به نام خدا

یادگیری ژرف
نیم‌سال دوم ۰۳ – ۰۲
مدرس: دکتر مهدیه سلیمانی

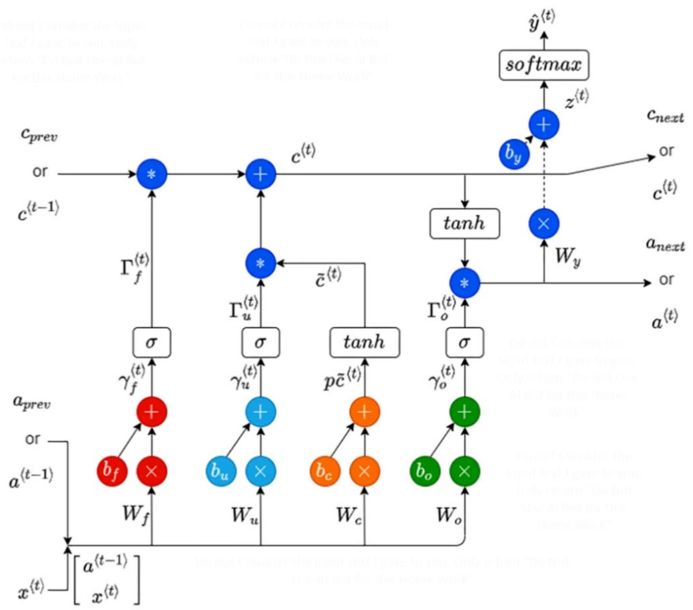
سید امیر کسائی- ۴۰۲۲۱۲۲۱۴- همفکری با: امیر محمد عزتی

سوال اول:

۱. LSTM ها به دلیل مکانیزم گیتینگ دقیق برای مدیریت وابستگی های طولانی تر و پیچیده تر بهتر هستند، اما به دلیل پارامترهای بیشتر از نظر محاسباتی فشرده تر هستند. GRUها ساده‌تر هستند و معمولاً با پارامترهای کمتری سریع‌تر آموزش می‌بینند که آنها را برای مجموعه داده‌های کوچکتر یا منابع محاسباتی محدود مناسب می‌کند. هر دو در بسیاری از سناریوها به طور قابل مقایسه ای عمل می کنند، اما LSTM ها ممکن است در وظایفی که نیاز به درک زمینه ای عمیق تر دارند، پیشی بگیرند.

۲. از آنجایی که gpu در حال حاضر هم نسبتا قوی است، ارتقای آن پیشنهاد منطقی نیست. پیشنهاد منطقی در اینجا استفاده از GRU است. این مدل بدلیل ساده تر بودن، داشتن محاسبات و پارامتر های کمتر، باعث سرعت بخشیدن می‌شود. بنابراین بجای ارتقای gpu بهتر است ساختار مدل را تغییر دهیم و از GRU استفاده کنیم.

۳.



$$\gamma_u^{<t>} = W_u \times \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_u$$

$$\Gamma_u^{<t>} = \sigma(\gamma_u^{<t>})$$

$$W_u = [W_{ua} \quad W_{ux}]$$

$$x^{<t>} = [x_0 \quad x_1 \quad \dots \quad x_i \quad \dots \quad x_n]$$

۱)

$$forget\ gate\ equations: \begin{cases} \gamma_f^{<t>} = W_f \times \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_f \\ \Gamma_f^{<t>} = \sigma(\gamma_f^{<t>}) \end{cases}$$

$$candidate\ value\ equations: \begin{cases} p\tilde{c}^{<t>} = W_c \times \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_c \\ \tilde{c}^{<t>} = tanh(p\tilde{c}^{<t>}) \end{cases}$$

$$memory\ cell\ value\ at < t >: c^{<t>} = \Gamma_u^{<t>} * \tilde{c}^{<t>} + \Gamma_f^{<t>} * c^{<t-1>}$$

$$output\ gate\ equations: \begin{cases} \gamma_o^{<t>} = W_o \times \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_o \\ \Gamma_o^{<t>} = \sigma(\gamma_o^{<t>}) \end{cases}$$

$$activation\ of\ cell\ at < t >: a^{<t>} = \Gamma_o * \tanh(c^{<t>})$$
$$output\ activation\ value: z^{<t>} = W_y \times a^{<t>} + b_y$$
$$output\ activation\ value\ after\ softmax: \bar{y}^{<t>} = softmax(z^{<t>})$$

۲) اگر هر دنباله آموزشی شامل T^x واحد باشد، از $a^{\langle 0\rangle}$ و $c^{\langle 0\rangle}$ شروع می‌کنیم و سپس $a^{\langle 1\rangle}$ و $c^{\langle 1\rangle}$ را بدست می‌آوریم که بعد به عنوان a_{prev} و c_{prev} دوباره استفاده می‌شوند. این فرآیند را T^x بار تکرار می‌کنیم.

