Towards a Neural Statistician

R. Comanescu, Y.C. Leung, T. Martin

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Motivation

We present an extension of a variational autoencoder (VAE), that learns to compute statistics of a dataset. The key idea is that we work with **datasets** rather than **datapoints**, by introducing a **context variable c** constant for items in the same dataset [Edwards and Storkey, 2017].

Vanilla VAE

VAE is a latent variable model $p(x|z;\theta)$ (the **decoder**) with parameters θ. Each observed x, can be decoded by its corresponding latent variable z as follows:

$$p(x) = \int p(x|z;\theta)p(z)dz. \tag{1}$$

To approximate the posterior distribution of z, i.e. $p(z)$, a recognition network (the **encoder**) $q(z|x;\phi)$ with parameters ϕ is introduced to set the standard variational lower bound on the single-datum log-likelihood, i.e. $log(p(x|\theta)) \geq \mathcal{L}_x$, where

$$\mathcal{L}_x = \mathbb{E}_{q(z|x;\phi)}[log\ p(x|z;\theta)] - D_{KL}(q(z|x;\phi||p(z)). \tag{2}$$

Neural Statistician

To extend the VAE model, we introduce a new latent variable **c**, the context, which varies between datasets but is constant for items from the same dataset. The likelihood of one dataset D with parameter θ is given by:

$$p(D) = \int p(c)\left[\prod_{x\in D}\int p(x|z;\theta)p(z|c;\theta)dz\right]dc. \tag{3}$$

The variational lower bound on the dataset can then be expressed as follows:

$$\mathcal{L}_D = \mathbb{E}_{q(c|D;\phi)}\left[\sum_{x\in d}\mathbb{E}_{q(z|c,x;\phi)}[\log p(x|z;\theta)] - D_{KL}(q(z|c,x;\phi)\|p(z|c;\theta))\right] - D_{KL}(q(c|D;\phi)\|p(c)). \tag{4}$$

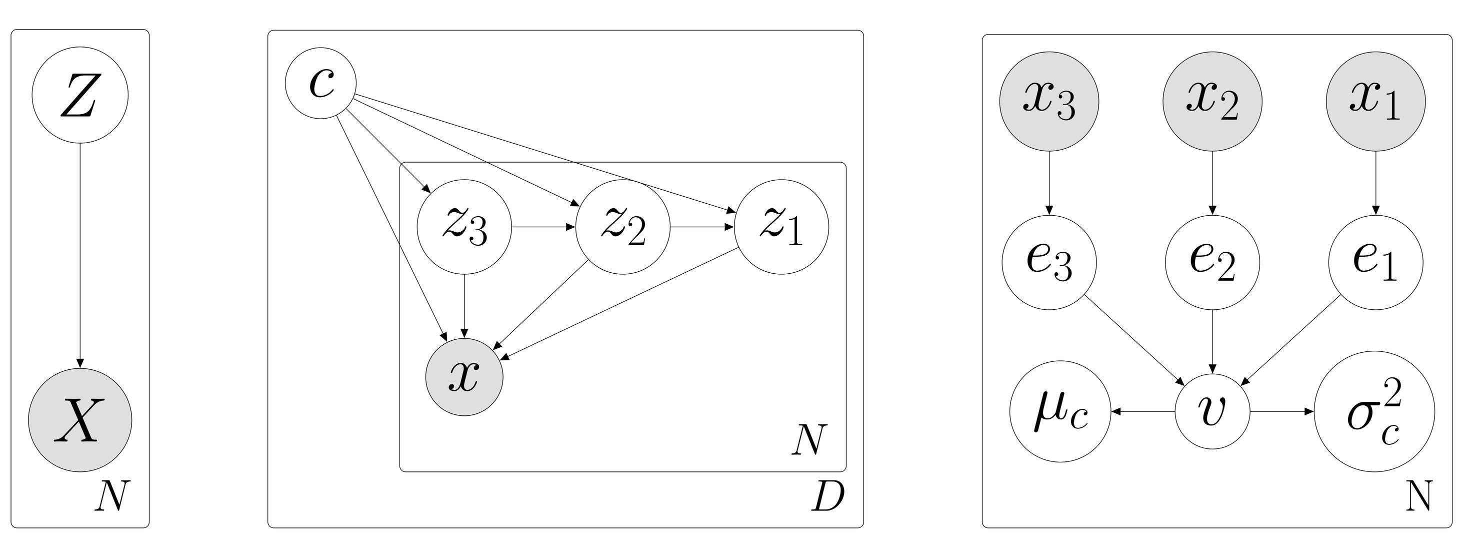

Figure: *Left*: Vanilla VAE. *Middle*: Neural statisctician with 3 layers. *Right*: Statistic network, combines datapoints in a dataset.

