Name - T.Vijay
Roll.no - CS17BTECH11040

Time complexity

Q1 a) RNN → $t * l * n^2$ (Both train & test)

Transformer → $t^2 * l * n$ (Both train & test)

(b) Space complexity

RNN (Train) → $t \times l \times n$
(Test) → $l \times n$

Transformer → $t \times l \times n$ (Train & test)

b) $$\frac{\text{Time of RNN}}{\text{Time of Transform}} = \frac{t l n^2}{t^2 l n} = \frac{n}{t}$$

∴ If $n < t$, RNN can be faster.

Scanned with CamScanner

c) Self attention layer looks across the tokens of a given input sequence parallelly. It copies the input sequence for each token and processes the copied sequence in parallel. Hence its not a bottleneck.

d) Since output of self attention layer can be calculated parallely, feedforward network can be parallelized because it is applied to each output vector of the self attention module.

Norm layer needs timestamp to normalize the output vector and it looks across tokens. So its a bottleneck.

Scanned with CamScanner

Q2 a) Given: $z = \sum_{i=1}^{M} v_i \alpha_i$

$\therefore z = v_j$ then

$$\alpha_i = \begin{cases} 0 & \forall i \neq j \\ 1 & i = j \end{cases}$$

$\therefore$ For j, $\alpha_j = 1 = \dfrac{\exp(k_i^T q)}{\sum_{i=1}^{m} \exp(k_i^T q)}$

$\therefore k_j^T q >>> k_i^T q \quad \forall i \neq j$

b) Given :- $z \simeq \frac{1}{2}(v_a + v_b)$ & $z = \sum_{i=1}^{M} v_i \alpha_i$

$\therefore \frac{1}{2} \approx \alpha_a$ & $\alpha_b \approx \frac{1}{2}$ and $\alpha_i = 0 \; \forall i \notin \{a, b\}$

Since $\{k_1 \dots k_m\}$ are orthogonal

We set $q = x_a k_a + x_b k_b$

where $x_a = x_b >>> 0$

$\therefore$ We can easily verify that

$\alpha_a \approx \alpha_b \approx \frac{1}{2}$ on substitution in the eq^n

$$\alpha_i = \frac{\exp(k_i^T q)}{\sum_{i=1}^{m} \exp(k_i^T q)}$$

Q3 $$L(q) = \int q(z/x) \log\left(\frac{p(x,z)}{q(z/x)}\right) dz$$

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

$$= \underbrace{\int q(z/x) \log(p(x/z)) dz}_{\text{Reconstruction loss}}$$

$$\underbrace{- \int q(z/x) \log\left(\frac{q(z/x)}{p(z)}\right) dz}_{\text{Regularization term}}$$

On calculation, this can be reduced to

$$L(q) = \underbrace{E_{z\sim q(z/x)} \left[\log p(x/z)\right]}_{\text{Reconstruction loss for given input}} - \underbrace{KL(q(z/x), p(z))}_{\text{KL divergence Regularization loss.}}$$

Reconstruction loss for given input

KL divergence Regularization loss.

Q4. a) Maximize $f(p_t, q_t) = p_t q_t$ wrt q_t

$$\frac{df}{dq_t} = p_t$$

$\because$ step size $= 1$

$$q_{t+1} = q_t + \frac{df}{dq_t} = p_t + q_t \quad \text{—} (1)$$

Minimise $f(p_t, q_{t+1}) = p_t q_{t+1}$ wrt p_t

$$\frac{df}{dp_t} = q_{t+1}$$

$$p_{t+1} = p_t - \frac{df}{dp_t} = p_t - q_{t+1} \quad (2)$$

$$= -q_t \quad (\text{from } (1))$$

$$f_{t+1}(p_{t+1}, q_{t+1}) = (p_{t+1})(q_{t+1}) = -q_t(p_t + q_t)$$

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

Scanned with CamScanner

b) It is not possible to find optimal value because $(p_0, q_0) = (p_6, q_6)$

It is periodic and does not converge
∴ We ~~could~~ can not find optimal solution

To find optimal solution, we need to reduce the step size

c) Equilibrium point $\Rightarrow$ $(0,0)$

as $f_t = f_{t+1}$

$q_t = q_{t+1}$, $p_t = p_{t+1}$

Scanned with CamScanner