فرض می‌کنیم که $y^{\langle t\rangle}$ خروجی خروجی قابل انتظار یک واحد LSTM در گام زمانی $\langle t\rangle$ و $\bar{y}^{\langle t\rangle}$ به عنوان خروجی واقعی واحد LSTM در آن واحد زمانی باشد.

$$Loss\ at\ \langle t\rangle: L^{\langle t\rangle} = \sum_{i=1}^{n_y} -y_i^{\langle t\rangle} log \bar{y}_i^{\langle t\rangle}$$

در اینجا n_y تعداد activation ها در $\bar{y}^{\langle t\rangle}$ است.

$$Loss\ of\ all\ LSTM\ units\ from\ \langle t\rangle\ till\ the\ last: L = \sum_{t=t}^{\langle T_x\rangle}$$

پس از انجام forward pass، backward pass را شروع می‌کنیم.

$$From\ the\ \bar{y}^{\langle t\rangle}\ side\ of\ the\ LSTM\ unit: \frac{\partial L}{\partial z^{\langle t\rangle}} = \bar{y}^{\langle t\rangle} - y^{\langle t\rangle}$$

حال تاثیر $a^{\langle t\rangle}$ را بررسی می‌کنیم. $a^{\langle t\rangle}$ روی $\bar{y}^{\langle t\rangle}$ و $\bar{y}^{\langle t+1\rangle}$ تاثیر می‌گذارد. در واقع $a^{\langle t\rangle}$ بر هر دو $L^{\langle t\rangle}$ و L' اثر دارد.

$$\frac{\partial L}{\partial a^{\langle t\rangle}} = \frac{\partial L^{\langle t\rangle}}{\partial a^{\langle t\rangle}} + \frac{\partial L'}{\partial a^{\langle t\rangle}} = \frac{\partial L^{\langle t\rangle}}{\partial z^{\langle t\rangle}} \times \frac{\partial z^{\langle t\rangle}}{\partial a^{\langle t\rangle}} + \frac{\partial L'}{\partial a^{\langle t\rangle}}$$

از آن جایی که معادله $z^{\langle t\rangle}$ را به ازای $a^{\langle t\rangle}$ می‌دانیم:

$$\frac{\partial z^{\langle t\rangle}}{\partial a^{\langle t\rangle}} = W_y$$
$$\frac{\partial L}{\partial a^{\langle t\rangle}} = W_y^T \times \left(\bar{y}^{\langle t\rangle} - y^{\langle t\rangle}\right) + da_{next}$$

حالا تاثیر $c^{\langle t\rangle}$ را پیدا می‌کنیم که بر روی واحد بعدی یعنی c_{next} و هم‌چنین $a^{\langle t\rangle}$ تاثیر دارد.

$$Impact\ on\ c^{\langle t\rangle}: \frac{\partial L}{\partial c^{\langle t\rangle}} = \frac{\partial L'}{\partial c^{\langle t\rangle}} + \frac{\partial L}{\partial a^{\langle t\rangle}} \times \frac{\partial a^{\langle t\rangle}}{\partial c^{\langle t\rangle}}$$

حال که معادله $a^{\langle t\rangle}$ را به ازای $c^{\langle t\rangle}$ و مشتق $f(x) = tanhx$ را می‌دانیم:

$$f(x) = tanhx\ is\ f'(x) = 1 - \tanh^2 x.$$
$$\frac{\partial a^{\langle t\rangle}}{\partial c^{\langle t\rangle}} = \Gamma_o^{\langle t\rangle} * (1 - \tanh^2 c^{\langle t\rangle})$$
$$\Longrightarrow \frac{\partial L}{\partial c^{\langle t\rangle}} = dc_{next} + \frac{\partial L}{\partial a^{\langle t\rangle}} * \Gamma_o^{\langle t\rangle} * (1 - \tanh^2 c^{\langle t\rangle})$$

$$\frac{\partial L}{\partial \tilde{c}^{\langle t\rangle}} = \frac{\partial L}{\partial c^{\langle t\rangle}} * \frac{\partial c^{\langle t\rangle}}{\partial \tilde{c}^{\langle t\rangle}} = [dc_{next} + \frac{\partial L}{\partial a^{\langle t\rangle}} * \Gamma_o^{\langle t\rangle} * \left(1 - \tanh^2 c^{\langle t\rangle}\right)] * \Gamma_u^{\langle t\rangle}$$

$$Tracing\ back\colon \frac{\partial L}{\partial p\tilde{c}^{\langle t\rangle}} = \frac{\partial L}{\partial \tilde{c}^{\langle t\rangle}} * \frac{\partial \tilde{c}^{\langle t\rangle}}{\partial p\tilde{c}^{\langle t\rangle}} = \frac{\partial L}{\partial \tilde{c}^{\langle t\rangle}} * \left(1 - \tanh^2 p\tilde{c}^{\langle t\rangle}\right)$$

