Assignment 5

Vinta Reethu
ES18BTECH11028

1.

a) Given that, t = sequence length
l = number of layers
n = number of neurons

Time complexity :

RNN - Train time $= tn^2l$
Test time $= tn^2l$

Transformer - Train time $= t^2nl$
Test time $= t^2nl$

Space complexity :

RNN - Train time $= tnl$
Test time $= nl$

Transformer - Train time $= tnl$
Test time $= tnl$

b)
- The time taken for computation is t^2nl in transformers and tn^2l in RNN.
- If the value of n is smaller than sequence length t, time taken by RNN will be less than time taken by transformer.
- Transformers in this case perform worstly because the computations at self attention layer are more than the normal feed forward network.

c)
- Self attention layer looking across the ~~text~~ tokens of a given input sequence is bottleneck for parallelism.
- Generally, the sequential operations in transformers are independent of sequence length. But these are expensive to decode.
- Parallel processing is what that makes transformer learn faster than RNN for longer sequences.

d)
- The feed forward network and layer norm do not look across the tokens.
- They only look at output content vector of self attention layer
- In this way parallelism is introduced as feed forward network work in parallel.

2.

$$z = \sum_{i=1}^{m} (v_i \alpha_i)$$

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

a) Given,

$z = v_j$

which means that $\alpha_j = 1$

and $\alpha_i = 0 \ \forall i \neq j$

This means that $k_j^T q >>> k_{i;i\neq j}^T q$

b) Given, $\{k_1, k_2, \ldots k_m\} \rightarrow k_i \perp k_j \ \forall i \neq j$ and $\|k_i\| = 1 \ \forall i$

Let $q = \sum_{j=1}^{m} B_j k_j$

Let's evaluate $k_i^T q = k_i^T \left(\sum_j B_j k_j\right) = \sum_j B_j k_i^T k_j$

$= B_i \underbrace{\|k_i\|^2}_{1} + B_i \left[\underbrace{\qquad}_{0}\right]$ as k_i's are $\perp r$ to $k_j \ \forall i \neq j$

$= B_i$

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

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

and $\alpha_i = 0 \ \forall i \neq a$ and $i = b$

- we can obtain $\alpha_a \approx \frac{1}{2}$, $\alpha_b \approx \frac{1}{2}$ by setting $B_a = B_b >> 0$ & $B_i << 0 \ \forall i \neq a$ & $i \neq b$

is the query vector q

and

$\alpha_i \approx 0 \ \forall i \neq a$ and $i \neq b$

3. $L(q) = \int q(\frac{z}{x}) \log \frac{P(x,z)}{q(\frac{z}{x})} dz$

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

$$= E_{z \sim q_\phi(\frac{z}{x})} \log \left[\frac{P_\theta(\frac{x}{z})\, P_\theta(z)}{q(\frac{z}{x})}\right]$$

$$= E_{z \sim q_\phi(\frac{z}{x})} \left[-\log \frac{q(\frac{z}{x})}{P_\theta(z)} + \log P_\theta(\frac{x}{z})\right]$$

$$= E_{z \sim q_\phi(\frac{z}{x})} [\log P_\theta(\frac{x}{z})] + E_{z \sim q_\phi(\frac{z}{x})} \left[-\log \frac{q(\frac{z}{x})}{P_\theta(z)}\right]$$

$$= \underbrace{E_{z \sim q_\phi(\frac{z}{x})} [\log P_\theta(\frac{x}{z})]}_{\text{reconstruction loss}} - \underbrace{KL[q(\frac{z}{x}), P(z)]}_{\text{regularisation term}}$$

- KLD measures similarity between two distributions
- reconstruction error here is the MLE of decoder.

4. Given that, $f(P,q) = Pq$

$\min_P \max_q f(P,q)$ — ?

a) $K = (1,1)$; $n = 6$; step $= 1$

Maximizing wrt q_t

$$\frac{\partial f}{\partial q_t} = \frac{\partial (P_t q_t)}{\partial q_t} = P_t$$

- $q_{t+1} = P_t + q_t$ [gradient update] —— ①
- Hence $f' = P_t q_{t+1}$
- Minimizing wrt P_t :-

$$\frac{\partial f'}{\partial P_t} = \frac{\partial (P_t q_{t+1})}{\partial P_t} = q_{t+1}$$

- $P_{t+1} = P_t - 1(q_{t+1})$ [gradient update]

↑ substituting in Eq ①

4.

- $p_{t+1} = (\cancel{q_{t+1}} - q_t) - \cancel{q_{t+1}}$

  $= -q_t$

- $\therefore f_{t+1} = (-q_t)(p_t + q_t)$

- Using the above to fill the table

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)
- With the step = 1, as we can see above there is no convergence. hence it is not possible to find the optimal value.
- To find the optimal value we can change the value of step size.

c)
- It attains equibrium when the values remain same (f)

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

$$\cancel{p_t q_t = (-q_t)(p_t + q_t)}$$

$$\cancel{-p_t = p_t + q_t}$$

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

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

$$\therefore q_t = 0 \quad \text{or} \quad p_t = \frac{-q_t}{2}$$

- Here at point 2, 5 $p_t = \frac{-q_t}{2}$ which didnot lead to equilibrium.
- Hence q_t should be 0, which means $p_{t+1} = 0$.

  $\downarrow$

  this or above condition should be satisfied to obtain equilibrium.