Akash Tadwai
Indian Institute of Technology Hyderabad
Deep Learning for Vision
ES18BTECH11019

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Assignment-V

Akash Tadwai - ES18BTECH11019

April 14, 2021

1 Computational Complexity and Parallelism

1.1 RNNs vs Transformers

We use the following notations:

$$\begin{aligned} n &= No.\ of\ neurons \\ l &= No.\ of\ layers \\ t &= No.\ of\ timesteps \end{aligned} \tag{1}$$

1.1.1 Time Complexity

Architecture	Train	Test
RNN	$t \times l \times n^2$	$t \times l \times n^2$
Transformer	$t^2 \times l \times n$	$t^2 \times l \times n$

1.1.2 Space Complexity

Architecture	Train	Test
RNN	$t \times l \times n$	$l \times n$
Transformer	$t \times l \times n$	$t \times l \times n$

1.2 Performance on Parallelism

- As in the case of RNNs the computational bottleneck is proportional to $t \times l \times n^2$ whereas in Transformers it is proportional to $t^2 \times l \times n$.
- When $n < t$ then the transformers perform worse than RNNs because there is a lot of computation going on at the self-attention layer than the feed-forward layer.

Akash Tadwai
Indian Institute of Technology Hyderabad
Deep Learning for Vision
ES18BTECH11019

1.3 Self Attention Layers and Bottleneck of Parallelism

- Yes, Self-attention layer looking across the tokens of a given input sequence is a bottleneck for Parallelism.There is a trade-off between the sequencial operations and decoding complexity.
- The sequential operations in transformers are independent of sequence length, but they are very expensive to decode.Transformers can learn faster than RNNs on parallel processing hardwards for longer sequences.

1.4 Feed-forward and norm layers

- **No**, the feed forward and norm layers doesn't look across the tokens, they only look at the output of context-vector of self-attention layer. So this is the way parallelism is induced in Transformers as feed forward layers work in parallel.

2 Attention Model and Orthogonality

2.1 Part A

- If $z = v_j$ then a possible scenario is the following:

$$\begin{aligned} \alpha_j &= 1 \\ \alpha_i &= 0; \forall i \neq j \end{aligned} \tag{2}$$

- This scenario is possible when the query vector is aligned with one of the key vectors and dot product with other vectors is very low.

$$\begin{aligned} k_j &\parallel q \\ k_{i:i\neq j}^T q &\ll k_j^T q \end{aligned} \tag{3}$$

2.2 Part B

- Let us assume that the query vector belongs to span of key vectors:

$$q = \sum_{i=1}^{m} \beta_i k_i \tag{4}$$

- Now, we can write

$$\begin{aligned} k_i^\top q &= k_i{}^\top \left(\sum_j B_j k_j \right) \\ &= \sum_j B_j k_i{}^\top k_j \\ &= B_i \left\| k_i \right\|^2 + 0 \\ &= B_i \times 1 \end{aligned}$$

- $$\begin{aligned} \alpha_i &= \frac{\exp{(B_i)}}{\sum_{j=1}^{m} \exp{(B_j)}} \\ z \approx \frac{1}{2}(v_a + v_b) &\Rightarrow \alpha_a = \alpha_b \gg \alpha_{i \neq a,b} \end{aligned} \tag{5}$$

- So we have the following and setting $\beta_{a,b} \gg \beta_{i:i \neq a,b}$ and $\beta_a \approx \beta_b$:

$$\begin{aligned} \alpha_a &= \frac{\exp{\beta_a}}{\exp{\beta_a} + \exp{\beta_b} + (\ldots)} \\ &\approx \frac{\exp{\beta_a}}{\exp{\beta_a} + \exp{\beta_b}} \\ &\approx \frac{1}{2} \end{aligned} \tag{6}$$

- Hence by setting $\beta_{a,b} \gg \beta_{i:i \neq a,b}$ and $\beta_a \approx \beta_b$ and $\beta_{i:i \neq a,b} \ll 0$ we can achieve $z \approx \frac{v_a + v_b}{2}$

