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 before [NG23] in [Kim+19] however according to [NG23] the attention mechanism used in [Kim+19] tends to make overconfident predictions and have poor performance on sequential decision-making problems.

1.2 Requirements

As NPs are a meta-learning method they 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 autoregressively, 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:

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

Where each conditional is modeled as a Gaussian distribution with mean and variance predicted by the model.

There are two important properties of TNPs that we need to build upon:

- **Context Invariance**: If we permute the context set, the model should not change its predictions
- **Target Equivariance** If we permute the target set, the model should not change its predictions, e.g. we shift the target set by one point to the left, and the model's predictions should also shift by one point to the left.

1.3 Autoregressive Transformer Neural Processes (TNP-A)

We can not implement the traditional Transformer as the inclusion of the positional encoding breaks the context invariance property since the positional encoding is dependent on the position of the points, but we need to allow the model to learn the relationships between x and y **pairs**. To do this, [NG23] concatenates the x and y pairs together into a single vector for the context set and target set. They then add auxiliary tokens from the target set but with $y=0$ such that the dataset looks like

$$(x_1,y_1),...,(x_N,y_N),(x_{N_c+1},0),...,(x_N,0)$$

As we can see the x values from the target set effectively appear twice, the auxiliary tokens are used to represent the target set and then the corresponding y values are used to train the model since those tokens are not auxiliary. To make this work, we must employ a masking scheme to allow:

- Context tokens to attend to themselves
- Target tokens to attend to context and previous target tokens

- Auxillary tokens to attend to context and previous target tokens

This model is referred to as the Autoregressive Transformer Neural Process (TNP-A) in [NG23].

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].