Question 1. Answers

1.1 Given

$$\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$$

If we let

$$\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$$

then

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$$

=

$$\sigma(\boldsymbol{W}\frac{(\boldsymbol{h}_t - \boldsymbol{U}\boldsymbol{x}_t - \boldsymbol{b})}{\boldsymbol{W}} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$$

=

$$\sigma((\boldsymbol{h}_t - \boldsymbol{U}\boldsymbol{x}_t - \boldsymbol{b}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$$

=

$$\sigma(\boldsymbol{h}_t)$$

*1.2 Since in Q1.1 we proved the

$$\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$$

results in an equivalent recurrence as the conventional way of applying the activation function. Hence, we use it to calculate the the gradient of the hidden state as :

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \frac{\boldsymbol{h}_t}{\sigma(\boldsymbol{h}_{t-1})} * \frac{\sigma(\boldsymbol{h}_{t-1})}{\boldsymbol{h}_{t-1}} = \boldsymbol{W}^T * diag(\sigma'(\boldsymbol{h}_{t-1}))$$

Therefore

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_k} = \prod_{k<i\leq t} \frac{\boldsymbol{h}_i}{\sigma(\boldsymbol{h}_{i-1})} * \frac{\sigma(\boldsymbol{h}_{i-1})}{\boldsymbol{h}_{i-1}} = \prod_{k<i\leq t} \boldsymbol{W}^T * diag(\sigma'(\boldsymbol{h}_{i-1}))$$

Hence

$$\left|\left|\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}\right|\right| = ||\boldsymbol{W}^T diag(\sigma'(\boldsymbol{h}_i))|| \leq ||\boldsymbol{W}^T||||diag(\sigma'(\boldsymbol{h}_i))||$$

since $|\sigma'(x)| \leq \gamma$ and $\lambda_1(\boldsymbol{W}^\top\boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2}$ we can get

$$||\boldsymbol{W}^T||||diag(\sigma'(\boldsymbol{h}_i))|| \leq \frac{\delta}{\gamma}\gamma \leq \delta < 1$$

Therefore,

$$\left|\left|\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}\right|\right| \leq \delta$$

Then

$$\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| = \left|\left|\prod_{i=0}^{T-1}\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}\right|\right| \leq \delta^T$$

since $0 < \delta < 1$, the term δ^T goes to zero exponentially fast with T.Thurs, we can conclude that$\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right|$ goes to zero as $T \to \infty$.

1.3 It will be the necessary condition for gradient to explode depends on the value of γ and δ. If $\frac{\delta^2}{\gamma^2}$ is larger than 1 then the gradient will explode.

Question 2. Answer

2.1 For $t \geq 1$,

- SGD with momentum :

$$\boldsymbol{v}_t = \alpha \boldsymbol{v}_{t-1} + \epsilon \boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$$

where $\epsilon > 0$ and $\alpha \in (0, 1)$.
we get :

$$\Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t = -\alpha \boldsymbol{v}_{t-1} - \epsilon \boldsymbol{g}_t$$

- SGD with running average of $\boldsymbol{g}_t$:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\delta \boldsymbol{v}_t$$

where $\beta \in (0, 1)$ and $\delta > 0$.
we get :

$$\Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t = -\delta(\beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t) = -\delta(\beta\Delta\boldsymbol{\theta}_{t-1} + (1-\beta)\boldsymbol{g}_t) = -\delta\beta\Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t$$

from above, we can see if we let

$$(\alpha, \epsilon) = (\delta\beta, \delta(1-\beta))$$

then these two update rules are equivalent.

2.2

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t$$

=

$$\beta(\beta\boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t$$

=

$$\beta^2\boldsymbol{v}_{t-2} + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t$$

=

$$\beta^2(\beta\boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2}) + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t$$

=

$$\beta^3\boldsymbol{v}_{t-3} + \beta^2(1-\beta)\boldsymbol{g}_{t-2} + \beta^1(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t$$

We can see that we can always represent $\boldsymbol{v}_t$ as

$$\beta^i\boldsymbol{v}_{t-i} + \beta^{i-1}(1-\beta)\boldsymbol{g}_{t-i+1} + \beta^{i-2}(1-\beta)\boldsymbol{g}_{t-i+2} + \ldots + \beta^0(1-\beta)\boldsymbol{g}_t$$

If we unroll till $i = t$, we get :

$$\beta^t\boldsymbol{v}_0 + \beta^{t-1}(1-\beta)\boldsymbol{g}_1 + \beta^{t-2}(1-\beta)\boldsymbol{g}_2 + \ldots + \beta^0(1-\beta)\boldsymbol{g}_t$$

