Self-Attention Between Datapoints

Going Beyond Individual Input-Output Pairs in Deep Learning






Jannik Kossen* @janundnik

Neil Band* @neilbband

Clare Lyle @clarelyle

Aidan N. Gomez @AidanNGomez

Tom Rainforth @tom_rainforth

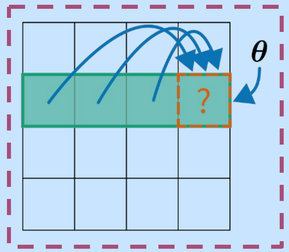
Yarin Gal @yaringal

(*equal contribution)

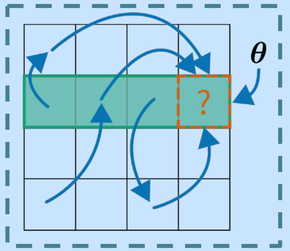
OATML

UNIVERSITY OF OXFORD

OxCSML



Most of deep learning relies on **parametric prediction**. But **what if** it didn't?



Non-Parametric Transformers can be parametric and/or non-parametric.

In practice, they **learn to rely on non-parametric mechanisms**, achieving SOTA performance on tabular datasets.

Summary

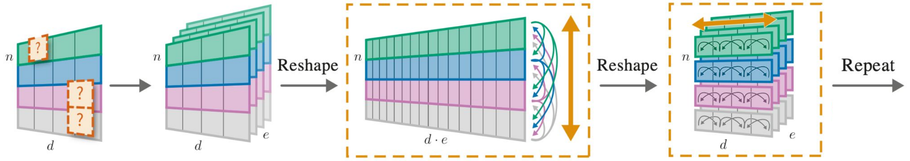
We introduce **Non-Parametric Transformers** (NPTs). NPTs ...

... take the **entire dataset as input**. (We approximate this with minibatches for large datasets.)
... use multi-head self-attention to **predict from relationships between datapoints**.
... rely on (stochastic/deterministic) **masking** to form a reconstruction loss objective.
... can be used for class./regression/missing data/self- and semi-supervised and transductive learning.

Experimentally, NPTs ...

... show **strong performance on tabular datasets**, and promising performance for image classification.
... **learn to rely** on other datapoints for prediction.
... can solve complex **reasoning tasks**.

Non-Parametric Transformers



Model Architecture

- **Goal:** predict $\mathbf{p(X^M \mid X^O)}$
- **Input:** entire dataset **X**, binary masking matrix **M**
- **Embedding:** per-attribute linear embedding, applied independently to each datapoint
- **Multi-Head Self-Attention:** applied between datapoints and independently between attributes of each datapoint (standard attention)

Properties

✅ **Learns complex relationships** between datapoints
✅ Learns **transformations** of **individual** datapoints
✅ **Equivariant to a permutation** of the datapoints

Objective

$$\mathcal{L}^{\text{NPT}} = \underbrace{(1-\lambda)\mathcal{L}^{\text{Targets}}}_{\text{Take targets of training datapoints as input}} + \underbrace{\lambda\mathcal{L}^{\text{Features}}}_{\text{Regularizer}}$$

Tabular Data Experiments

Average Rank Performance

Binary Classification

Method	AUROC
NPT	**2.50 ± 0.87**
CatBoost	2.75 ± 0.85
LightGBM	3.50 ± 1.55
XGBoost	4.75 ± 1.25
Gradient Boosting	5.00 ± 0.71
MLP	5.75 ± 1.49
Random Forest	6.00 ± 0.71
TabNet	6.50 ± 1.32
k-NN	8.25 ± 0.48

Multi-Class Classification

Method	Accuracy
NPT	**2.50 ± 0.50**
XGBoost	**2.50 ± 1.50**
MLP	3.00 ± 2.00
CatBoost	3.50 ± 0.50
Gradient Boosting	3.50 ± 1.50
Random Forest	6.50 ± 0.50
TabNet	7.50 ± 0.50
LightGBM	7.50 ± 1.50
k-NN	8.50 ± 0.50

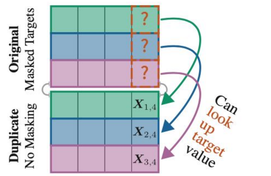
Regression

Method	RMSE
CatBoost	**3.00 ± 0.91**
XGBoost	3.25 ± 0.63
NPT	3.25 ± 1.31
Gradient Boosting	4.00 ± 1.08
Random Forest	4.50 ± 0.87
MLP	5.00 ± 1.22
LightGBM	6.50 ± 1.55
TabNet	6.75 ± 0.95
k-NN	8.75 ± 0.25

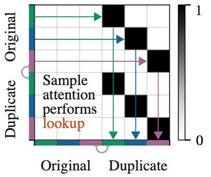
Permutation Test of Non-Parametricity

Δ *Accuracy*	CIFAR-10	Poker	Income	Higgs	MNIST	Forest	Kick	Breast Cancer
	−1.2	−1.1	−1.1	−0.5	−0.4	−0.1	−0.1	0.0
$\Delta_{RMSE}/_{RMSE}$ (%)	Yacht	Protein	Boston	Concrete				
	−52%	−21%	−20%	−7%				

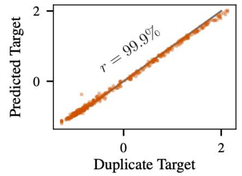
Semi-Synthetic Experiments



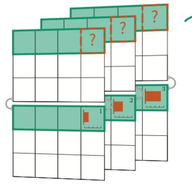
(a) Semi-Synthetic Input



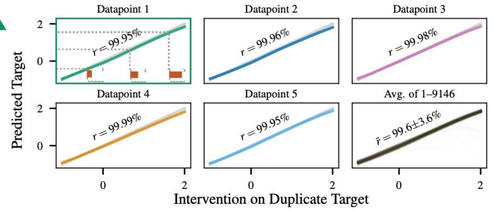
(b) Attention Weights



(c) Model Predictions



(d) Interventions on Duplicates



(e) Model Responses to Interventions