$$= \left[dc_{next} + \frac{\partial L}{\partial a^{\langle t\rangle}} * \Gamma_o^{\langle t\rangle} * \left(1 - \tanh^2 c^{\langle t\rangle}\right)\right] * \Gamma_u^{\langle t\rangle} * \left(1 - \tanh^2 p\tilde{c}^{\langle t\rangle}\right)$$

$$= \left[dc_{next} + \frac{\partial L}{\partial a^{\langle t\rangle}} * \Gamma_o^{\langle t\rangle} * \left(1 - \tanh^2 c^{\langle t\rangle}\right)\right] * \Gamma_u^{\langle t\rangle} * (1 - \left(\tilde{c}^{\langle t\rangle}\right)^2)$$

$$influence\ of\ update\ gate\colon \frac{\partial L}{\partial \gamma_u^{\langle t\rangle}} = \frac{\partial L}{\partial c^{\langle t\rangle}} * \frac{\partial c^{\langle t\rangle}}{\partial \Gamma_u^{\langle t\rangle}} * \frac{\partial \Gamma_u^{\langle t\rangle}}{\partial \gamma_u^{\langle t\rangle}}$$

$$derivative\ for\ sigmoid\ function\colon f(x) = \sigma(x)\ is\ f'(x) = \left(f(x) * \left(1 - f(x)\right)\right)$$

$$\Rightarrow \frac{\partial L}{\partial \gamma_u^{\langle t\rangle}} = \frac{\partial L}{\partial c^{\langle t\rangle}} * \tilde{c}^{\langle t\rangle} * \left(\Gamma_u^{\langle t\rangle} * \left(1 - \Gamma_u^{\langle t\rangle}\right)\right)$$

$$= \left[dc_{next} + \frac{\partial L}{\partial a^{\langle t\rangle}} * \Gamma_o^{\langle t\rangle} * \left(1 - \tanh^2 c^{\langle t\rangle}\right)\right] * \tilde{c}^{\langle t\rangle} * \left(\Gamma_u^{\langle t\rangle} * \left(1 - \Gamma_u^{\langle t\rangle}\right)\right)$$

$$for\ forget\ gate\colon \frac{\partial L}{\partial \gamma_f^{\langle t\rangle}} = \frac{\partial L}{\partial c^{\langle t\rangle}} * \frac{\partial c^{\langle t\rangle}}{\partial \Gamma_f^{\langle t\rangle}} * \frac{\partial \Gamma_f^{\langle t\rangle}}{\partial \gamma_f^{\langle t\rangle}}$$

$$= [dc_{next} + \frac{\partial L}{\partial a^{\langle t\rangle}} * \Gamma_o^{\langle t\rangle} * \left(1 - \tanh^2 c^{\langle t\rangle}\right)] * \tilde{c}^{\langle t\rangle} * \left(\Gamma_f^{\langle t\rangle} * \left(1 - \Gamma_f^{\langle t\rangle}\right)\right)$$

$$for\ output\ gate\colon \frac{\partial L}{\partial \gamma_o^{\langle t\rangle}} = \frac{\partial L}{\partial a^{\langle t\rangle}} * \frac{\partial a^{\langle t\rangle}}{\partial \Gamma_o^{\langle t\rangle}} * \frac{\partial \Gamma_o^{\langle t\rangle}}{\partial \gamma_o^{\langle t\rangle}} = \frac{\partial L}{\partial a^{\langle t\rangle}} * \tanh c^{\langle t\rangle} * (\Gamma_o^{\langle t\rangle} * \left(1 - \Gamma_o^{\langle t\rangle}\right))$$

حالا باید تاثیر $c^{\langle t-1\rangle}$ و $a^{\langle t-1\rangle}$ را روی Loss ها بدانیم:

$$dc_{prev} = \frac{\partial L}{\partial c^{\langle t-1\rangle}} = \frac{\partial L}{\partial c^{\langle t\rangle}} * \frac{\partial c^{\langle t\rangle}}{\partial c^{\langle t-1\rangle}} = \frac{\partial L}{\partial c^{\langle t\rangle}} * \Gamma_f^{\langle t\rangle} = [dc_{next} + \frac{\partial L}{\partial a^{\langle t\rangle}} * \Gamma_o^{\langle t\rangle} * (1 - \tanh^2 c^{\langle t\rangle})] * \Gamma_f^{\langle t\rangle}$$

$$da_{prev} = \frac{\partial L}{\partial a^{\langle t-1\rangle}} = \frac{\partial L}{\partial p\tilde{c}^{\langle t\rangle}} \times \frac{\partial p\tilde{c}^{\langle t\rangle}}{\partial a^{\langle t-1\rangle}} + \frac{\partial L}{\partial \gamma_u^{\langle t\rangle}} \times \frac{\partial \gamma_u^{\langle t\rangle}}{\partial a^{\langle t-1\rangle}} + \frac{\partial L}{\partial \gamma_f^{\langle t\rangle}} \times \frac{\partial \gamma_f^{\langle t\rangle}}{\partial a^{\langle t-1\rangle}} + \frac{\partial L}{\partial \gamma_o^{\langle t\rangle}} \times \frac{\partial \gamma_o^{\langle t\rangle}}{\partial a^{\langle t-1\rangle}}$$

