishingly small, the probability density at $\mathcal{N}(x = \mu_i, \Sigma_i)$ is extremely high. It is reasonable to assume $k_i \approx \mu_i$. Since μ_i's are orthogonal unit vectors, the requirement on q for making $c \approx \frac{1}{2}(v_a + v_b)$ is similar to that in Part 1.(b). Therefore, $q = w(\mu_a + \mu_b)$ with scalar $w \gg 0$ should satisfy the requirement.

ii. The distributions of k_i for $i \neq a$ remains the same as in Part i, the samples k_i will continue to be very close to μ_i (i.e. $k_i \approx \mu_i$). While $k_a \sim \mathcal{N}(\mu_a, \alpha I + \frac{1}{2}(\mu_a \mu_a^\top))$, since α is vanishingly small, we can assume $k_a \approx \epsilon \mu_a$ with ϵ being approximately sampled from $\mathcal{N}(1, 1/2)$. When we sample $\{k_1, \ldots, k_n\}$ multiple times, for $i \notin \{a, b\}$, $k_i^\top q \approx 0$ will always be true. And $k_b^\top q \approx w \gg 0$ will still be true since $k_b \approx \mu_b$. However, $k_a^\top q \approx \epsilon w$, and $k_a^\top q \gg 0$ is no longer always true, since $\epsilon \sim \mathcal{N}(1, 1/2)$ will mostly oscillate between 0 and 2. Then vector

$$\begin{aligned} c &\approx \frac{\exp(\epsilon w)}{\exp(\epsilon w) + \exp(w)} v_a + \frac{\exp(w)}{\exp(\epsilon w) + \exp(w)} v_b \\ &= \frac{1}{\exp((1-\epsilon)w) + 1} v_a + \frac{1}{\exp((\epsilon - 1)w) + 1} v_b. \end{aligned}$$

When $\epsilon \to 0$, $c \to v_b$; when $\epsilon \to 2$, $c \to v_a$. Vector c will approximately oscillate between v_a and v_b, for different samples of $\{k_1, \ldots, k_n\}$.

(d) i. $q_1 = q_2 = w(\mu_a + \mu_b)$ with scalar $w \gg 0$.

ii. The vector c will be an average of $c(q_1)$ and $c(q_2)$. As derived in Part 1.(c).ii, for $j \in \{1, 2\}$

$$c(q_j) \approx \frac{1}{\exp((1-\epsilon)w) + 1} v_a + \frac{1}{\exp((\epsilon - 1)w) + 1} v_b.$$

Then $c = \frac{1}{2}\left(c(q_1) + c(q_2)\right)$ will reduce the variance of its oscillation between v_a and v_b for different samples of $\{k_1, \ldots, k_n\}$ since two instances of $c(q_j)$'s are averaged to get c. If we have more attention heads, we can approach the expectation of ϵ, which is 1. Then we can have $c \approx \frac{1}{2}(v_a + v_b)$ as desired.

(e) i.

$$
\begin{aligned}
c_2 &= \alpha_{21}v_1 + \alpha_{22}v_2 + \alpha_{23}v_3 \\
&= \frac{\exp(x_1^\top x_2)x_1 + \exp(x_2^\top x_2)x_2 + \exp(x_3^\top x_2)x_3}{\exp(x_1^\top x_2) + \exp(x_2^\top x_2) + \exp(x_3^\top x_2)} \\
&= \frac{\exp((u_d+u_b)^\top u_a)(u_d+u_b) + \exp(u_a^\top u_a)u_a + \exp((u_c+u_b)^\top u_a)(u_c+u_b)}{\exp((u_d+u_b)^\top u_a) + \exp(u_a^\top u_a) + \exp((u_c+u_b)^\top u_a)} \\
&= \frac{\exp(0)(u_d+u_b) + \exp(\beta^2)u_a + \exp(0)(u_c+u_b)}{\exp(0) + \exp(\beta^2) + \exp(0)} \\
&\approx \frac{\exp(\beta^2)u_a}{\exp(\beta^2)} \\
&= u_a
\end{aligned}
$$

It is impossible for c_2 to approximate u_b by adding u_d or u_c to x_2. For example, we add u_d, c_2 becomes:

$$
\begin{aligned}
c_2 &= \frac{\exp(\beta^2)(u_d+u_b) + \exp(2\beta^2)(u_a+u_d) + \exp(0)(u_c+u_b)}{\exp(\beta^2) + \exp(2\beta^2) + \exp(0)} \\
&\approx \frac{\exp(2\beta^2)(u_a+u_d)}{\exp(2\beta^2)} \\
&= u_a + u_d
\end{aligned}
$$

Similar for adding u_c to x_2.

ii. Let $V = (u_b u_b^\top - u_c u_c^\top)/\beta^2$, then
$v_1 = Vx_1 = (u_b u_b^\top - u_c u_c^\top)(u_d+u_b)/\beta^2 = u_b u_b^\top u_b/\beta^2 = u_b$,
$v_2 = Vx_2 = (u_b u_b^\top - u_c u_c^\top)u_a/\beta^2 = 0$,
$v_3 = Vx_3 = (u_b u_b^\top - u_c u_c^\top)(u_c+u_b)/\beta^2 = u_b u_b^\top u_b/\beta^2 - u_c u_c^\top u_c/\beta^2 = u_b - u_c$.

