Puneet - Mangla
CS17BTECH11029.

Q.1 a)

Time complexity	RNN	Transformer
train	tlx^2	t^2lx
test	tlx^2	t^2lx.

Space complexity	RNN	Transformer
train	tlx	tlx
test	lx	tlx.

b) RNNs is sequential ~~because of~~ bottleneck depends on sequence length. whereas for a transformer it only depends on number of layers. Hence when sequences are longer - small-layer transformers can learn much faster than vanilla RNNs

c) Yes, it is bottleneck. The sequential operations in transformers are independent of length but are quite complex / expensive to decode. It make it a tradeoff.

d) No, feed forward network and layer norm don't look across tokens. However, various operations / paths can be executed parallely while flowing through FF and norm layers.

Date : / /
Page No.:

Q.2 a) $z = v_j$ implies $\alpha_j = 1$ and $\alpha_{i:\, i \neq j} = 0$

This means that $K_j^T q >>> K_{i:\, i \neq j}^T q$

b) ~~Assume~~

$$q = \sum_{j=1,m} \beta_j K_j$$

now $K_i^T q = K_i^T \left(\sum_j \beta_j K_j \right)$

$$= \sum_j \beta_j K_i^T K_j$$

$$= \beta_i \|K_i\|^2 + 0 \quad \text{---} \quad \{ \text{as } K_i \text{ are orthonormal}$$

$$= \beta_i \times 1$$

so $\alpha_i = \dfrac{\exp(\beta_i)}{\sum_{j=1}^{m} \exp(\beta_j)}$

$z \approx \frac{1}{2}(v_a + v_b) \Rightarrow \alpha_a = \alpha_b = \frac{1}{2}$

and $\alpha_{i:\, i \neq a \text{ and } i \neq b} = 0$

setting $\beta_a = \beta_b >> 0$ and $\beta_{i:\, i \neq a \text{ and } i \neq b} << 0$

in query vector q will achieve the same.

This will give $\alpha_a = \alpha_b \approx \frac{1}{2}$

and $\alpha_{i:\, i \neq a \text{ and } i \neq b} \approx 0$

Date : / /
Page No.:

Q.3

$$\mathcal{L}(q) = \int q(z|x) \log \frac{p(x,z)}{q(z|x)} dz$$

$$= \int q(z|x) \log \frac{p(x|z)\,p(z)}{q(z|x)} dz$$

{ Approximating q by q_ϕ, p by p_θ

$$= \mathop{E}_{z \sim q_\phi(z|x)} \left[\log \frac{p_\theta(z)}{q_\phi(z|x)} + \log p_\theta(x|z) \right]$$

$$= \mathop{E}_{z \sim q_\phi(z|x)} \left[-\log \frac{q_\phi(z|x)}{p_\theta(z)} \right] + \mathop{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x|z)]$$

$$= -KL\left(q_\phi(z|x) \,||\, p_\theta(z)\right) + \mathop{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x|z)]$$

$$= \mathop{E}_{z \sim q_\phi(z|x_i)} [\log p_\theta(x|z)] - KL\left(q_\phi(z|x) \,||\, p(z)\right)$$

reconstruction term — regularization term. reconstruction

Here $\mathop{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x|z)]$ is the ~~regularization~~ term which ~~aligns the~~ forces the generated image to be similar to original one.

Here $KL\left(q_\phi(z|x) \,||\, p_\theta(z)\right)$ is the regularization term which aligns the distribution $q_\phi(z|x)$ with prior $p_\theta(z)$

Date : / /
Page No.:

Q.4 $f(p,q) = pq$ $\min_p \max_q f(p,q)$

$\frac{d f(p,q)}{dq} = p$ will evaluate to 0 as $\max_q pq \geq 0$

hence only setting $p \geq 0$ will minimize $\max_q pq$

a)

q_0	q_1	q_2	q_3	q_4	q_5	q_6
1	2	1	-1	-2	-1	1
p_0	p_1	p_2	p_3	p_4	p_5	p_6
1	-1	-2	-1	1	2	1

b) No it is not possible to reach optimal value as after 6 iterations the new value of p, q are same as p_0, q_0. make it oscillatory sequence.

It is because the step size which is 1 is too large. Since min-max problems are unstable, small step size is preferable.

c) Nash equilibrium point is 0.

$$f_t = f_{t+1}$$

$$p_t q_t = -(q_t)(p_t + q_t)$$

$$q_t(2p_t + q_t) = 0$$

$\Rightarrow 2p_t = -q_t$ or $q_t = 0$

↓ Not valid solution (see part a)

↓ Hence $q_t = 0 \Rightarrow p_{t+1} = 0$ and p, q from then $= 0$

$$\begin{cases} p_{t+1} = -q_t \\ q_{t+1} = p_t + q_t \end{cases}$$