CS 224N Spring 2024

Assignment 4

Andrew Hinh

9/8/24

1. a. i) For α_i to be ~ 1, $k_{i \neq j}^T q$ must be ≈ 0 & $k_i^T q$ must be $\gg 0$, i.e.:

ii) $q_{i \neq j} = -k_{i \neq j}$, $q_i = k_i$. ☒

$c \approx v_i \cancel{\alpha_i}^{1} \approx v_i$ ☒

b.) $q_{j \neq a,b} = -k_{j \neq a,b}$ & $q_{j = a,b} = k_{j = a,b}$ s.t.

$1 \leq j, a, b \leq n$, $j \neq a \neq b$,

normalize: $\|q\| = \sqrt{(k_a + k_b)^T (k_a + k_b)}$

$= \sqrt{\cancel{k_a^T k_a}^{1} + \cancel{k_a^T k_b}^{0} + \cancel{k_b^T k_a}^{0} + \cancel{k_b^T k_b}^{1}}$

$= \sqrt{2}$

$\Rightarrow q = \frac{k_a + k_b}{\sqrt{2}}$,

check: $k_a^T q = \frac{1}{\sqrt{2}} \cancel{k_a^T k_a}^{1} + \frac{1}{\sqrt{2}} \cancel{k_a^T k_b}^{0} = \frac{1}{\sqrt{2}}$ ✓,

" $k_b^T q$ "

✓ ☒

c. i.) $q = \frac{\mu_a + \mu_b}{\sqrt{2}}$ b/c $k_i \approx \mu_a$ given small cov. ☒

ii.) Since k_a could $\approx \frac{1}{2} \mu_a^2$, q could be much larger, which when mat mult w/ k_a would cause $\alpha_a \sim 1$ & c var to be wildly diff. ☒

d i.) Since $c \approx \frac{1}{2}(v_a + v_b) = \frac{1}{2}(c_1 + c_2)$, pick v_a, v_b = c_1, c_2 by setting $q_1 = \frac{k_a + k_b}{2}$, $q_2 = \frac{k_a - k_b}{2}$. ☒

ii) Same var in c_1, c_2 as before but now avg out, so not as much ☒

e) By avg. head outputs, there is reduced var despite high var attn scores. ☒

2. a i.) $Q_{perm} = X_{perm} \cdot W_Q = PXW_Q = PQ$; $K_{perm} = PK$;

$V_{perm} = PV$; $H_p = \text{softmax}\left(\frac{Q_p K_p^T}{\sqrt{d}}\right) V_p$

$= \text{"} \left(\frac{PQ(PK)^T}{\sqrt{d}}\right) PV$

$= \text{"} \left(P \frac{QK^T}{\sqrt{d}} P^T\right) PV$

$= P\,\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) P^T PV$

$= PH$,

$Z_p = \text{ReLU}(H_p W_1 + 1 \cdot b_1) W_2 + 1 \cdot b_2$

$= \text{ReLU}(PHW_1 + 1 \cdot b_1) W_2 + 1 \cdot b_2$

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

$= PZ$ ☒

ii.) Weights don't use permutation in calculation, so everything after gets shifted by perm & text would be jumbled. ☒

b. i.) Yes, since there is defined step increase in pos. emb., which correlates with actual text. ☒

ii.) Since denom so large (i.e. $\gg T = 1$.), values never so $> 2\pi$, so safe. ☒

3. a-c.) code

d) Dev acc: 1.2%

Baseline: 5% ☒

e.) code

f) Dev acc: 27.6% ☒

g. i) $\begin{pmatrix} \cos t\theta_1 + i\sin t\theta_1 \\ \vdots \\ \cos t\theta_{d/2} + i\sin t\theta_{d/2} \end{pmatrix} \odot \begin{pmatrix} x_t^{(1)} + i x_t^{(2)} \\ \vdots \\ x_t^{(d-1)} + i x_t^{(d)} \end{pmatrix}$

$= \begin{pmatrix} \cos t\theta_1 x_t^{(1)} + \cos t\theta_1 i x_t^{(2)} + \\ i\sin t\theta_1 x_t^{(1)} - \sin t\theta_1 x_t^{(2)} \\ \vdots \end{pmatrix}$

$= \begin{pmatrix} (\cos t\theta_1 x_t^{(1)} - \sin t\theta_1 x_t^{(2)}) + \\ i(\sin t\theta_1 x_t^{(1)} + \cos t\theta_1 x_t^{(2)}) \\ \vdots \end{pmatrix}$ ■

ii) $\langle RoPE(z_1, t_1), RoPE(z_2, t_2) \rangle$

$= \langle z_1 e^{it_1\theta}, z_2 e^{it_2\theta} \rangle$

$= Re(z_1 e^{it_1\theta} \overline{z_2 e^{it_2\theta}})$

$= Re(z_1 \overline{z_2} e^{i\theta(t_1 - t_2)})$,

$\langle RoPE(z_1, t_1 - t_2), RoPE(z_2, 0) \rangle$

$= \langle z_1 e^{i(t_1 - t_2)\theta}, z_2 \cancel{e^{i0\theta}} \rangle$

$= Re(z_1 e^{i(t_1 - t_2)\theta} \overline{z_2})$ ■

iii) Dev acc: 19.2%

■

4 a.) Via pretraining, the model better learned the facts. ■

b.) 1: lose trust in systems (increasing cost/risk to knowledgable),

2: spread misinfo.  (increasing un) ■

c.) Remembering most similar vector to current, & return that, which can be utterly wrong. ■