Unifying Transformers and Convolutional Neural Processes

Name: Lakee Sivaraya, College: Emmanuel College, Supervisor: Prof. Richard E. Turner

Neural Processes (NPs) are a class of general purpose models designed for uncertaincy-aware meta learning. Knowledge of the uncertaincy in a prediction is crucial for real-world applications such as climate modelling, time series forecasting and decision-making in healthcare.

We highlight the limitations of the standard DeepSet based NP model, particularly its inability to model complicated relationships in the data in high dimensions. To extend the capabilities of NPs, we replace the DeepSet with two powerful architectures: Convolutional Neural Networks (CNNs) and Transformers giving rise to Convolutional Neural Processes (ConvNPs) and Transformer Neural Processes (TNPs) respectively. Both models are bound to have their own strengths and weaknesses.

Our experiments demonstrate that the TNP outperforms the ConvNP, especially in higher dimensions. We observed that far outside the training data, the TNP behaves unpredictably unlike the ConvNP which is more stable. We hypothesize the stability of the ConvNP is due to the *translation equivariance* property of CNNs. Translation equivariance is the property of a model to produce the same output when the input is translated, such property is greatly beneficial when modelling stationary or periodic data. As the TNP does not have this property it generalizes poorly to data far from the training data.

To address the limitations of the TNP, we introduce translation equivariance to the TNP, by modifying the self-attention mechanism in the TNP. We call this model the *Translation Equivariant Transformer Neural Process* (TETNP). Our experiments show the TETNP outperforming the TNP and ConvNP on datasets that exhibit strong discontinuities and periodicity. On smooth datasets, the TETNP performs comparably to the TNP which is expected since these datasets do not require translation equivariance.

The quadratic computational complexity of full self attention is a key bottleneck in the TNP and TETNP models, limiting their scalability to large datasets. This thesis explores several methods to reduce the computational complexity of the TNP by the use of pseduotokens, and approximations to the self-attention mechanism via kernelization and MLPs. While our experiments demonstrates a substantial reduction in computational complexity, the approximations come at the cost of performance. All linear models struggled to model the discontinuous datasets, highlighting the loss of expressiveness in the model. Future work could explore implementing a hybrid model that combines the strengths of pseduotokens and attention approximations to achieve even greater computational efficiency.

Overall this thesis provides a comprehensive study of the TNP and ConvNP models. We establish the importance of translation equivariance in the TNP and demonstrate its benefits. Overall the TETNP model is the most promising model, with strong performance across a range of datasets especially in higer dimensions. ConvNPs scale terribly with dimensionality, making the TNP and TETNP a clear choice for large scale applications in high dimensions. Future work could explore the use of TETNPs in real-world applications

such as climate modelling and image processing.