UNIVERSITY OF CAMBRIDGE
Department of Engineering

Towards A Neural Statistician

Bourgin, V., Chang, Y., Shysheya, A.

GitHub Repo

Introduction

Neural Statistician [1] is an extension of variational autoencoder (VAE) as an unsupervised generative model that introduces a **dataset-level** latent variable $c_i \in \mathbb{R}^l$, referred to as a **context**. The context is used to learn *summary statistics* of **unordered datasets** $D_i = \{x_1, ..., x_j\}$.

Model Description

Vanilla Variational Autoencoder

VAE uses a latent variable model for **data-point** x, with latent $z \sim p(z)$ s.t.:

$$p(x) = \int p(x|z;\theta)p(z)dz$$

- **Encoder network:** $q(z|x;\phi)$
- **Decoder network:** $p(x|z;\theta)$

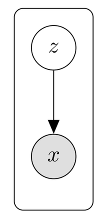
Figure: VAE model.

$\rightarrow$ Get **variational lower bound** (ELBO):

$$\log P(x|\theta) \geq \mathcal{L}_x = \mathbb{E}_{q(z|x,\phi)}[\log p(x|z;\theta)] - D_{KL}(q(z|x;\phi)||p(z))$$

Neural Statistician

Basic model: latent variable c shared for items in same **dataset**, s.t.:

$$p(D) = \int p(c)\left[\prod_{x\in D}\int p(x|z;\theta)p(z|c;\theta)dz\right]dc$$

The variational lower bound on the dataset:

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

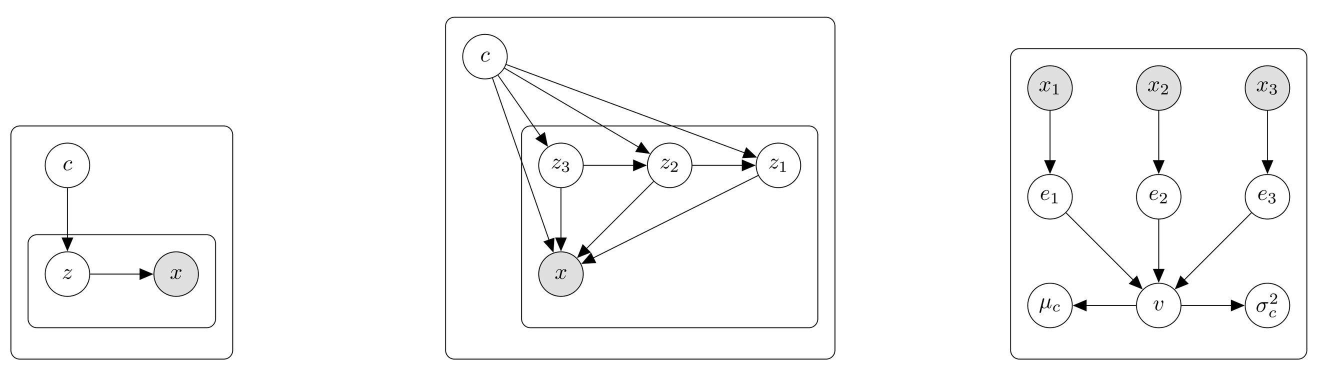
(a) Basic model (b) Full model (c) Statistic network

Figure: The Neural Statistician model.

Full model: for complex datasets, use multiple stochastic layers $z_{1:k}$ and skip-connections:

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

The full approximate posterior is now:

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

The variational lower bound for the **full model**:

$\mathcal{L}_D = R_D$ (reconstruction) $+C_D$ (context divergence) $+L_D$ (latent divergence)

Neural Statistician Building Blocks

- **Shared encoder** $x \rightarrow h$ *optional*
- **Statistic network** $q(c|D;\phi) : \{h_1, \ldots, h_m\} \rightarrow \mu_{c|D}, \sigma^2_{c|D}$
- **Inference network** $q(z|x, c;\phi) : h, c \rightarrow \mu_{z|x,c}, \sigma^2_{z|x,c}$
- **Latent decoder network** $p(z|c;\theta) : c \rightarrow \mu_{z|c}, \sigma^2_{z|c}$
- **Observation decoder network** $p(x|c, z;\theta) : c, z \rightarrow \mu_{x|c,z}, \sigma^2_{x|c,z}$

Synthetic 1-D Distributions

Aim: Demonstrate clustering of similar datasets.

We generate synthetic datasets consisting of samples from different distributions and we plot the summary statistics $\mu_{c|D}$ learned by the model. The distribution families cluster, with the mean and variance mapped to orthogonal directions.