Multiple stochastic layers $z_1, ..., z_k$ and skip connections are introduced:

$$p(D) = \int p(c)\prod_{x\in D}\int p(x|c,z_{1:L};\theta)p(z_L|c;\theta)\prod_{i=1}^{L-1}p(z_i|z_{i+1},c;\theta)\,dz_{1:L}dc. \tag{5}$$

where $p(x|c, z_{1:L})$ is the observation decoder and $p(z_i|z_{i+1}, c)$ is the latent decoder. The full approximate posterior factorises analogously as,

$$q(c, z_{1:L}|D;\phi) = q(c|D;\phi)\prod_{x\in D}q(z_L|x,c;\phi)\prod_{i=1}^{L-1}q(z_i|z_{i+1},x,c;\phi), \tag{6}$$

where $q(c|D)$ the statistic network and $q(z_i|z_{i+1}, x, c)$ the inference network.

Experiments

Given input $x_1 \dots x_k$ use the statistics network to calculate the approximate posterior $q(c|x_1 \dots x_k;\phi)$. Set c to the mean of the approximate posterior and sample from $p(x|c,\theta)$.

Summarizing datasets: SPATIAL MNIST

The Spatial MNIST dataset is created by sampling coordinate values from each original MNIST image based on pixel intensity. This generates a set of 50 (x, y) coordinates for each image (black digits in top row).

The dataset $\mathcal{D}$ can be summarized into $\mathcal{S} \subseteq \mathcal{D}$, by minimizing $D_{KL}(q(c|\mathcal{D})||q(c|\mathcal{S}))$. In the bottom row, the dataset is summarized into 6 samples (red dots).

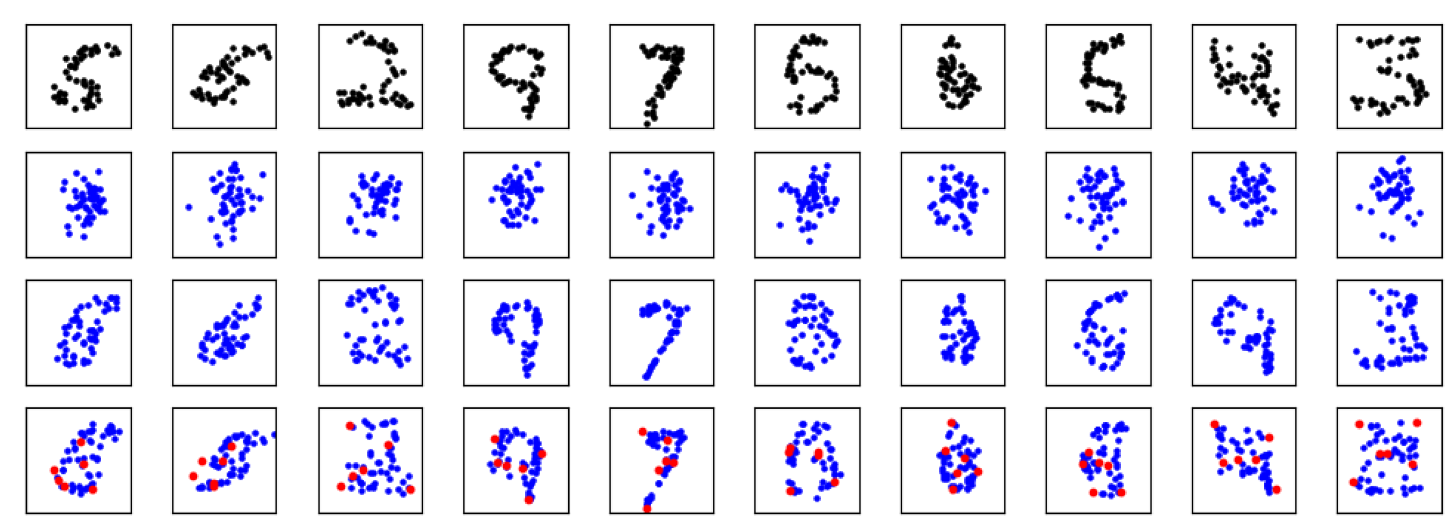

Figure: Spatial MNIST. First row: input. Bottom row: conditioned sampled. Red points correspond to a 6-sample summary.