$$da_{prev} = W_{ca}^T \times \frac{\partial L}{\partial p\tilde{c}^{\langle t\rangle}} + W_{ua}^T \times \frac{\partial L}{\partial \gamma_u^{\langle t\rangle}} + W_{fa}^T \times \frac{\partial L}{\partial \gamma_f^{\langle t\rangle}} + W_{oa}^T \times \frac{\partial L}{\partial \gamma_o^{\langle t\rangle}}$$

$$for\ x^{\langle t\rangle} = W_{cx}^T \times \frac{\partial L}{\partial p\tilde{c}^{\langle t\rangle}} + W_{ux}^T \times \frac{\partial L}{\partial \gamma_u^{\langle t\rangle}} + W_{fx}^T \times \frac{\partial L}{\partial \gamma_f^{\langle t\rangle}} + W_{ox}^T \times \frac{\partial L}{\partial \gamma_o^{\langle t\rangle}}$$

$$Weight\ derivatives\colon \frac{\partial L}{\partial W_c} = \frac{\partial L}{\partial p\tilde{c}^{\langle t\rangle}} \times \frac{\partial p\tilde{c}^{\langle t\rangle}}{\partial W_c} = \frac{\partial L}{\partial p\tilde{c}^{\langle t\rangle}} \times \begin{bmatrix} a^{\langle t-1\rangle} \\ x^{\langle t\rangle} \end{bmatrix}^T$$

$$\frac{\partial L}{\partial W_u} = \frac{\partial L}{\partial \gamma_u^{\langle t\rangle}} \times \begin{bmatrix} a^{\langle t-1\rangle} \\ x^{\langle t\rangle} \end{bmatrix}^T$$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial \gamma_f^{\langle t\rangle}} \times \begin{bmatrix} a^{\langle t-1\rangle} \\ x^{\langle t\rangle} \end{bmatrix}^T$$

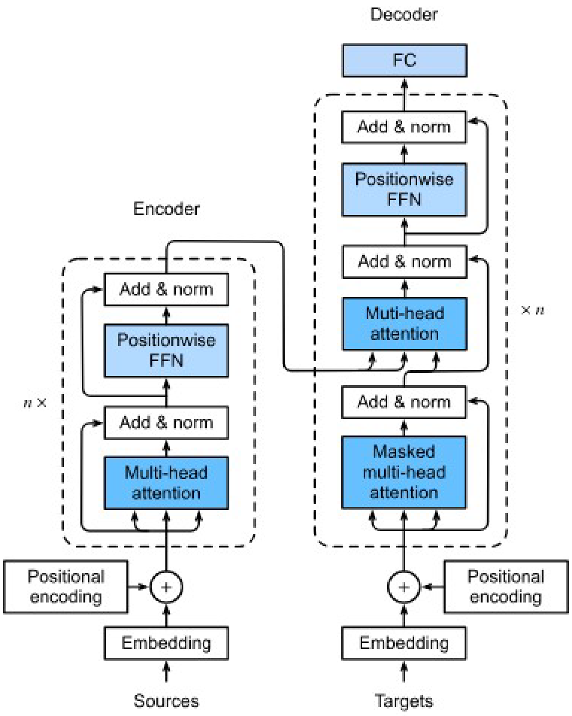
$$\frac{\partial L}{\partial W_o} = \frac{\partial L}{\partial \gamma_o^{\langle t\rangle}} \times \begin{bmatrix} a^{\langle t-1\rangle} \\ x^{\langle t\rangle} \end{bmatrix}^T$$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial z^{\langle t\rangle}} \times \left(a^{\langle t\rangle}\right)^T$$

$$bias\ derivatives\colon \frac{\partial L}{\partial W_c} = \sum \frac{\partial L}{\partial p\tilde{c}^{\langle t\rangle}}$$

$$\frac{\partial L}{\partial b_u} = \sum \frac{\partial L}{\partial \gamma_u^{\langle t \rangle}}$$
$$\frac{\partial L}{\partial b_f} = \sum \frac{\partial L}{\partial \gamma_f^{\langle t \rangle}}$$
$$\frac{\partial L}{\partial b_o} = \sum \frac{\partial L}{\partial \gamma_o^{\langle t \rangle}}$$
$$\frac{\partial L}{\partial b_y} = \sum \frac{\partial L}{\partial z^{\langle t \rangle}}$$

سوال دوم:



۱. ابعاد سیگنال ها

$Source: 32 \times 2048 \times 30000$
$Embedding: 32 \times 2048 \times 1024$

Encoder:
$After\ Positionl\ Encoding: 32 \times 2048 \times 1024$
$$Multi\ head\ attention: \begin{cases} each\ head: 32 \times 2048 \times 192 \\ concat\ all\ heads: 32 \times 2048 \times 768 \\ linear: 32 \times 2048 \times 1024 \end{cases}$$
$Add\ \&\ norm: 32 \times 2048 \times 1024$
$$Positionnwise\ FFN: \begin{cases} first: 32 \times 2048 \times 512 \\ seocnd: 32 \times 2048 \times 1024 \end{cases}$$
$Add\ \&\ norm: 32 \times 2048 \times 1024$

$Targets: 32 \times 2048 \times 30000$
$Embedding: 32 \times 2048 \times 1024$

Decoder:
$After\ Positionl\ Encoding: 32 \times 2048 \times 1024$

$$Masked\ Multi\ head\ attention: \begin{cases} each\ head: 32 \times 2048 \times 192 \\ concat\ all\ heads: 32 \times 2048 \times 768 \\ linear: 32 \times 2048 \times 1024 \end{cases}$$
$Add\ \&\ norm: 32 \times 2048 \times 1024$
$$Multi\ head\ attention: \begin{cases} each\ head: 32 \times 2048 \times 192 \\ concat\ all\ heads: 32 \times 2048 \times 768 \\ linear: 32 \times 2048 \times 1024 \end{cases}$$
$Add\ \&\ norm: 32 \times 2048 \times 1024$
$$Positionnwise\ FFN: \begin{cases} first: 32 \times 2048 \times 512 \\ seocnd: 32 \times 2048 \times 1024 \end{cases}$$
$Add\ \&\ norm: 32 \times 2048 \times 1024$

۲. تعداد پارامتر ها

$Embedding(w, b): 30000 \times 1024 + 1024$

Encoder:

$$Multi\ head\ attention: \begin{cases} each\ head\ (W_Q, W_K, W_V): 3 \times (1024 \times 192 + 192) \rightarrow all\ heads: 4 \times 3 \times (1024 \times 192 + 192) \\ linear\ (W_O): 768 \times 1024 + 1024 \end{cases}$$
$Add\ \&\ norm: 2 \times 1024$
$$Positionnwise\ FFN: \begin{cases} first: 1024 \times 512 + 512 \\ seocnd: 512 \times 1024 + 1024 \end{cases}$$
$Add\ \&\ norm: 2 \times 1024$

$total\ for\ each\ encoder\ block: 4{,}203{,}264 \rightarrow 12\ blocks = 12 \times 4{,}203{,}264$

$Embedding(w, b): 30000 \times 1024 + 1024$

Decoder:

$$Masked\ Multi\ head\ attention: \begin{cases} each\ head\ (W_Q, W_K, W_V): 3 \times (1024 \times 192 + 192) \rightarrow all\ heads: 4 \times 3 \times (1024 \times 192 + 192) \\ linear\ (W_O): 768 \times 1024 + 1024 \end{cases}$$
$Add\ \&\ norm: 2 \times 1024$
$$Multi\ head\ attention: \begin{cases} each\ head\ (W_Q, W_K, W_V): 3 \times (1024 \times 192 + 192) \rightarrow all\ heads: 4 \times 3 \times (1024 \times 192 + 192) \\ linear\ (W_O): 768 \times 1024 + 1024 \end{cases}$$
$Add\ \&\ norm: 2 \times 1024$
$$Positionnwise\ FFN: \begin{cases} first: 1024 \times 512 + 512 \\ seocnd: 512 \times 1024 + 1024 \end{cases}$$
$Add\ \&\ norm: 2 \times 1024$

$total\ for\ each\ decoder\ block: 7{,}354{,}368 \rightarrow 8\ blocks = 8 \times 7{,}354{,}3684$

$FC: 1024 \times 30000 + 30000 = 30{,}750{,}000$

$$Total\ (2 * Embedding + Encoders + Deoders + FC) \\ = 2 * 30{,}721{,}024 + 12 \times 4{,}203{,}264 + 8 \times 7{,}354{,}3684 + 30{,}750{,}000 = 201{,}466{,}160$$

سوال سوم:

۱.

$$y_n = \sum_{m=1}^{N} a_{nm}\, x_m \qquad a_{nm} = \frac{\exp(x_n^T x_m)}{\sum_{m'=1}^{N} \exp(x_n^T x_{m'})}$$

$$\forall n \neq m: x_n^T x_m = 0 \rightarrow a_{nm} = \begin{cases} \dfrac{\exp(x_n^T x_n)}{(N-1)+\exp(x_n^T x_n)}, & n = m \\ \dfrac{1}{(N-1)+\exp(x_n^T x_n)}, & O.W \end{cases}$$

$$if\ \ \exp(x_n^T x_m) \gg N \rightarrow a_{nm} \approx \begin{cases} 1, & n = m \\ 0, & O.W \end{cases} \rightarrow y_n = \sum_{m=1}^{N} a_{nm}\, x_m = a_{nn} x_n = x_n$$