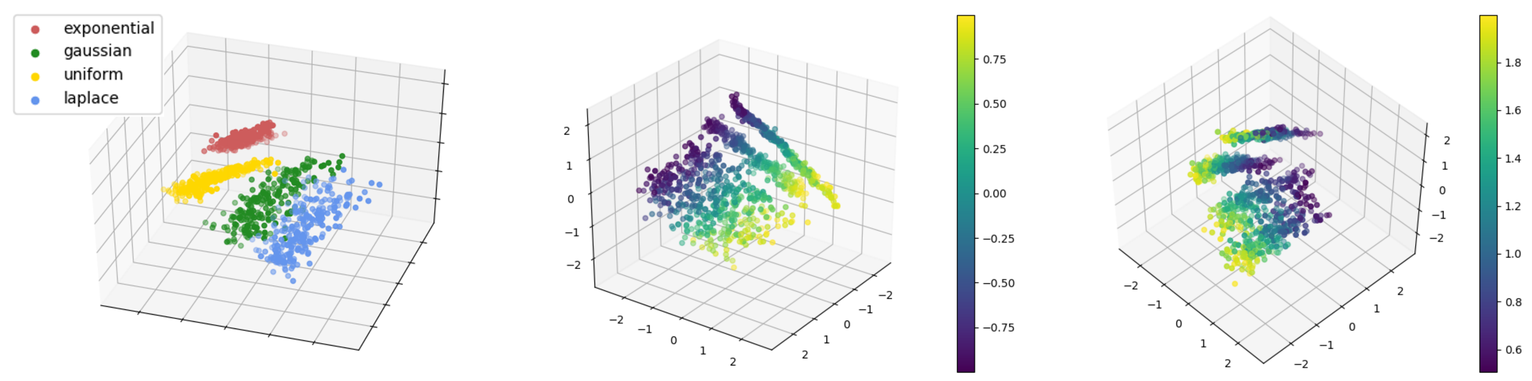

Figure: Mean of $q(c|D;\phi)$, coloured by distribution (left) type, (center) mean, (right) variance.

Spatial MNIST 2-D Experiments

Aim: Model complex datasets and identify representative samples.

Spatial MNIST is obtained by sampling 50 coordinate values from a probability density specified by the pixel intensity of MNIST digits [3]. We are able to sample new datasets conditioned on a set of inputs, and also summarise sensible datasets by choosing a subset $S \subseteq D$ that minimises $KL\left(q(c|D;\phi)||q(c|S;\phi)\right)$.

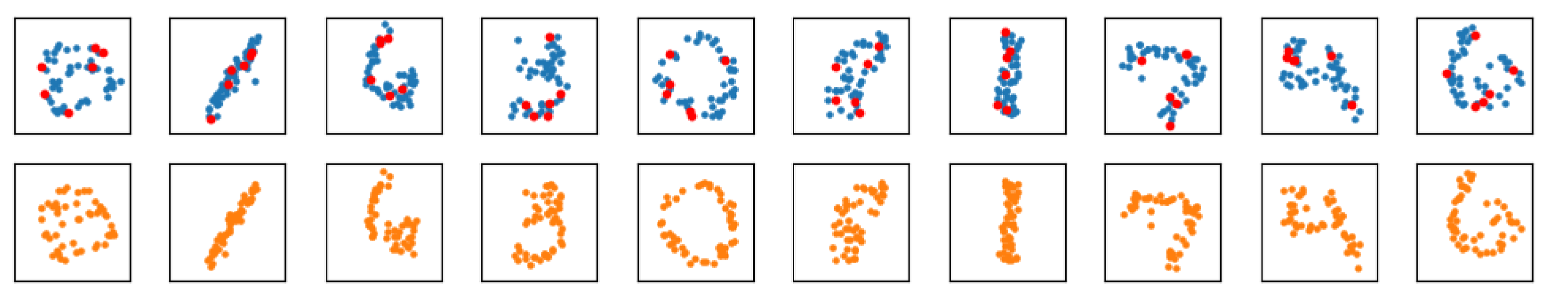
Figure: Blue and red dots are the input digits as well as 6-sample summaries. Orange digits are the conditioned samples from spatial MNIST data.

YouTube Faces

Aim: Specify complex distributions and generate conditioned / new samples.

We train the model on cropped and resized images from the YouTube Faces Database [6] to generate new frames conditioned on input faces and show reasonable similarity. We also generate new samples with a consistent identity.


Figure: (left) Inputs, (center) faces conditioned on input and (right) generated from sampled c.