Generating new samples: YOUTUBE FACES

The YouTube Faces Database (Wolf et. al [2011]) contains 3245 videos of 1595 different people, which have been cropped to contain only faces and resized to 64×64 pixels.

Figure: YouTube Faces Dataset.

Using these data, we trained the network model with a single stochastic layer with 512 dimensional latent c and 16 dimensional z variable to encode information into the latent variable spaces. The decoder produced the following artificial faces.

(a) (b) (c) (d) (e)

Figure: The evolution of artificial faces. (a) is the input image. (b), (c), (d) and (e) are the output images after 1, 2, 50 and 150 epochs.

Few-shot learning: OMNIGLOT

We performed few-shot classification of unseen OMNIGLOT(1628 classes of handwritten characters) and MNIST(10 digits) characters, after training on OMNIGLOT. We classify input image x as class i using $\text{argmin}_i\, D_{KL}(q(c|D_i)||q(c|x;\phi))$.

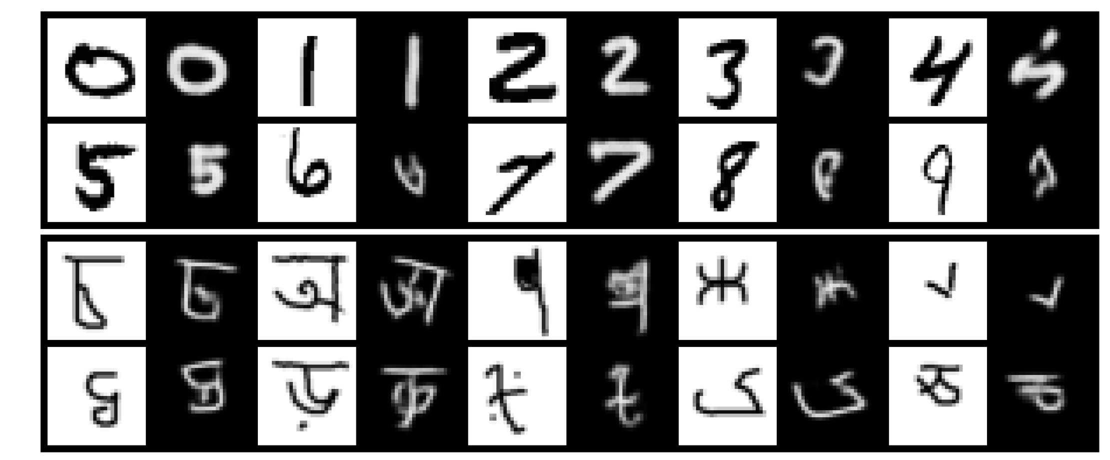

Figure: 1-shot learning on MNIST (upper) and OMNIGLOT(bottom). White is unseen input, black is generated.

	Task		Method		
Test Dataset	**K Shot**	**K Way**	**Siamese**	**NeuStat**	**Ours**
MNIST	1	10	70	78.6	71.0
MNIST	5	10	-	93.2	88.2
OMNIGLOT	1	5	97.3	98.1	95.9
OMNIGLOT	5	5	98.4	99.5	98.2
OMNIGLOT	1	20	88.1	93.2	85.4
OMNIGLOT	5	20	97.0	98.1	94.0

Table: Classification accuracy of various few-shot learning tasks averaged over 100 runs. Training is done with 20 inputs from each class.

Extension: Unsupervised Sentence Embeddings

We adapt the model to learn sentence representations, by representing each sentence as a dataset of word embeddings. We learn unsupervised sentence embddings by training a Neural Statistician on 2 million Wikipedia sentences and we test the sentence embeddings on a sentence similarity task (SentEval), defining similarity as the divergence between posteriors given sentences s_1 and s_2 as $D_{KL}(q(c|s_2)||q(c|s_1))$.

Preliminary results: Using 300-dimenstional GloVe embeddings, we obtain **0.28** Pearson-correlation on STS 2016, compared to **0.51** for Skip-thought [Kiros et. al, 2015].

References

Harrison Edwards and Amos Storkey. Towards a neural statistician. In *ICLR*, 2017.

Wolf et. al. Face recognition in unconstrained videos with matched background similarity. IEEE Computer Society, 2011.

Kiros et. al. Skip-thought vectors. In *NIPS*, 2015.