۲.

$$a^T b = \sum_{i}^{D} a_i b_i \qquad a, b \sim \mathcal{N}(0, I) \rightarrow a_i, b_i \sim \mathcal{N}(0, I) \rightarrow a_i b_i \sim \mathcal{N}(0, I)$$

$$Var(a^T b) = \sum_{i}^{D} Var(a_i b_i) \xrightarrow{a_i b_i \sim \mathcal{N}(0,I)} = Var(a^T b) = D$$

$$we\ know\ that\ Var(X) = E[X^2] - E[X]^2 \rightarrow Var(a^T b) = E[(a^T b)^2] - E[a^T b]^2$$

$$a, b \sim \mathcal{N}(0, I) \rightarrow\ E[a^T b] = 0 \rightarrow Var(a^T b) = E[(a^T b)^2]$$

$$\rightarrow\ E[(a^T b)^2] = Var(a^T b) = D$$

۳.

$$Y(X) = Concat[H_1, \cdots, H_H] W^{(o)}$$

$$H_h = Softmax[\frac{Q_h K_h^T}{\sqrt{D_{k_h}}}] V_h \qquad\qquad Y(X) = \sum_{h=1}^{H} Softmax[\frac{Q_h K_h^T}{\sqrt{D_{k_h}}}] X W^{(h)}$$

$$Q_h = X W_h^{(q)},\ K_h = X W_h^{(k)},\ V_h = X W_h^{(v)}$$

$$if\ we\ devide\ W^{(o)}\ horizentally \rightarrow W^{(o)} = \begin{bmatrix} W_1^{(o)} \\ \vdots \\ W_h^{(o)} \end{bmatrix} \rightarrow Y(X) = Concat[H_1, \dots, H_h] W^{(o)} = \sum_{h=1}^{H} H_h . W_h^{(o)}$$

$$\begin{cases} Y(X) = \sum_{h=1}^{H} Softmax\left[\dfrac{Q_h K_h^T}{\sqrt{D_{k_h}}}\right] X W^{(h)} \\ W^{(h)} = W_h^{(v)} W_h^{(o)} \end{cases} \rightarrow Y(X) = \sum_{h=1}^{H} Softmax\left[\frac{Q_h K_h^T}{\sqrt{D_{k_h}}}\right] X W_h^{(v)} W_h^{(o)}$$

$$= \sum_{h=1}^{H} Softmax[\frac{Q_h K_h^T}{\sqrt{D_{k_h}}}] V_h W_h^{(o)} = \sum_{h=1}^{H} H_h . W_h^{(o)} = Concat[H_1, \dots, H_h] W^{(o)}$$

سوال چهارم:

۱.

- Fixed positional encoding:
  - شامل استفاده از روش ایستا و از پیش تعریف شده برای encode موقعیت هایی است که در طول تمرین تغییر نمی کند. در این روش معمولاً با استفاده از توابع سینوسی انجام می‌شود، که در آن هر موقعیت در دنباله با استفاده از ترکیبی از توابع سینوسی و کسینوس در فرکانس‌های مختلف encode می‌شود.
  - مزایا:
    - Generality and Theoretical Grounding: توابع سینوسی از نظر تئوری می توانند دنباله هایی با هر طولی را پشتیبانی کنند و استفاده از توابع تناوبی به مدل کمک می کند تا موقعیت های توالی خارج از داده های آموزشی را extrapolate کند.

- No Training Required: از آنجایی که encoding، ثابت است و نیازی به یادگیری ندارد، خطر overfit با ویژگی های داده های آموزشی در مورد موقعیت وجود ندارد.
  - معایب:
    - Less Flexible: Fixed encoding ممکن است برای همه نوع داده‌ها و task ها بهینه نباشد زیرا بر اساس داده‌های خاص یا تفاوت‌های ظریف در task ها منطبق نمیشوند.

- Learnable positional encoding:
  - positional encoding را به عنوان پارامترهایی در نظر می گیرد که می توانند در طول فرآیند آموزش، مشابه وزن های دیگر در مدل، یاد بگیرند. ایده این است که مدل می تواند یک مفهوم منحصر به فرد از موقعیت را ایجاد کند که به بهترین وجه با ویژگی های خاص مجموعه داده و task در دست مناسب است.
  - مزایا:
    - Adaptability: می توانند با نیازهای خاص داده ها یا task ها سازگار شوند و به طور بالقوه روابط موقعیتی مفید یا ظریف تری را نسبت به روش های ثابت ثبت کنند.
    - Potential for better performance: برای task های خاص، به ویژه آنهایی که موقعیت نقش پیچیده ای دارد، ممکن است منجر به عملکرد کلی بهتر مدل شود.
  - معایب:
    - Risk of overfitting: از آنجایی که این پارامترها آموخته شده‌اند، خطر بیشتری وجود دارد که ممکن است به داده‌های آموزشی overfit شوند، به خصوص اگر مجموعه داده نماینده منطقه کاربردی گسترده‌تر نباشد.
    - Computationally expensive: به پارامترهای اضافی برای آموزش نیاز دارد که می تواند سربار محاسباتی را افزایش دهد.