Since $\boldsymbol{v}_0$ is a vector of zero, the above equation is equal to

$$\boldsymbol{v}_t = \beta^{t-1}(1-\beta)\boldsymbol{g}_1 + \beta^{t-2}(1-\beta)\boldsymbol{g}_2 + \beta^{t-i}(1-\beta)\boldsymbol{g}_i \ldots + \beta^0(1-\beta)\boldsymbol{g}_t$$

2.3 From the equation above, we can see that the $\boldsymbol{v}_t$ is depends on the value of the previous $\boldsymbol{g}_i$ with the β^{t-i} as its weights. Since the $\beta \in (0,1)$, we can see as t increases, the older value of the $\boldsymbol{g}_i$ get smaller and contributes less for the overall value of the current $\boldsymbol{v}_t$. When the t increase to certain point, we can see the β^t will be close to zero and the older value will be "forgotten". We can see it mathematically from the following formula,

$$\mathbb{E}[\boldsymbol{v}_t] = \mathbb{E}[\sum_{i=1}^{t}(\beta^{t-i}(1-\beta)\boldsymbol{g}_i)] = (1-\beta)\mathbb{E}[\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i)] = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}[\boldsymbol{g}_i] \neq \mathbb{E}[\boldsymbol{g}_i]$$

Since $\boldsymbol{g}_t$ has a stationary distribution independent of t, it means $\mathbb{E}[\boldsymbol{g}_i] = E[\boldsymbol{g}], \forall i$.

$$\mathbb{E}[\boldsymbol{v}_t] = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}[\boldsymbol{g}_i] = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}[\boldsymbol{g}]$$

Since $\sum_{i=1}^{t}\beta^{t-i} = \frac{1-\beta^t}{1-\beta}$ according to the geometric series rule, we have

$$\mathbb{E}[\boldsymbol{v}_t] = \mathbb{E}[\boldsymbol{g}](1-\beta)\frac{1-\beta^t}{1-\beta} = \mathbb{E}[\boldsymbol{g}](1-\beta^t)$$

Thus

$$\mathbb{E}[\frac{\boldsymbol{v}_t}{1-\beta^t}] = E[\boldsymbol{g}]$$

Therefore,a bias-corrected version of $\boldsymbol{v}_t$ is to replace

$$\boldsymbol{v}_t = \frac{\boldsymbol{v}_t}{1-\beta^t}$$

Hence, as the value of t increase, the value of β^t will decrease close to zero then not change the current $\boldsymbol{v}_t$.

Question 3. Answer

3.1 (a) Given that queries, keys, and values are

$$\bar{\boldsymbol{Q}}, \bar{\boldsymbol{K}}, \bar{\boldsymbol{V}} \in \mathbb{R}^{n\times d}$$

$$\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K \in \mathbb{R}^{d\times d}, \boldsymbol{W}_i^V \in \mathbb{R}^{d\times d} \forall i$$

we can have the dimension of

$$\boldsymbol{Q} \in \mathbb{R}^{n\times d}$$

$$\boldsymbol{K} \in \mathbb{R}^{n\times d}$$

$$\boldsymbol{V} \in \mathbb{R}^{n\times d}$$

In this step, we have the time complexity $\Theta(ndd){=}\Theta(nd^2)$; The space complexity is $\Theta(nd + d^2 + nd) = \Theta(d^2 + 2nd)$ for each $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$.
The dot production between $\boldsymbol{Q}\boldsymbol{K}^\top$ ends up a matrix with

$$\boldsymbol{Q}\boldsymbol{K}^\top \in \mathbb{R}^{n\times n}$$

In this step, we have the time complexity $\Theta(nnd){=}\Theta(n^2d)$; The space complexity is $\Theta(nd + nd + n^2) = \Theta(n^2)$ by assuming we remove the intermediate variables $\boldsymbol{Q}, \boldsymbol{K}$.
In the softmax operation step,

$$\text{softmax}_{\text{row}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)$$

The time complexity is $\Theta(nn){=}\Theta(n^2)$; The space complexity is $\Theta(n^2) = \Theta(n^2)$ by assuming we remove the intermediate variables $\boldsymbol{Q}\boldsymbol{K}^\top$.

The last step is to times the $\boldsymbol{V} \in \mathbb{R}^{n\times d}$ and resulting a matrix with

$$\text{softmax}_{\text{row}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V} \in \mathbb{R}^{n\times d}$$

The time complexity is $\Theta(nnd){=}\Theta(n^2d)$; The space complexity is $\Theta(nd) = \Theta(nd)$ by assuming we remove the intermediate variables $\text{softmax}_{\text{row}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)$ and $\boldsymbol{V}$.

