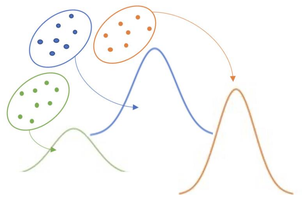
Towards a Neural Statistician

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

Trevor Clark, Lucia Lopez Rivilla, Ryan Anderson

Introduction

The **neural statistician** is a network that computes summary statistics of unordered sets of data. To do so, it extends the variational autoencoder (VAE) by including a latent context variable c, shared among items in a dataset.



The Neural Statistician

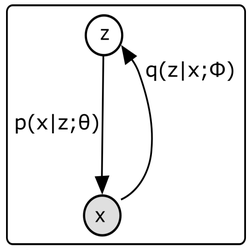


Figure: VAE

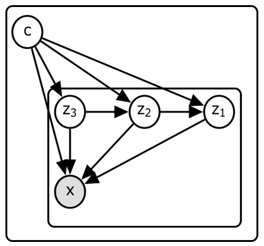


Figure: Full model

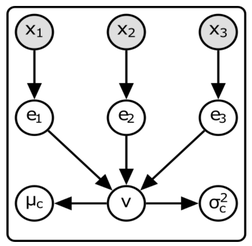


Figure: Statistic Network

Probabilistic model: $p(D) = \int p(c) \prod_{x \in D} \int p(x|c, z_{1:L}; \theta) p(z_L|c; \theta) \prod_{i=1}^{L-1} p(z_i|z_{i+1}|c; \theta)\, d_{z_{1:L}}\, dc$

Approximate posterior: $q(c, z_{1:L}|D; \phi) = q(c|D; \phi) \prod_{x \in D} q(z_L|x, c; \phi) \prod_{i=1}^{L-1} q(z_i|z_{i+1}, c, x; \phi)$

ELBO: $\mathcal{L}_D = \mathbb{E}_{q(c|D;\phi)} \left[\sum_{x \in D} \mathbb{E}_{q(z|c,x;\phi)} \log p(x|z;\theta) - D_{KL}\big(q(z|c,x;\phi) \| p(z|c;\theta)\big) \right] - D_{KL}\big(q(c|D;\phi) \| p(c)\big)$

$\mathcal{L}_D =$ Reconstruction Term - Context Divergence – Latent Divergence

Components

Shared encoder (optional): $x \mapsto h$

Statistic Network: $q(c|D;\phi) : \{h_1, h_2, \ldots, h_m\} \mapsto \mu_{c|D}, \sigma^2_{c|D}$

Inference Network: $q(z|x, c; \phi) : h, c \mapsto \mu_{z|x,c}, \sigma^2_{z|x,c}$

Latent Decoder Network: $p(z|c;\theta) : c \mapsto \mu_{z|c}, \sigma^2_{z|c}$

Observation Decoder Network: $p(x|c, z; \theta) : c, z \mapsto \mu_{x|c,z}, \sigma^2_{x|c,z}$

Experiments: Overview

1. Visualising the context.
2. Conditional sampling: Unlike training, use the mean of c instead of sampling c.
3. Specific tasks: Transfer learning, few-shot classification and summarising datasets.

Synthetic 1-D Distributions

Datasets consist of 200 samples from either an Exponential, Gaussian, Uniform or Laplacian distribution with equal probability.

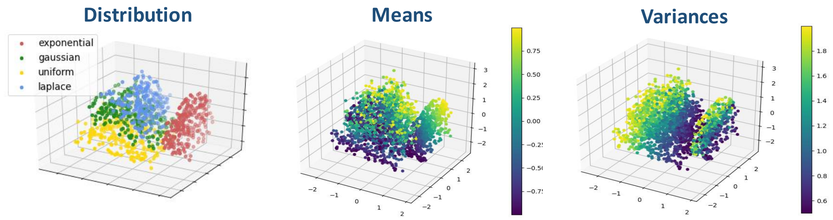


Figure: Each point is the mean of the approximate posterior over the context, coloured by distribution (left), means (middle) and variance (right).