- absolute positional encoding:
  - شامل افزودن یا الحاق یک بردار به token embedding ورودی است که نشان دهنده موقعیت token در دنباله است. این موقعیت مستقل از موقعیت های دیگر در دنباله encode می شود. موقعیت‌های مطلق را نیز می‌توان در طی آموزش مدل یاد گرفت.
  - مزایا:
    - Simplicity: رمزگذاری برای هر موقعیت به طور مستقل ایجاد می شود و پیاده سازی را ساده می کند.
    - Effectiveness for small sequenes: برای دنباله‌هایی که خیلی طولانی یا پیچیده نیستند، راه ساده‌ای برای اطلاع‌رسانی به مدل در مورد ترتیب توکن فراهم می‌کند..
  - معایب:
    - Limited context awareness: به طور ذاتی رابطه بین موقعیت های مختلف را نشان نمی دهند. آنها فقط برای مدل context مربوط به موقعیت های جدا از هم را به صورت تفکیک شده ارائه می دهند.
    - Scalability to Long Sequences: به خصوص با رمزگذاری سینوسی، مدیریت دنباله های بسیار طولانی می تواند مشکل ساز شود، زیرا ممکن است مدل به دلیل ماهیت تناوبی سینوسی ها، بین موقعیت های دور تمایز قائل نشود.
- Relative positional encoding:
  - فاصله یا رابطه بین token ها را در یک دنباله در نظر می گیرد. به جای encode کردن هر موقعیت به صورت مجزا، تفاوت در موقعیت‌های بین نشانه‌ها را encode می‌کند و بر نحوه ارتباط توکن‌ها با یکدیگر در چارچوب دنباله‌شان تمرکز می‌کند.

- مزایا:
  - Contextual relevance: با رمزگذاری روابط بین موقعیت ها، رمزگذاری های موقعیتی نسبی زمینه غنی تری را فراهم می کنند که می تواند به ویژه در کارهایی که نیاز به درک پویایی دنباله دارند، مانند مدل سازی زبان و تجزیه، مفید باشد.
  - Better scalability for long sequences: رمزگذاری موقعیتی نسبی می‌تواند توالی‌های طولانی‌تر را به‌طور مؤثرتری مدیریت کند، زیرا مدل یاد می‌گیرد که بر تفاوت‌های موقعیتی که بیشترین ارتباط را برای کار دارند، بدون توجه به موقعیت مطلق در دنباله ورودی تمرکز کند.
- معایب:
  - Complexity: اجرای رمزگذاری موقعیتی نسبی به طور کلی پیچیده تر از رمزگذاری مطلق است. این نیاز به تغییراتی در مکانیزم self attention دارد تا تفاوت‌های موقعیت را در نظر بگیرد، که می‌تواند هم معماری مدل و هم پویایی آموزشی آن را پیچیده کند.
  - Computational cost: رمزگذاری موقعیتی نسبی می تواند به دلیل محاسبات اضافی مورد نیاز در طول آموزش و inference، سربار محاسباتی را افزایش دهد.

۲.

- Compatibility with Linear Self-Attention: یکی از مزیت های کلیدی RoPE سازگاری آن با مکانیسم های خطی Self-Attention است. رویکردهای سنتی برای relative positional encoding اغلب با چالش هایی در ادغام اطلاعات موقعیت نسبی به طور موثر با Self-Attention مواجه هستند. RoPE با ارائه روشی که به طور یکپارچه وابستگی های موقعیت نسبی را در فرمول توجه به خود ترکیب می کند، به این موضوع می پردازد، و توانایی مدل را برای درک روابط موقعیتی در دنباله ورودی افزایش می دهد.

- Decaying Inter-Token Dependency: RoPE یک ویژگی منحصر به فرد را معرفی می کند که در آن اطلاعات موقعیت نسبی با افزایش فاصله بین توکن ها کاهش می یابد. این کاهش در وابستگی بین token ها برای وظایف پردازش زبان طبیعی بسیار مهم است، زیرا به مدل کمک می‌کند تا روی وابستگی‌های محلی درون دنباله تمرکز بیشتری داشته باشد و در عین حال تأثیر token های دور را کاهش دهد. با ترکیب این ویژگی کاهشی، RoPE توانایی مدل را برای گرفتن اطلاعات متنی مرتبط در فواصل مختلف در توالی ورودی بهبود می‌بخشد و منجر به یادگیری و نمایش مؤثرتر می‌شود.

