Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning

https://arxiv.org/pdf/2106.02584.pdf

tl;dr: instead of predicting the output based on model parameters and a single input, use the entire dataset

Important Points

- 1. Parametric: $p(y * \mid x * ; \theta)$ Non-parametric: $p(y * \mid x * , Dtrain)$
- 2. This paper can be seen as a combination of the two
- 3. Predicts entire masked features/targets matrix (self-supervised learning and supervised learning, respectively)
- 4. Must use minibatch for large datasets

Overall idea

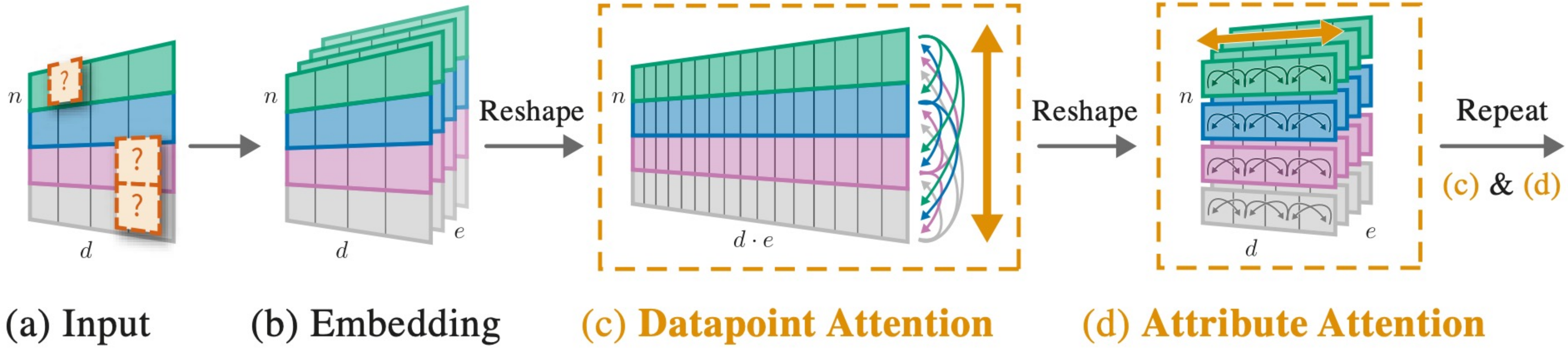

Figure 2: Overview of the Non-Parametric Transformer. (a) The input dataset and mask matrix are stacked and (b) linearly embedded for all datapoints independently. NPT then applies (c) Attention Between Datapoints (ABD, §2.4) across all n samples of hidden dimension $h = d \cdot e$. (d) Attention Between Attributes (ABA, §2.5) then attends between the attributes for each datapoint independently. We repeat steps (c) and (d) and obtain a final prediction from a separate linear projection (not shown).

My thoughts

- NPT achieves 68.2% accuracy on CIFAR-10 and 98.3% accuracy on MNIST
- Even though experiments are conducted on CIFAR-10 and MNIST, there are no comparisons with convolutional architectures or Transformers, but only comparisons with non-parametric models.
- The paper’s official explanation is “we perform no pre-training, and therefore a direct comparison of our results to this line of work is inappropriate”

A quote from Ferenc Huszar

- Unfortunately I was not able to find this thread again on twitter, ergo I could only paraphrase rather than screenshot
- “I would like to know that this project began with accidentally setting (dim=0) to (dim=1) somewhere”