Homework 7

Deep Learning 2025 Spring

Due on 2025/4/21

1 Q&A

Problem 1. (Transformer) Consider a vanilla transformer proposed in [1].

1. Which block is the most computationally expensive part of a vanilla transformer?
2. In practice, training a vanilla transformer requires a much larger GPU memory compared to RNN. We have learned in the lecture that we can reduce the computation cost of RNN via truncated BPTT. How can we utilize this practical trick for transformers? Try to design an algorithm and write down the pseudo-code.

   Hint: You can read this paper.

Problem 2. Parallel WaveNet models a sequence $x = (x_1, x_2, ..., x_L)$ using a sequence of latent variables $z = (z_1, z_2, ..., z_L) \sim \mathcal{N}(0, I)$, where each output token is generated as:

$$x_i = \mu_\theta(z_{1:i-1}) + z_i \cdot \exp(\alpha_\theta(z_{1:i-1}))$$

This defines a non-autoregressive generation process using a reparameterization trick. Construct this process as a normalizing flow with L layers, where each layer transforms a single variable z_i into x_i using the formula above. Explicitly define the transformation at each layer, and explain how this composition results in a valid flow.

Problem 3. Prove that the speculative decoding procedure preserves the target model's distribution. Specifically, show that for any candidate sequence $y_{1:k}$ generated by the draft model q, the joint distribution after speculative decoding satisfies:

$$\pi(y_{1:k} \mid x) = p(y_{1:k} \mid x),$$

where $p(y_{1:k} \mid x)$ is the target model's distribution, $q(y_{1:k} \mid x)$ is the draft model's distribution, $\pi(y_{1:k} \mid x)$ is the final distribution after speculative decoding.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.