OMNIGLOT - Transfer Learning

1628 classes of hand-written characters with 20 examples per class. Each class is divided into datasets of size 5. Transfer learning to unseen classes by training the model on a subset of OMNIGLOT'S 1628 classes.

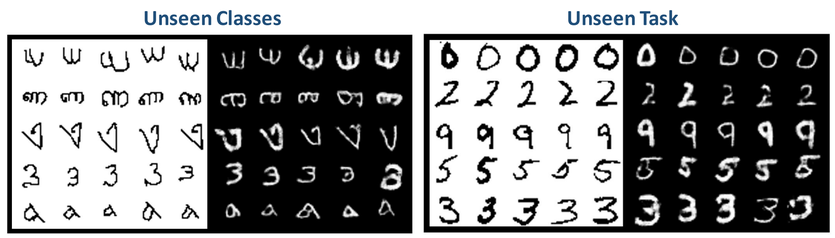


Figure: Generated samples (black) conditioned on unseen input classes (white) from OMNIGLOT (left) and MNIST (right).

YouTube Faces

Dataset: Images of 1595 faces. Each epoch, 5 images of each person in the training set are randomly sampled.

Proof of concept for generating faces.

Figure: Input faces — Faces conditioned on input — Generated faces from sampled contexts

K-Shot Classification

Trained on OMNIGLOT, test image *x* is classified to a seen dataset:

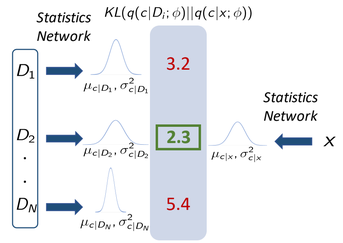


Table: The K-shot classification results for our replication and the Neural Statistician paper.

Task			Method	
Test Dataset	K-shot	K-way	Paper	Ours
MNIST	1	10	78.6	64.9
MNIST	5	10	93.2	84.7
OMNIGLOT	1	5	98.1	92.4
OMNIGLOT	5	5	99.5	97.8
OMNIGLOT	1	20	93.2	79.4
OMNIGLOT	5	20	98.1	92.9

Spatial MNIST

50 coordinates sampled using pixel intensities as probability densities:

T-sample summary of a dataset:

$S^* = \arg\min_{S \subseteq D} KL(q(c|D;\phi)||q(c|S;\phi))$

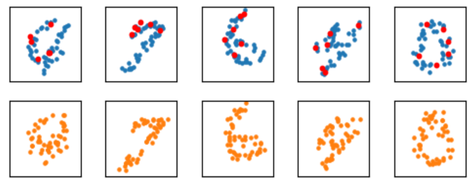


Figure: Orange points are samples conditioned on the blue points. Red points are 6-sample summaries.

"Sensible" Summaries?

200 coordinates sampled, 20-sample summaries.

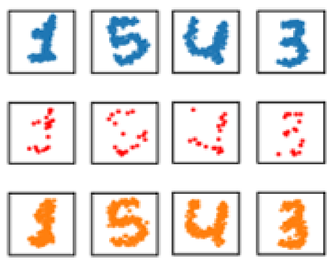


Figure: Orange points are samples conditioned on the blue points. Red points are 20-sample summaries.

Alternative Sampling

Sampling using *mean* of observation decoder.

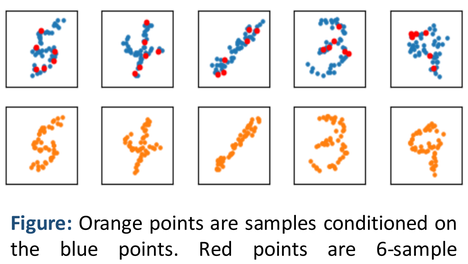


Figure: Orange points are samples conditioned on the blue points. Red points are 6-sample summaries.

Strengths/Weaknesses

The model is:

\+ Unsupervised, data efficient, parameter efficient, capable of few-shot learning.

\- Dataset hungry.

References

[1] H. Edwards and A. J. Storkey, "Towards a neural statistician," ArXiv, vol. abs/1606.02185, 2017.