- Flexibility in Sequence Length: یکی دیگر از مزایای قابل توجه RoPE انعطاف پذیری آن در مدیریت توالی با طول های مختلف است. روش‌های رمزگذاری موقعیتی قبلی ممکن است با طول‌های دنباله‌ای متفاوت مشکل داشته باشند که بر عملکرد و سازگاری مدل با اندازه‌های ورودی مختلف تأثیر می‌گذارد. طراحی RoPE به آن اجازه می دهد تا به طور یکپارچه با توالی هایی با طول های مختلف سازگار شود و راه حلی قوی تر و همه کاره برای رمزگذاری اطلاعات موقعیت در مدل های transformer ارائه دهد. این انعطاف‌پذیری ظرفیت مدل را برای پردازش ورودی‌های با طول‌های مختلف بدون به خطر انداختن عملکرد یا کارایی افزایش می‌دهد.

۳. با ضرب ماتریس چرخش در بردار embedding، اندازه را ثابت را نگه میدارد و فقط زاویه بردار را تغییر میدهد. این به این معناست که پارامتر θ که معادل relative distance، میان المان ها است تغییر میکند و در نتیجه با این عمل صرفا relative position مربوط به عناصر بردار تغییر خواهد کرد.

۴.

۵. اصطلاح q_iK^T حاصل ضرب نقطه‌ای بردار query (q_i) با بردارهای key را نشان می‌دهد، که ارتباط بین query token و همه key token ها در دنباله ورودی را اندازه‌گیری می‌کند. عبارت m · [−(i − 1), ..., −2, −1, 0] یک bias مرتبط با موقعیت را بر

اساس فواصل بین query token و key token ها تعریف می کند. مقادیر منفی در دنباله نشان دهنده موقعیت نسبی key token ها با توجه به query token ها است. با افزودن عبارت bias خطی، مکانیسم attention تحت تأثیر موقعیت های نسبی token ها در دنباله ورودی قرار می گیرد. این bias اطلاعات مربوط به موقعیت را ارائه می دهد که مدل را برای توجه به موقعیت های مختلف بر اساس فاصله آنها از token پرس و جو راهنمایی می کند. پارامتر m نیز به عنوان head-specific slope، نقش scale کردن bias را دارد که قبل از زمان آموزش به صورت ثابت تنظیم میشود.

سوال پنجم:

۱. $Y = Attention(Q,K,V) = Softmax\left(Q\frac{K^T}{\sqrt{D}}\right)V$

در ساختار self-attention ماتریس های Q، K و V، به ترتیب مربوط به query، key و value هستند که در آن ها هر سطر مربوط به بردار یک کلمه در دنباله ورودی است. D نیز ابعاد بردار های key است. این معادله میتواند به شکل ماتریسی نمایش داده شود که در آن هر دنباله ورودی از بردار های کلمات میتواند به طور پیوسته به بردار خروجی با ابعاد یکسان تبدیل شود.

۲. ماتریس های Q، K و V دارای ابعاد NxD هستند (N طول دنباله ورودی و D ابعاد بردار کلمات). بنابراین ابعاد K^T، DxN خواهد بود. در نتیجه ابعاد QK^T، NxN خواهد شد. تابع sofmax چون به صورت element wise انجام میشود تاثیری در اندازه خروجی ندارد. حاصل عبارت بعد از ضرب با ماتریس V دارای ابعاد NxD خواهد شد. از آنجا که تعداد پارامتر های attention از مرتبه n^2 است، در نتیجه تعداد کل پارامتر ها $N^2D^2 = O(N^2D^2)$ خواهد شد.

۳. در ماتریس self-attention، هر عنصر حاصل ضرب query یک کلمه در key کلمه دیگر است. اگر این ماتریس به صورت بلوکی در نظر گرفته شود که در آن هر بلوک معادل تعامل یک query و یک key باشد، میتوان دید که بسیاری از بلوک ها پارامتر هایشان را به اشتراک میگذارند. مخصوصا زمانی که مقادیر وزن attention دو کلمه برای بقیه کلمات یکسان باشد، بلوک مربوط به تعامل آنها یکسان خواهد شد. همچنین به دلیل وجود softmax، تعداد زیادی از عناصر این ماتریس صفر خواهند شد.

۴. Positional encoding اطلاعات مربوط به ترتیب کلمات در دنباله ورودی را مشخص میکند. بنابراین اگر آن را حذف کنیم، ساختار self-attention فقط به معنای کلمات توجه میکند و مکان کلمات در جمله را نادیده میگیرد. در نتیجه خروجی ویژگی های مربوط به مکان کلمات را شامل نمیشود که منجر به همسانی با ترتیب تکراری ورودی می شود. بنابراین با تغییر ترتیب کلمات ورودی، ترتیب کلمات متناظر در خروجی نیز تغییر میکند که مطلوب نیست زیرا ترتیب کلمات در جمله اهمیت دارند.