My solutions to

Deep Learning: Foundations and Concepts

Dario Miro Konopatzki

12 Transformers

12.2

For any $x_k, x_l \in \mathbb{R}^D$, $x_k^\top x_l \in \mathbb{R}$ and thus $e^{x_k^\top x_l} > 0$. Hence $a_{nm} = \frac{\overbrace{e^{x_n^\top x_m}}^{>0}}{\sum_{m'=1}^N \underbrace{e^{x_n^\top x_{m'}}}_{>0 \text{ f.a. } m'}} > 0$.

$$\begin{aligned}
\sum_{m=1}^N a_{nm} &= \sum_{m=1}^N \frac{e^{x_n^\top x_m}}{\sum_{m'=1}^N e^{x_n^\top x_{m'}}} \\
&= \frac{\cancel{\sum_{m=1}^N e^{x_n^\top x_m}}}{\cancel{\sum_{m'=1}^N e^{x_n^\top x_{m'}}}} \\
&= 1
\end{aligned}$$

12.4

$$\begin{aligned}
\mathbb{E}\left[\left(a^\top b\right)^2\right] &= \mathbb{E}\left[\left(\sum_{d=1}^D a_d b_d\right)^2\right] \\
&= \mathbb{E}\left[\sum_{d=1}^D (a_d b_d)^2 + \sum_{d=1}^D \sum_{\substack{d'=1 \\ d' \neq d}}^D a_d b_d a_{d'} b_{d'}\right] \\
&= \sum_{d=1}^D \left(\mathbb{E}\left[a_d^2 b_d^2\right] + \sum_{\substack{d'=1 \\ d' \neq d}}^D \mathbb{E}\left[a_d b_d a_{d'} b_{d'}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&\stackrel{\star}{=} \sum_{d=1}^{D} \left(\mathbb{E}\left[a_d^2\right] \mathbb{E}\left[b_d^2\right] + \sum_{\substack{d'=1 \\ d' \neq d}}^{D} \mathbb{E}\left[a_d\right] \mathbb{E}\left[b_d\right] \mathbb{E}\left[a_{d'}\right] \mathbb{E}\left[b_{d'}\right] \right) \\
&\stackrel{\dagger}{=} \sum_{d=1}^{D} \left(1 \cdot 1 + \sum_{\substack{d'=1 \\ d' \neq d}}^{D} 0 \cdot 0 \cdot 0 \cdot 0 \right) \\
&= D
\end{aligned}
$$

- $\star$ By assumption $a, b \sim \mathcal{N}(0, \mathbb{I})$. Diagonal covariance implies components of a (resp. b) independent. Taken together with a, b independent this gives $a_1, \ldots, a_D, b_1, \ldots, b_D$ independent.
- $\dagger$ By assumption $a \sim \mathcal{N}(0, \mathbb{I})$, so $\mathbb{E}[a_d] = 0$ and $\mathbb{E}[a_d^2] = \mathbb{E}\left[aa^\top\right]_{dd} = \mathbb{I}_{dd} = 1$ for all d. The analogous result holds for b.

12.5

Let σ denote softmax and Att attention. Then

$$
\begin{aligned}
Y &= \left[\mathrm{Att}\left(Q_h, K_h, V_h\right)\right]_h W^{(o)} \\
&= \sum_{h=1}^{H} \mathrm{Att}\left(Q_h, K_h, V_h\right) W_h^{(o)} \\
&= \sum_{h=1}^{H} \sigma\left(\frac{Q_h K_h^\top}{\sqrt{D_k}}\right) V_h W_h^{(o)} \\
&= \sum_{h=1}^{H} \sigma\left(\frac{Q_h K_h^\top}{\sqrt{D_k}}\right) X W_h^{(v)} W_h^{(o)} \\
&= \sum_{h=1}^{H} \sigma\left(\frac{Q_h K_h^\top}{\sqrt{D_k}}\right) X W^{(h)} \\
&= \sum_{h=1}^{H} \mathrm{Att}\left(Q_h, K_h, X W^{(h)}\right)
\end{aligned}
$$

12.7

Let $\pi : N \to N$ be any permutation, $P := \left(\delta_{\pi(n)n'}\right)_{nn'}$. Note that $P^\top P = \left(\delta_{\pi(n)n'}\right)_{n'n} \left(\delta_{\pi(n)n'}\right)_{nn'} = \left(\sum_n \delta_{\pi(n)n'} \delta_{\pi(n)n''}\right)_{n'n''} = \left(\delta_{n'n''}\right)_{n'n''} = \mathbb{I}_N$.