To have $c_2 = \alpha_{21}v_1 + \alpha_{22}v_2 + \alpha_{23}v_3 = \alpha_{21}u_b + \alpha_{23}(u_b - u_c) \approx u_b$, we need $\alpha_{21} \approx 1$ and $\alpha_{23} \approx 0$. Therefore, $\exp(k_1^\top q_2) \gg \exp(k_3^\top q_2)$.
To have $c_1 = \alpha_{11}v_1 + \alpha_{12}v_2 + \alpha_{13}v_3 = \alpha_{11}u_b + \alpha_{13}(u_b - u_c) \approx u_b - u_c$, we need $\alpha_{11} \approx 0$ and $\alpha_{13} \approx 1$. That is $\exp(k_3^\top q_1) \gg \exp(k_1^\top q_1)$.
Let $K = u_a u_b^\top + u_a u_d^\top - u_c u_c^\top$ and $Q = u_a u_a^\top - u_c u_b^\top + u_c u_a^\top$, then

$$
\begin{aligned}
k_1^\top q_2 &= x_1^\top K^\top Q x_2 = (u_d+u_b)^\top (u_b u_a^\top + u_d u_a^\top - u_c u_c^\top)(u_a u_a^\top - u_c u_b^\top + u_c u_a^\top)u_a \\
&= u_b^\top u_b u_a^\top u_a u_a^\top u_a + u_d^\top u_d u_a^\top u_a u_a^\top u_a \\
&= 2\beta^6 \\
k_3^\top q_2 &= x_3^\top K^\top Q x_2 = (u_c+u_b)^\top (u_b u_a^\top + u_d u_a^\top - u_c u_c^\top)(u_a u_a^\top - u_c u_b^\top + u_c u_a^\top)u_a \\
&= u_b^\top u_b u_a^\top u_a u_a^\top u_a - u_c^\top u_c u_c^\top u_c u_a^\top u_a \\
&= 0 \\
k_3^\top q_1 &= x_3^\top K^\top Q x_1 = (u_c+u_b)^\top (u_b u_a^\top + u_d u_a^\top - u_c u_c^\top)(u_a u_a^\top - u_c u_b^\top + u_c u_a^\top)(u_d+u_b) \\
&= u_c^\top u_c u_c^\top u_c u_b^\top u_b \\
&= \beta^6 \\
k_1^\top q_1 &= x_1^\top K^\top Q x_1 = (u_d+u_b)^\top (u_b u_a^\top + u_d u_a^\top - u_c u_c^\top)(u_a u_a^\top - u_c u_b^\top + u_c u_a^\top)(u_d+u_b) \\
&= 0
\end{aligned}
$$

Then $\exp(k_1^\top q_2) \gg \exp(k_3^\top q_2)$ and $\exp(k_3^\top q_1) \gg \exp(k_1^\top q_1)$ are both satisfied.

2 Pretrained Transformer models and knowledge access

(a) None

(b) None

(c) None

(d) Model accuracy on the dev set: `Correct: 9.0 out of 500.0: 1.799%`
London baseline accuracy on the dev set: `Correct: 25.0 out of 500.0: 5.0%`

(e) None

(f) The accuracy on the dev set: `Correct: 133.0 out of 500.0: 26.6%`

(g) i. Report the accuracy of your synthesizer attention model on birth-place prediction on `birth_dev.tsv` after pretraining and fine-tuning.
The accuracy on the dev set: `Correct: 35.0 out of 500.0: 7.0%`

ii. Why might the synthesizer self-attention not be able to do, in a single layer, what the key-query-value self-attention can do?
The single layer synthesizer cannot capture the inter-dimension relative importance as the key-query-value self-attention can.

3 Considerations in pretrained knowledge

(a) First the pretrained model has a lot more training data (that might contain birth-place information) than the non-pretrained model. Secondly, the pretrained model uses the span corruption dataset from wiki to learn the pattern of the structural question answering task.

(b) Such user-facing models with indeterminacy 1. might provide misleading information, and 2. might cause bias or stereotype.

(c) The model might relate a unseen name to some close already seen name in terms of its internal latent space metric, then output the birth place of that seen person. While in reality closeness between names does not reflect the relationship between birthplaces. The model is very data hungry to be accurate.