The total time complexity is $\Theta(3nd^2 + n^2d + n^2 + n^2d) = \Theta(3nd^2 + 2n^2d + n^2)$. The space time complexity is $\Theta(3d^2 + 3 * 2nd + n^2 + n^2 + nd) = \Theta(3d^2 + 7nd + 2n^2)$

(b) In terms of multi-head dot-product attention, we have $\text{Concat}(\text{head}_1, \ldots, \text{head}_h) \in \mathbb{R}^{n\times dh}$. The last linear layer results in a matrix of $\in \mathbb{R}^{n\times d}$ from $\text{Concat}(\text{head}_1, \ldots, \text{head}_h) \in \mathbb{R}^{n\times dh}$ and $\boldsymbol{W}_O \in \mathbb{R}^{hd\times d}$. The additional time complexity is $\Theta(ndhd){=}\Theta(nhd^2)$; The total time complexity is $\Theta((3d^2 + 7nd + 2n^2 + nd^2) * h)$

Since it is assumed that the heads are computed sequentially,therefore, the space complexity is the same as the single head attention $\Theta(3d^2 + 7nd + 2n^2)$.

(c) For very long sequences,the n will be large. Bottleneck is the computation of the softmax attention. The

$$\text{softmax}_{\text{row}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V} \in \mathbb{R}^{n\times d}$$

will need more time $\Theta((2d+1)*n^2)$ to compute and more space $\Theta(2n^2)$ since the time complexity and space complexity are increasing as quadratic in the length of input n.

3.2 (a)

$$\text{softmax}_{row}(\boldsymbol{Q})$$

$$= \begin{bmatrix} \text{softmax}(\boldsymbol{q}_{0,0}, \boldsymbol{q}_{0,1}, \boldsymbol{q}_{0,2}, \dots, \boldsymbol{q}_{0,d}) \\ \text{softmax}(\boldsymbol{q}_{1,0}, \boldsymbol{q}_{1,1}, \boldsymbol{q}_{1,2}, \dots, \boldsymbol{q}_{1,d}) \\ \dots \\ \text{softmax}(\boldsymbol{q}_{n,0}, \boldsymbol{q}_{n,1}, \boldsymbol{q}_{n,2}, \dots, \boldsymbol{q}_{n,d}) \end{bmatrix}$$

$$= \begin{bmatrix} s^{q_{0,0}} & s^{q_{0,1}} & s^{q_{0,2}} & \dots & s^{q_{0,d}} \\ s^{q_{1,0}} & s^{q_{1,1}} & s^{q_{1,2}} & \dots & s^{q_{1,d}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s^{q_{n,0}} & s^{q_{n,1}} & s^{q_{n,2}} & \dots & s^{q_{n,d}} \end{bmatrix}$$

For each row $i \in (0, ..., n)$, as a result of the softmax operation, we have :

$$\sum_{j=0}^{d} s^{q_{i,j}} = 1$$

for

$$\text{softmax}_{col}(\boldsymbol{K})$$

We have

$$\begin{bmatrix} \text{softmax}(\boldsymbol{k}_{0,0}, \boldsymbol{k}_{1,0}, \boldsymbol{k}_{2,0}, \dots, \boldsymbol{k}_{n,0}) \\ \text{softmax}(\boldsymbol{k}_{0,1}, \boldsymbol{k}_{1,1}, \boldsymbol{k}_{2,1}, \dots, \boldsymbol{k}_{n,1}) \\ \dots \\ \text{softmax}(\boldsymbol{k}_{0,d}, \boldsymbol{k}_{1,d}, \boldsymbol{k}_{2,d}, \dots, \boldsymbol{k}_{n,d}) \end{bmatrix}$$

$$= \begin{bmatrix} s^{k_{0,0}} & s^{k_{0,1}} & s^{k_{0,2}} & \dots & s^{k_{0,d}} \\ s^{k_{1,0}} & s^{k_{1,1}} & s^{k_{1,2}} & \dots & s^{k_{1,d}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s^{k_{n,0}} & s^{k_{n,1}} & s^{k_{n,2}} & \dots & s^{k_{n,d}} \end{bmatrix}$$

With transpose, we have

$$\begin{bmatrix} s^{k_{0,0}} & s^{k_{1,0}} & s^{k_{2,0}} & \dots & s^{k_{n,0}} \\ s^{k_{0,1}} & s^{k_{1,1}} & s^{k_{2,1}} & \dots & s^{k_{n,1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s^{k_{0,d}} & s^{k_{1,d}} & s^{k_{2,d}} & \dots & s^{k_{n,d}} \end{bmatrix}$$

for each row $i \in n$, as a result of the softmax operation, we have :

