0.1 Introduction

TODO

Refer to tables

From (todo), it is clear that the Transformer models are memory-intensive due to their quadratic complexity in terms of the dataset size. This is a significant bottleneck for scaling up the Transformer models to larger datasets with limited compute resources. To address this issue, several methods have been proposed to reduce the memory complexity of the Transformer models. In this chapter, we will discuss some of the methods that have been proposed to reduce the memory complexity of the Transformer NPs.

0.2 Pseudotokens

Pseudotokens (aka Inducing Points) are a set of tokens that are used to approximate the full set of tokens, it can be thought of as a lower dimensional representation of the data set. Consider projecting the original tokens $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ into a lower-dimensional space $\boldsymbol{U} \in \mathbb{R}^{M \times D}$ where $M << N$ through some translation equivariant network [Ashman et al. 2024]. The pseudotokens $\boldsymbol{I} \in \mathbb{R}^{M \times D}$ can then be used to perform cross-attention with the original tokens $\boldsymbol{X}$ thus reducing the memory complexity to $\mathcal{O}(M(N_c + M_t))$ making the model linear with respect to context and target set size.

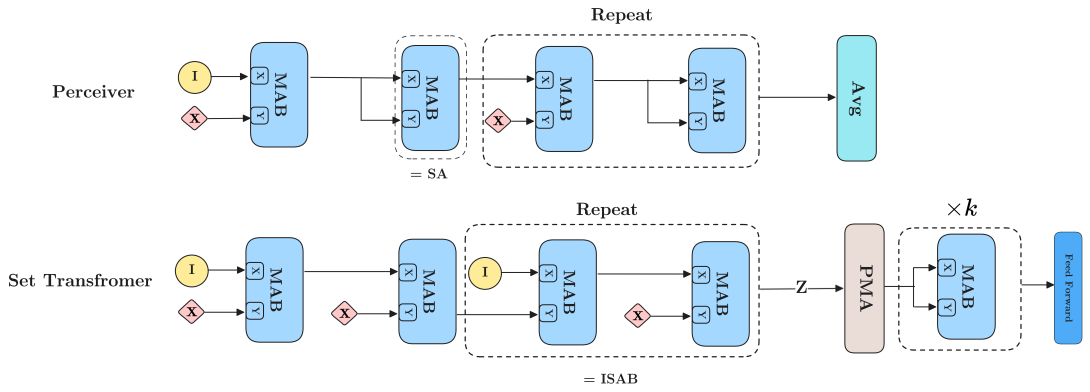

Figure 0.2.1: Perceiver vs Set Transformer. MAB is equivalent to Multi-Head Cross Attention, refer to [Ashman et al. 2024] for more details.

We consider two implementations of a pseudotokens based Transformer, one is the Set Transformer [Lee et al. 2019] and the other is the Perceiver [Jaegle et al. 2021]. Both models are very similar and only really differ in the way they implement the cross-attention mechanism between the orginal and pseudo tokens. [Feng et al. 2023] implemented the Perciever model in the context of NPs creating the 'Latent Bottled Attention Neural Process' (LBANP) model. Ashman et al. 2024 implemented the Set Transformer model into a NP creating the 'Inducing Set Transformer' (IST) model. Due to the similarity of the models, we expect both to have very similar performance on the tasks they are evaluated on.

0.3 Linear Transformer

0.4 HyperMixer

0.5 Experimental Results

Bibliography

Ashman, Matthew et al. (2024). "Translation-Equivariant Transformer Neural Processes". In: *Forty-first International Conference on Machine Learning*. URL: https://openreview.net/forum?id=pftXzp6Yn3.

Feng, Leo et al. (2023). *Latent Bottlenecked Attentive Neural Processes*. arXiv: 2211.08458 [cs.LG].

Jaegle, Andrew et al. (2021). *Perceiver: General Perception with Iterative Attention*. arXiv: 2103.03206 [cs.CV].

Lee, Juho et al. (2019). *Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks*. arXiv: 1810.00825 [cs.LG].