Consider any $A \in \mathbb{R}^{N\times N}$. Then

$$\begin{aligned}
\sigma(PA) &= \sigma\left(\left(\sum_{n'} \delta_{\pi(n)n'} A_{n'n''}\right)_{nn''}\right) \\
&= \sigma\left(\left(A_{\pi(n)n''}\right)_{nn''}\right) \\
&= \left(\frac{e^{A_{\pi(n)n''}}}{\sum_{n'} e^{A_{\pi(n)n'}}}\right)_{nn''} \\
&= \left(\sigma(A)_{\pi(n)n''}\right)_{nn''} \\
&= \left(\sum_{n'} \delta_{\pi(n)n'} \sigma(A)_{n'n''}\right)_{nn''} \\
&= P\sigma(A)
\end{aligned}$$

and

$$\begin{aligned}
\sigma\left(AP^\top\right) &= \sigma\left(\left(\sum_{n'} A_{nn'} \delta_{\pi(n'')n'}\right)_{nn''}\right) \\
&= \sigma\left(\left(A_{n\pi(n'')}\right)_{nn''}\right) \\
&= \left(\frac{e^{A_{n\pi(n'')}}}{\sum_{n'} e^{A_{n\pi(n')}}}\right)_{nn''} \\
&= \left(\sigma(A)_{n\pi(n'')}\right)_{nn''} \\
&= \left(\sum_{n'} \sigma(A)_{nn'} \delta_{\pi(n'')n'}\right)_{nn''} \\
&= \sigma(A)P^\top
\end{aligned}$$

Since A arbitrary, for every h where $(W^{(q)}, W^{(k)}, W^{(v)} := W_h^{(q)}, W_h^{(k)}, W_h^{(v)})$

$$\begin{aligned}
\mathrm{Att}\left(PXW^{(q)}, PXW^{(k)}, PXW^{(v)}\right) &= \sigma\left(\frac{PXW^{(q)}\left(PXW^{(k)}\right)^\top}{\sqrt{D_k}}\right) PXW^{(v)} \\
&= \sigma\left(\frac{PXW^{(q)}W^{(k)\top}X^\top P^\top}{\sqrt{D_k}}\right) PXW^{(v)} \\
&= P\sigma\left(\frac{XW^{(q)}W^{(k)\top}X^\top}{\sqrt{D_k}}\right) P^\top PXW^{(v)} \\
&= P\sigma\left(\frac{XW^{(q)}\left(XW^{(k)}\right)^\top}{\sqrt{D_k}}\right) \mathbb{I}_N XW^{(v)} \\
&= \mathrm{Att}\left(XW^{(q)}, XW^{(k)}, XW^{(v)}\right)
\end{aligned}$$

and since the above holds for all h

$$\left[\mathrm{Att}\left(PXW_h^{(q)}, PXW_h^{(k)}, PXW_h^{(v)}\right)\right]_h W^{(o)} = P\left[\mathrm{Att}\left(Q_h, K_h, V_h\right)\right]_h W^{(o)}$$

12.9

Let $x \in \mathbb{R}^D, e \in \mathbb{R}^K, W \in \mathbb{R}^{(D+K)\times M}$ and define

$$\bar{x} := \begin{bmatrix} x \\ e \end{bmatrix}, \text{ i.e. } \bar{x}_l = \begin{cases} x_l & \text{if } 1 \le l \le D, \\ e_{l-D} & \text{if } D+1 \le l \le D+K \end{cases} \quad \text{for all } 1 \le l \le D+K.$$

Then for any $1 \le m \le M$:

$$\begin{aligned}
\left(\bar{x}^\top W\right)_m &= \sum_{l=1}^{D+K} \bar{x}_l W_{l,m} \\
&= \sum_{l=1}^{D} \bar{x}_l W_{l,m} + \sum_{l=D+1}^{D+K} \bar{x}_l W_{l,m} \\
&= \sum_{l=1}^{D} x_l W_{l,m} + \sum_{l=D+1}^{D+K} e_{l-D} W_{l,m} \\
&= \sum_{l=1}^{D} x_l W_{l,m} + \sum_{l=1}^{K} e_l W_{D+l,m} \\
&= x^\top W_{1:D,m} + e^\top W_{D+1:D+K,m}
\end{aligned}$$

i.e.

$$\begin{aligned}
&\begin{bmatrix} x_1 & \cdots & x_D & e_1 & \cdots & e_K \end{bmatrix} \begin{bmatrix} W_{1,1} & \cdots & W_{1,M} \\ \vdots & \ddots & \vdots \\ W_{D,1} & \cdots & W_{D,M} \\ W_{D+1,1} & \cdots & W_{D+1,M} \\ \vdots & \ddots & \vdots \\ W_{D+K,1} & \cdots & W_{D+K,M} \end{bmatrix} \\
&= \bar{x}^\top W \\
&= x^\top W_{1:D,:} + e^\top W_{D+1:D+K,:} \\
&= \begin{bmatrix} x_1 & \cdots & x_D \end{bmatrix} \begin{bmatrix} W_{1,1} & \cdots & W_{1,M} \\ \vdots & \ddots & \vdots \\ W_{D,1} & \cdots & W_{D,M} \end{bmatrix} + \begin{bmatrix} e_1 & \cdots & e_K \end{bmatrix} \begin{bmatrix} W_{D+1,1} & \cdots & W_{D+1,M} \\ \vdots & \ddots & \vdots \\ W_{D+K,1} & \cdots & W_{D+K,M} \end{bmatrix}
\end{aligned}$$