$$\sum_{i=0}^{n} s^{k_{i,j}} = 1$$

$$\text{softmax}_{row}(\boldsymbol{Q})\text{softmax}_{col}(\boldsymbol{K})^T$$

$$=$$

$$\begin{bmatrix} s^{q_{0,0}} & s^{q_{0,1}} & s^{q_{0,2}} & \dots & s^{q_{0,d}} \\ s^{q_{1,0}} & s^{q_{1,1}} & s^{q_{1,2}} & \dots & s^{q_{1,d}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s^{q_{n,0}} & s^{q_{n,1}} & s^{q_{n,2}} & \dots & s^{q_{n,d}} \end{bmatrix} \begin{bmatrix} s^{k_{0,0}} & s^{k_{1,0}} & s^{k_{2,0}} & \dots & s^{k_{n,0}} \\ s^{k_{0,1}} & s^{k_{1,1}} & s^{k_{2,1}} & \dots & s^{k_{n,1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s^{k_{0,d}} & s^{k_{1,d}} & s^{k_{2,d}} & \dots & s^{k_{n,d}} \end{bmatrix}$$

$$=$$

$$\begin{bmatrix} \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{0,j}} & \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{1,j}} & \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{2,j}} & \dots & \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{n,j}} \\ \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{0,j}} & \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{1,j}} & \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{2,j}} & \dots & \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{n,j}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{0,j}} & \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{1,j}} & \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{2,j}} & \dots & \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{n,j}} \end{bmatrix}$$

For each row $i \in n$, We can show that

$$\sum_{j=0}^{d} s^{q_{i,j}} s^{k_{0,j}} + \sum_{j=0}^{d} s^{q_{i,j}} s^{k_{1,j}} + \sum_{j=0}^{d} s^{q_{i,j}} s^{k_{2,j}} + \cdots + \sum_{j=0}^{d} s^{q_{i,j}} s^{k_{n,j}}$$

$$=$$

$$\sum_{j=0}^{d} s^{q_{i,j}} \left(\sum_{i=0}^{n} s^{k_{i,j}} \right) = 1$$

Therefore, we can prove $\text{softmax}_{row}(\boldsymbol{Q})\text{softmax}_{col}(\boldsymbol{K})^T$ prove valid categorical distribution in every row.

(b) The $\text{softmax}_{row}(\boldsymbol{Q})$ result a time complexity as $\Theta(nd)$ and space complexity $\Theta(2nd)$. Same for the $\text{softmax}_{col}(\boldsymbol{K})$.

In the $attention_{separable}$, instead of calculating the $\text{softmax}_{row}(\boldsymbol{Q})\text{softmax}_{col}(\boldsymbol{K}) \in \mathbb{R}^{n\times n}$. we can calculate $\boldsymbol{K}^T \in \mathbb{R}^{d\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times d}$ first. This step results a time complexity as $\Theta(nd^2)$ and space complexity $\Theta(2nd + d^2)$.

After the multiplication with $\boldsymbol{Q} \in \mathbb{R}^{n\times d}$results a matrix has the time complexity as (nd^2) and space complexity as $(d^2 + 2nd)$. We can see that the time and space complexity increase linearly as n increase. While in $attention_{std}$, the time and space complexity increase quadratically with n.

So, if $n \gg d$, the $attention_{separable}$ is easier to compute in terms of time and space efficiency(both increase quadratic of the d rather than n) than $\text{Attention}_{\text{std}}$.

(c) The $attention_{separable}$ is not as expressive as the $attention_{std}$ since for each element in the ith row, $attention_{separable}$ use the information up to dth value.

$$\begin{bmatrix} \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{0,j}} & \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{1,j}} & \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{2,j}} & \dots & \sum_{j=0}^{d} s^{q_{0,j}} s^{k_{n,j}} \\ \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{0,j}} & \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{1,j}} & \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{2,j}} & \dots & \sum_{j=0}^{d} s^{q_{1,j}} s^{k_{n,j}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{0,j}} & \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{1,j}} & \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{2,j}} & \dots & \sum_{j=0}^{d} s^{q_{n,j}} s^{k_{n,j}} \end{bmatrix}$$

while in $attention_{std}$, the $\text{softmax}_{row}(\boldsymbol{QK}^T)$ use information between $\boldsymbol{Q}, \boldsymbol{K}$ up to nth value.

3.3 We need to prove that writing out the individual rows of Q and K

$$\boldsymbol{Q} = \begin{bmatrix} -\boldsymbol{q}_0- \\ -\boldsymbol{q}_1- \\ \vdots \\ -\boldsymbol{q}_n- \end{bmatrix}$$

$$\boldsymbol{K}^T = \begin{bmatrix} -\boldsymbol{k}_0- \\ -\boldsymbol{k}_1- \\ \vdots \\ -\boldsymbol{k}_n- \end{bmatrix}^T$$

