1 Transformer Neural Processes

1.1 Introduction

Transformer Neural Processes (TNPs) [NG23] were introduced to overcome the limitations of the Latent Neural Process (LNP) [Gar+18] which have the following limitations:

- LNPs have intracatable likelihoods hence they have to be trained using variational inference by optimizing for the ELBO. It has been shown that the ELBO can be a poor approximation of the true likelihood hence the model can be poorly trained.
- LNPs tend to underfit the data and are unable to capture the complexity of the data.

TNPs overcome these limitations by using the Transformer [Vas+17] to utilize attention to learn the relationships between the context points. This allows the model to capture the complexity of the data and also allows the model to be trained using maximum likelihood estimation (MLE) instead of variational inference as the likelihood is tractable.

Attention has already been implemented prior to [NG23] in [Kim+19] however according to [NG23] the attention mechanism used in [Kim+19] 'tend to make overconfident predictions and have poor performance on sequential decision making problems'.

1.2 Requirements

As NPs are a meta learning method it should be trained on a context set which is a subset of a larger dataset. In this example, say we sample N $x-y$ pairs from a process $\mathcal{F}$, we designate the first N_c pairs as the context set and the remaining $N-N_c$ pairs as the target set, or mathematically $\mathcal{C}=\{(x_i,y_i)\}_{i=1}^{N_c}$ and $\mathcal{T}=\{(x_i,y_i)\}_{i=N_c+1}^{N}$. The objective function is the maximum likelihood estimation of the target set given the context set autorergressively, or mathematically on the target set.

$$\mathcal{L}(\theta)=\mathbb{E}_{(x,y)\sim\mathcal{F}}\left[\log p(y_{N_c+1:N}|x_{N_c+1:N},\mathcal{C};\theta)\right] \tag{1}$$

Where θ are the parameters of the model and $y_{N_c+1:N}$ and $x_{N_c+1:N}$ are the target set. Since we pass the target set through the model autoregressively, the objective function can be rewritten as:

$$\mathcal{L}(\theta)=\mathbb{E}_{(x,y)\sim\mathcal{F}}\left[\sum_{i=N_c+1}^{N}\log p(y_i|x_i,\mathcal{C},x_{N_c+1:i-1},y_{N_c+1:i-1};\theta)\right] \tag{2}$$

References

[Gar+18] Marta Garnelo et al. *Neural Processes.* 2018. arXiv: 1807.01622 [cs.LG].

[Kim+19] Hyunjik Kim et al. *Attentive Neural Processes.* 2019. arXiv: 1901.05761 [cs.LG].

[NG23] Tung Nguyen and Aditya Grover. *Transformer Neural Processes: Uncertainty-Aware Meta Learning Via Sequence Modeling.* 2023. arXiv: 2207.04179 [cs.LG].

[Vas+17] Ashish Vaswani et al. *Attention Is All You Need.* 2017. arXiv: 1706.03762 [cs.CL].