3 VAE Loss

$$\begin{aligned} \mathcal{L}(q) &= \int q(z|x) \log \frac{p(x,z)}{q(z|x)} \, dz \\ &= \int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} \, dz \\ &= \mathbb{E}_{z \sim q_\phi(z|x_i)}[\log \frac{p_\theta(x|z)p_\theta(z)}{q(z|x)}] \\ &= \mathbb{E}_{z \sim q_\phi(z|x_i)}[-\log \frac{q(z|x)}{p_\theta(z)} + \log p_\theta(x|z)] \\ &= \mathbb{E}_{z \sim q_\phi(z|x_i)}[\log p_\theta(x|z)] + \mathbb{E}_{z \sim q_\phi(z|x_i)}[-\log \frac{q(z|x)}{p_\theta(z)}] \\ &= \mathbb{E}_{z \sim q_\phi(z|x_i)}[\log p_\theta(x|z)] - KL(q(z|x), p(z)) \\ &= \underbrace{\mathbb{E}_{z \sim q_\phi(z|x_i)}[\log p_\theta(x|z)]}_{\text{Reconstruction error}} - \underbrace{KL(q(z|x), p(z))}_{\text{Regularization term}} \end{aligned} \tag{7}$$

- The $\mathbb{E}_{z \sim q_\phi(z|x_i)}[\log p_\theta(x|z)]$ term acts reconstruction since it is the maximum likelihood estimate of decoder.

- The $-KL(q(z|x), p(z))$ term acts like a regularizer here, because KL divergence measures the similarity between two distributions.

Akash Tadwai
Indian Institute of Technology Hyderabad
Deep Learning for Vision
ES18BTECH11019

4 GAN

- The optimization problem is:

$$\begin{aligned} f(p,q) &= pq \\ Obj &= min_p max_q(f(p,q)) \end{aligned} \tag{8}$$

4.1 Table

- Expressing the values of p_{t+1} and q_{t+1} in terms of p_t and q_t.
- First let us maximize with respect to q_t:

$$\begin{aligned} \frac{\partial f}{\partial q_t} &= \frac{\partial (p_t q_t)}{\partial q_t} \\ &= p_t \end{aligned} \tag{9}$$

$$\Rightarrow q_{t+1} = p_t + q_t \tag{10}$$

- Since we take a unit step, the final form will be:

$$f' = p_t q_{t+1} \tag{11}$$

- Now minimizing wrt p_t:

$$\begin{aligned} \frac{\partial f'}{\partial p_t} &= \frac{\partial p_t (q_{t+1})}{\partial p_t} \\ &= q_{t+1} \end{aligned} \tag{12}$$

$$\begin{aligned} p_{t+1} &= p_t - (q_{t+1}) \\ &= (q_{t+1} - q_t) - q_{t+1} \, (From\ eqn\ 7) \\ &= -q_t \end{aligned} \tag{13}$$

- Since we take a unit step, the final form will be:

$$f_{t+1} = -(q_t)(p_t + q_t) \tag{14}$$

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

Akash Tadwai
Indian Institute of Technology Hyderabad
Deep Learning for Vision
ES18BTECH11019

4.2 Reaching Optimal Value

- It is evident from the above table that the values oscillate and becomes periodic. Hence they do not converge.
- With the given step size, it is not possible to find out the optimal value.In order to find out the optimal value we need to change the step size.

4.3 Equilibrium Point

- In the min-max game the condition for equilibrium is that the product remains constant .
- Hence mathematically, we can show as follows:

$$\begin{aligned} f_t &= f_{t+1} \\ p_t q_t &= -(q_t)(p_t + q_t) \\ 2p_t q_t &= -q_t^2 \\ q_t(2p_t + q_t) &= 0 \end{aligned} \tag{15}$$

  So either $2p_t = -q_t$ or $q_t = 0$.

- We saw from the table that $2p_t = -q_t$ does not lead to equilibrium, i.e., (p_1, q_1). Hence $q_t = 0$, so $p_{t+1} = 0$. So one of them should be zero.

For the LaTeXtyped version of this Report visit : https://www.overleaf.com/read/tngzhnsrrnjf

******THE END******