$$\begin{bmatrix} \text{softmax}(\boldsymbol{q}_0\boldsymbol{k}_0, \boldsymbol{q}_0\boldsymbol{k}_1, \boldsymbol{q}_0\boldsymbol{k}_2, \dots, \boldsymbol{q}_0\boldsymbol{k}_n) \\ \text{softmax}(\boldsymbol{q}_1\boldsymbol{k}_0, \boldsymbol{q}_1\boldsymbol{k}_1, \boldsymbol{q}_1\boldsymbol{k}_2, \dots, \boldsymbol{q}_1\boldsymbol{k}_n) \\ \dots \\ \text{softmax}(\boldsymbol{q}_n\boldsymbol{k}_0, \boldsymbol{q}_n\boldsymbol{k}_1, \boldsymbol{q}_n\boldsymbol{k}_2, \dots, \boldsymbol{q}_n\boldsymbol{k}_n) \end{bmatrix}$$

$=$

$$\begin{bmatrix} s^{q_{0,0}k_{0,0}} & s^{q_{0,1}k_{0,1}} & s^{q_{0,2}k_{0,2}} & \dots & s^{q_{0,n}k_{0,n}} \\ s^{q_{1,0}k_{1,0}} & s^{q_{1,1}k_{1,1}} & s^{q_{1,2}k_{1,2}} & \dots & s^{q_{1,n}k_{1,n}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s^{q_{n,0}k_{n,0}} & s^{q_{n,1}k_{n,1}} & s^{q_{n,2}k_{n,2}} & \dots & s^{q_{n,n}k_{n,n}} \end{bmatrix}$$

$$\boldsymbol{A} = \exp\left(\boldsymbol{QK}^\top\right)$$

$=$

$$\begin{bmatrix} exp(q_{0,0}k_{0,0}) & exp(q_{0,1}k_{0,1}) & exp(q_{0,2}k_{0,2}) & \dots & exp(q_{0,d}k_{0,d}) \\ exp(q_{1,0}k_{1,0}) & exp(q_{1,1}k_{1,1}) & exp(q_{1,2}k_{1,2}) & \dots & exp(q_{1,d}k_{1,d}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ exp(q_{n,0}k_{n,0}) & exp(q_{n,1}k_{n,1}) & exp(q_{n,2}k_{n,2}) & \dots & exp(q_{n,d}k_{n,d}) \end{bmatrix}$$

$\boldsymbol{A}$ has shape $n * d$

$$\boldsymbol{A1}$$

$=$

$$\begin{bmatrix} exp(q_{0,0}k_{0,0}) & exp(q_{0,1}k_{0,1}) & exp(q_{0,2}k_{0,2}) & \dots & exp(q_{0,d}k_{0,d}) \\ exp(q_{1,0}k_{1,0}) & exp(q_{1,1}k_{1,1}) & exp(q_{1,2}k_{1,2}) & \dots & exp(q_{1,d}k_{1,d}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ exp(q_{n,0}k_{n,0}) & exp(q_{n,1}k_{n,1}) & exp(q_{n,2}k_{n,2}) & \dots & exp(q_{n,d}k_{n,d}) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$=$

$$\begin{bmatrix} exp(q_{0,0}k_{0,0}) + exp(q_{0,1}k_{0,1}) + exp(q_{0,2}k_{0,2}) + \dots + exp(q_{0,d}k_{0,d}) \\ exp(q_{1,0}k_{1,0}) + exp(q_{1,1}k_{1,1}) + exp(q_{1,2}k_{1,2}) + \dots + exp(q_{1,d}k_{1,d}) \\ \vdots \\ exp(q_{n,0}k_{n,0}) + exp(q_{n,1}k_{n,1}) + exp(q_{n,2}k_{n,2}) + \dots + exp(q_{n,d}k_{n,d}) \end{bmatrix}$$

$$\boldsymbol{D} = \text{diag}(\boldsymbol{A}\boldsymbol{1})$$

$=$

$$\begin{bmatrix} \sum_{j=0}^{d} exp(q_{0,j}k_{0,j}) & 0 & 0 & \dots & 0 \\ 0 & \sum_{j=0}^{d} exp(q_{1,j}k_{1,j}) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sum_{j=0}^{d} exp(q_{n,j}k_{n,j}) \end{bmatrix}$$

$$\boldsymbol{D}^{-1}$$

$=$

$$\begin{bmatrix} \frac{1}{\sum_{j=0}^{d} exp(q_{0,j}k_{0,j})} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sum_{j=0}^{d} exp(q_{1,j}k_{1,j})} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{\sum_{j=0}^{d} exp(q_{n,j}k_{n,j})} \end{bmatrix}$$