OMNIGLOT and Few-shot Learning

Aim: Transfer generative model to new datasets and classify unseen classes.

We demonstrate few-shot learning capabilities by training on OMNIGLOT and generating samples conditioned on *unseen* OMNIGLOT characters or MNIST digits. We also test k-shot classification of unseen examples x by minimising $KL(q(c|D_i;\phi)||q(c|x;\phi))$, with k labelled examples of each class D_i.

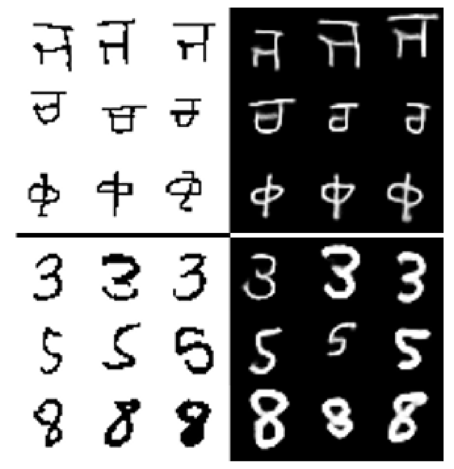
Figure: Few-shot learning from OMNIGLOT to unseen class / MNIST. (left) Inputs, (right) conditioned samples.

Task			Method	
Test Dataset	K Shot	K way	Paper	Ours
MNIST	1	10	78.6	70.2
MNIST	5	10	93.2	87.6
OMNIGLOT	1	5	98.1	95.7
OMNIGLOT	5	5	99.5	98.5
OMNIGLOT	1	20	93.2	85.6
OMNIGLOT	5	20	98.1	95.5

Table: Few-shot learning classification accuracies.

Extension: Emotion-specified Expression

Aim: Generate samples conditioned on label.

We change the proposed graphical model by introducing observed variable y [5]. The context prior is now conditioned on a dataset label y_D, i.e. $p(c) \rightarrow p(c|y_D)$ and $D = \{x_i, \ldots, x_m, y_D\}$. We train the model on the CK+ emotions database [2, 4]. Sample images generated from a context prior given emotion labels are consistent with the desired emotion.

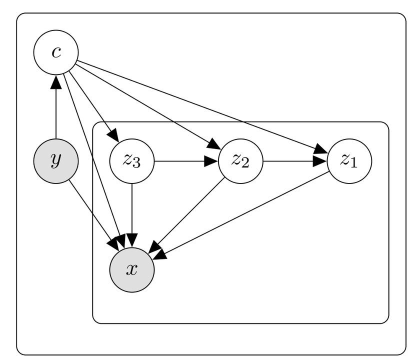
Figure: Extended model using labels for training and sampling.

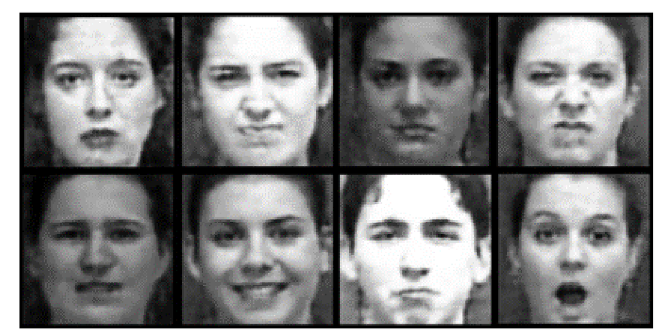
Figure: Sample faces conditioned on emotion label. Top-left to bottom-right: neutral, anger, contempt, disgust, fear, happiness, sadness, and surprise.

Conclusion

The Neural Statistician is a highly flexible generative model that can be used to learn representations of datasets, with applications in a wide variety of tasks. The model is:

+ Unsupervised, data efficient, parameter efficient, capable of few-shot learning, processes datasets of variable length.
– Dataset hungry, limited to datasets of relatively small size during training.

References

[1] Edwards, H. and Storkey, A., 2016. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*.
[2] Kanade, T., Cohn, J. F., Tian, Y., 2000. Comprehensive database for facial expression analysis. *Proceedings of FG'00*, pp.46-53.
[3] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), pp.2278-2324.
[4] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. *Proceedings of CVPR4HB 2010*, pp.94-101.
[5] Sohn, K., Lee, H., Xinchen, Y., 2015. Learning Structured Output Representation using Deep Conditional Generative Models.
[6] Wolf, L. and Hassner, T. and Maoz, I., 2011. Face Recognition in Unconstrained Videos with Matched Background Similarity. *IEEE CVPR*.