$$\boldsymbol{D}^{-1}\boldsymbol{A}$$

$=$

$$\begin{bmatrix} \frac{1}{\sum_{j=0}^{d} exp(q_{0,j}k_{0,j})} & 0 & \dots & 0 \\ 0 & \frac{1}{\sum_{j=0}^{d} exp(q_{1,j}k_{1,j})} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & \dots & \frac{1}{\sum_{j=0}^{d} exp(q_{n,j}k_{n,j})} \end{bmatrix} \begin{bmatrix} exp(q_{0,0}k_{0,0}) & exp(q_{0,1}k_{0,1}) & \dots & exp(q_{0,d}k_{0,d}) \\ exp(q_{1,0}k_{1,0}) & exp(q_{1,1}k_{1,1}) & \dots & exp(q_{1,d}k_{1,d}) \\ \vdots & \vdots & \vdots & \ddots \\ exp(q_{n,0}k_{n,0}) & exp(q_{n,1}k_{n,1}) & \dots & exp(q_{n,d}k_{n,d}) \end{bmatrix}$$

$=$

$$\begin{bmatrix} \frac{exp(q_{0,0}k_{0,0})}{\sum_{j=0}^{d} exp(q_{0,j}k_{0,j})} & \frac{exp(q_{0,1}k_{0,1})}{\sum_{j=0}^{d} exp(q_{0,j}k_{0,j})} & \dots & \frac{exp(q_{0,d}k_{0,d})}{\sum_{j=0}^{d} exp(q_{0,j}k_{0,j})} \\ \frac{exp(q_{1,0}k_{1,0})}{\sum_{j=0}^{d} exp(q_{1,j}k_{1,j})} & \frac{exp(q_{1,1}k_{1,1})}{\sum_{j=0}^{d} exp(q_{1,j}k_{1,j})} & \dots & \frac{exp(q_{1,d}k_{1,d})}{\sum_{j=0}^{d} exp(q_{1,j}k_{1,j})} \\ \vdots & \vdots & \vdots & \ddots \\ \frac{exp(q_{n,0}k_{n,0})}{\sum_{j=0}^{d} exp(q_{n,j}k_{n,j})} & \frac{exp(q_{n,1}k_{n,1})}{\sum_{j=0}^{d} exp(q_{n,j}k_{n,j})} & \dots & \frac{exp(q_{n,d}k_{n,d})}{\sum_{j=0}^{d} exp(q_{n,j}k_{n,j})} \end{bmatrix} = \text{softmax}_{\text{row}}\left(\boldsymbol{Q}\boldsymbol{K}^\top\right)$$

Therefore,$\text{softmax}_{\text{row}}\left(\boldsymbol{Q}\boldsymbol{K}^\top\right)\boldsymbol{V} = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{V}$

We can decompose the matrix $\boldsymbol{A}$ into $f(\boldsymbol{Q})$ and $f(\boldsymbol{K})$. By implementing the mapping $\boldsymbol{q}_i\boldsymbol{k}_j$ to $f(\boldsymbol{q}_i)^\top f(\boldsymbol{k}_j)$, we can change the order of the matrix multiplications as follwoing : $f(\boldsymbol{Q})(f(\boldsymbol{K})f(\boldsymbol{V}))$ the time dimension can be reduced from $\Theta(n^2d)$ to $\Theta(nmd)$ and space complexity can be reduced from $\Theta(n^2+2nd)$ to $\Theta(2nm+nd)$. We avoid the usage of the quadratic of the input length.

*3.4

$$\boldsymbol{x}^\top\boldsymbol{y} = -\frac{1}{2}(\boldsymbol{x}^\top\boldsymbol{x} - (\boldsymbol{x}+\boldsymbol{y})^\top(\boldsymbol{x}+\boldsymbol{y}) + \boldsymbol{y}^\top\boldsymbol{y})$$

Hence

$$\exp(\boldsymbol{x}^\top\boldsymbol{y}) = \exp(-\frac{1}{2}(\boldsymbol{x}^\top\boldsymbol{x}))\exp(\frac{1}{2}(\boldsymbol{x}+\boldsymbol{y})^\top(\boldsymbol{x}+\boldsymbol{y}))\exp(-\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{y})$$

=

$$\exp(-\frac{||\boldsymbol{x}||^2}{2})\exp(\frac{||\boldsymbol{x}+\boldsymbol{y}||^2}{2})\exp(-\frac{||\boldsymbol{y}||^2}{2})$$

Let $\boldsymbol{x} = \boldsymbol{q}_i$ and $\boldsymbol{y} = \boldsymbol{k}_j$, we have $a_{ij} = \exp(\boldsymbol{q}^\top\boldsymbol{k}) = \exp(-\frac{||\boldsymbol{q}_i||^2}{2})\exp(\frac{||\boldsymbol{q}_i+\boldsymbol{k}_j||^2}{2})\exp(-\frac{||\boldsymbol{k}_j||^2}{2})$

Since

$$p(\boldsymbol{x}) = (2\pi)^{-d/2}\exp\left(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$$

and

$$\int_{\boldsymbol{x}} p(\boldsymbol{x})d\boldsymbol{x} = 1$$

we have

$$\int_{\boldsymbol{x}}(2\pi)^{-d/2}\exp\left(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)d\boldsymbol{x} = (2\pi)^{-d/2}\int_{\boldsymbol{x}}\exp\left(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)d\boldsymbol{x} = 1$$

for any $u \in d$, we can have

$$\exp(\frac{||\boldsymbol{q}_i+\boldsymbol{k}_j||^2}{2}) = (2\pi)^{-d/2}\exp(\frac{||\boldsymbol{q}_i+\boldsymbol{k}_j||^2}{2})\int_{\boldsymbol{x}}\exp\left(-\frac{1}{2}\|\boldsymbol{x}-(\boldsymbol{q}_i+\boldsymbol{k}_j)\|^2\right)d\boldsymbol{x}$$

=

$$(2\pi)^{-d/2}\int_{\boldsymbol{x}}\exp\left(-\frac{1}{2}\|\boldsymbol{x}-(\boldsymbol{q}_i+\boldsymbol{k}_j)\|^2 + (\frac{||\boldsymbol{q}_i+\boldsymbol{k}_j||^2}{2})\right)d\boldsymbol{x}$$

=

$$(2\pi)^{-d/2}\int_{\boldsymbol{x}}\exp\left(-\frac{||\boldsymbol{x}||^2}{2} - \frac{||\boldsymbol{q}_i+\boldsymbol{k}_j||^2}{2} + \boldsymbol{x}^T(\boldsymbol{q}_i+\boldsymbol{k}_j) + \frac{||\boldsymbol{q}_i+\boldsymbol{k}_j||^2}{2}\right)d\boldsymbol{x}$$

=

$$(2\pi)^{-d/2}\int_{\boldsymbol{x}}\exp\left(-\frac{||\boldsymbol{x}||^2}{2} + \boldsymbol{x}^T(\boldsymbol{q}_i+\boldsymbol{k}_j)\right)d\boldsymbol{x}$$

=

$$(2\pi)^{-d/2}\int_{\boldsymbol{x}}\exp(-\frac{||\boldsymbol{x}||^2}{2})\exp(\boldsymbol{x}^T\boldsymbol{q}_i)\exp(\boldsymbol{x}^T\boldsymbol{k}_j)d\boldsymbol{x}$$

=

$$\mathbb{E}_{\boldsymbol{x}\in\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\exp(\boldsymbol{x}^\top\boldsymbol{q}_i)\exp(\boldsymbol{x}^\top\boldsymbol{k}_j)\right]$$

Therefore,

$$a_{ij} = \exp(\boldsymbol{q}^\top \boldsymbol{k}) = \exp(-\frac{||\boldsymbol{q}_i||^2}{2})\exp(\frac{||\boldsymbol{q}_i + \boldsymbol{k}_j||^2}{2})\exp(-\frac{||\boldsymbol{k}_j||^2}{2})$$

$$=$$

$$\exp\left(\frac{-\|\boldsymbol{q}_i\|^2}{2}\right) \cdot \mathbb{E}_{\boldsymbol{x}\in\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\exp(\boldsymbol{x}^\top \boldsymbol{q}_i)\exp(\boldsymbol{x}^\top \boldsymbol{k}_j)\right] \cdot \exp\left(\frac{-\|\boldsymbol{k}_j\|^2}{2}\right)$$

We can get m numbers of $\boldsymbol{x}$ from a normal distribution with mean 0 and variance $\mathbf{I}$ to get the approximate value for $\boldsymbol{A}$ with $f(\boldsymbol{q}_i)^T f(\boldsymbol{k}_j)$, where

$$f(\boldsymbol{q}_i) = \exp\left(\frac{-\|\boldsymbol{q}_i\|^2}{2}\right) \cdot \mathbb{E}_{\boldsymbol{x}\in\mathcal{N}(\mathbf{0},\mathbf{I})}\exp(\boldsymbol{w}^\top \boldsymbol{x})$$

$$=$$

$$\exp\left(\frac{-\|\boldsymbol{q}_i\|^2}{2}\right) \frac{1}{m}(\exp(\boldsymbol{w}_1^\top \boldsymbol{q}_i), \ldots, \exp(\boldsymbol{w}_m^\top \boldsymbol{q}_i))$$

$$f(\boldsymbol{k}_j) = \exp\left(\frac{-\|\boldsymbol{k}_i\|^2}{2}\right) \cdot \mathbb{E}_{\boldsymbol{x}\in\mathcal{N}(\mathbf{0},\mathbf{I})}\exp(\boldsymbol{w}^\top \boldsymbol{x})$$

$$=$$

$$\exp\left(\frac{-\|\boldsymbol{k}_j\|^2}{2}\right) \frac{1}{m}(\exp(\boldsymbol{w}_1^\top \boldsymbol{k}_i), \ldots, \exp(\boldsymbol{w}_m^\top \boldsymbol{k}_j))$$

This will give us an approximation of $\boldsymbol{A}$ rather than calculate Attention$_{std}$. By doing this, we can get $f(\boldsymbol{q}_i), f(\boldsymbol{k}_j)$ independently and allow $f(\boldsymbol{k}_j)$ to multiply the $\boldsymbol{V}$ first to reduce the time and space complexity.

3.5 When the m for Attention$_{\text{approx}}$ is larger, the $\mathbb{E}_{\boldsymbol{x}\in\mathcal{N}(\mathbf{0},\mathbf{I})}\exp(\boldsymbol{w}^\top \boldsymbol{x})$ will be more accurate so that the $f(\boldsymbol{q}_i), f(\boldsymbol{k}_j)$ more close to the real $\boldsymbol{Q}, \boldsymbol{K}$,but the time complexity $\Theta(nmd)$ and space complexity $\Theta(2nm + nd)$ will increase linearly as well, which makes the operation slower and need more space.

Question 4. Answer

4.1 The *L2 regularization* scheme uses a standard SGD update rule is :

$$\theta_{i+1} = \theta_i - \eta \frac{\partial L(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)}) + \gamma \|\theta\|_2^2}{\partial \theta_i}$$

$=$

$$\theta_{i+1} = \theta_i - \eta (\frac{\partial L(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)})}{\partial \theta_i} + 2\gamma\theta_i)$$

$=$

$$\theta_{i+1} = \theta_i - \eta \frac{\partial L(f(\mathbf{x}^{(i)}, \theta), \mathbf{y}^{(i)})}{\partial \theta_i} - 2\eta\gamma\theta_i$$

Let $2\eta\gamma = \lambda$, the weight decay scheme employs the modified SGD update is the same as the L2 regularization scheme that employs a standard SGD update rule.

4.2 The one-line in Algorithm 8.7 of the deep learning book that needs to change to give us Adam with an L2 regularization scheme is :

$$\text{Compute gradient : } \boldsymbol{g} \leftarrow \frac{1}{m} \nabla_\theta \sum_i L(f(\mathbf{x}^{(i)}, \boldsymbol{\theta}), \mathbf{y}^{(i)})$$

The corresponding modifications is

$$\text{Compute gradient : } \boldsymbol{g} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}_{\mathbf{t-1}}} \sum_i L(f(\mathbf{x}^{(i)}, \boldsymbol{\theta}_{t-1}), \mathbf{y}^{(i)}) + \lambda \boldsymbol{\theta}_{t-1}$$

4.3 (a) θ_{small} will be regularized more strongly than the other. With L_2 regularization, the sums of the gradient of the loss function and the gradient of the regularizer (i.e., the L_2 norm of the weights) are adapted. In this scenario,

$$\boldsymbol{\Delta}\theta \leftarrow -\eta \frac{\alpha \hat{\mathbf{m}}}{\sqrt{\hat{v} + \epsilon}}$$

when t is large, then

$$\boldsymbol{\Delta}\theta \leftarrow -\eta \frac{\alpha(\beta_1 m_{t-1} + (1-\beta_1)(\nabla f_t + \lambda \boldsymbol{\theta}_{t-1})}{\sqrt{\hat{v} + \epsilon}}$$

With **Adam-L2-reg**, both types of gradients are normalized by $\sqrt{\hat{v}}$. For θ_{large}, the corresponding $\sqrt{\hat{v}}$ is large too and the weight is regularized less than θ_{small}.. Therefore θ_{large} are regularized by a smaller relative amount than θ_{small}.

(b) Yes. There are not difference between the regularization between θ_{small} and θ_{large} with the **Adam-weight-decay**. Since only the gradients of the loss function are adapted (with the weight decay step separated from the adaptive gradient mechanism). **Adam-weight-decay** regularizes all weights with the same rate λ, effectively regularizing θ_{large} than standard L2 regularization.

4.4 Weight-decay is better for the following reasons :

(a) Weight Decay performs equally well on both SGD and adaptive gradient methods like Adam.

(b) While L2 regularization can only work with standard SGD. If combined with adaptive gradients, L2 regularization leads to weights with large historic parameter and/or gradient amplitudes being regularized less than they would be when using weight decay.

(c) Regularization is to prevent over-fitting. It is one method of reducing overfitting by allowing weights to be close to zero. the model's complexity. In the context of an adaptive gradient based optimizer, the goal of using L2 Regularization is not achieved since the parameters with large magnitudes will be updated slowly and regulated less.