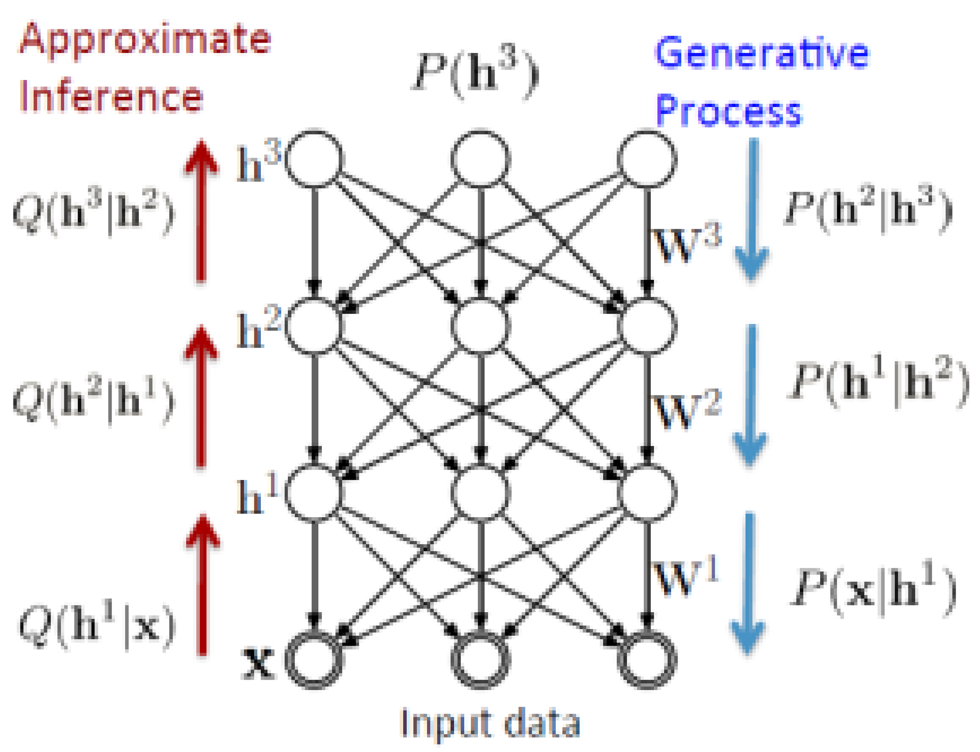
Importance Weighted Autoencoders

J. Rampersad, C.Tegho & S. Pascual Diaz

UNIVERSITY OF CAMBRIDGE

The Inference Problem

Variational Auto-Encoders (VAEs) can efficiently perform approximate inference and learning in deep directed probabilistic models even in the presence of continuous latent variables with intractable posterior distributions. VAEs rely on the optimisation of a lower-bound of the log-likelihood, the proximity of which to the true likelihood strongly determines the richness of posterior distributions that can be approximated. We present the **Importance Weighted Auto-Encoder** - this model results in significant improvements on density modelling benchmarks by optimising a strictly tighter lower bound on the log likelihood than the VAE .



Variational Auto-Encoder

The **VAE** fundamentally consists of two components, a *recognition* model over the latent variables **z**: $P_\theta(\mathbf{z}|\mathbf{x})$ and a *generative* model over the observed data $P_\theta(\mathbf{x}|\mathbf{z})$. Typically the posterior distribution is intractable, but we would like to learn and infer θ and **z** respectively so that we can perform data generation, or marginal likelihood estimation. The **VAE** introduces an approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ parametrised by a new set of parameters ϕ, these parameters are varied over the course of learning to drive up the likelihood.

Neural networks are often used in the parametrisation of both generative and recognition distributions. In the case of the recognition model, hidden layers of the neural network can be factorised parameter updates back-propagated in an analogous fashion to standard neural networks.

$$q_\phi(\mathbf{h}|\mathbf{x}) = q_\phi(\mathbf{h}^1|\mathbf{x})q_\phi(\mathbf{h}^1|\mathbf{h}^2)...q_\phi(\mathbf{h}^L|\mathbf{h}^{L-1}) \quad (1)$$

Similarly, in the generative distribution where $h = h^1...h^L$ denotes the stochastic hidden units:

$$p(x|\theta) = \sum_{h_1,...,h_L} p(h^L|\theta)p(h^{L-1}|h_L,\theta)...p(x|h^1,\theta) \quad (2)$$

The VAE is trained to maximize a variational lower bound on log$p(x)$ according to the variational posterior $q(h|x)$.

$$\mathcal{L}(x) = E_{q(h|x)}[\log\frac{p(x,h)}{q(h|x)}] \quad (3)$$

The lower bound can be approximated through sampling but is analytically intractable due to dependence of $p(x,h)$ on the intractable posterior. Sampling directly from the variational posterior and averaging returns leads to highly varied estimates of $\mathcal{L}(x)$.

VAEs use a re-parametrisation trick to compute latent variables **h** determinitically with the help of auxilliary variables ϵ^l independently sampled from $\mathcal{N}(0,I)$. Assume $q(h^l|h^{l-1},\phi) = \mathcal{N}(h^l|\mu(h^{l-1},\phi),\sigma(h^{l-1},\phi))$, we can write:

$$h^l(\epsilon^l,h^{l-1},\phi) = \mu(h^{l-1},\phi) + \sigma(h^{l-1},\phi)^{0.5}\epsilon^l \quad (4)$$

Using (1), latent variables h can be expressed as $h(x,\phi,\epsilon)$. This parametrisation allows a Monte-Carlo expectation to be written w.r.t $q_\phi(h|x)$ such that it is differentiable by ϕ.

Importance Weighted Auto-Encoder

IWAE uses the same architecture as **VAE** but optimises a tighter bound on $logp(x)$ corresponding to the k-sample importance weighting estimate of the log-likelihood [1]:

$$\mathcal{L}_k(x) = E_{h_1,...,h_k\sim q(h|x)}[\log\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,h_i)}{q(h_i|x)}] \quad (5)$$

Training Procedure: Gradients of the lower-bound $L_{\theta,\phi}$ w.r.t θ and ϕ are estimated in both cases and used to update the parameters until convergence. Define $f(x,h_i) = \frac{p(x,h_i)}{q(h_i|x)}$ and $\tilde{w}_i = \frac{f(x,h_i)}{\sum_{i=1}^{k} f(x,h_i)}$

Discrepancies in the gradient estimators:

$$\textbf{VAE:} \quad \frac{1}{k}\sum_{i=1}^{k}\nabla_\theta \log f(x,h(\epsilon_i,x,\theta),\theta) \quad (6)$$

$$\textbf{IWAE:} \quad \sum_{i=1}^{k}\tilde{w}_i\nabla_\theta \log f(x,h(\epsilon_i,x,\theta),\theta) \quad (7)$$

Visualisation of learned manifolds

We show random samples from learned generative models for MNIST, trained with VAE (right) and IWAE (left) with 1 layer and 5 samples.

Results on density estimation

The generative performance of IWAEs improved with increasing k, and increasing number of stochastic layers. Improvements were less significant with VAEs, and IWAEs outperformed VAEs with all models, for both the MNIST and the OMNIGLOT datasets. IWAEs learned more latent dimensions than VAEs.
The table below shows results for the MNIST dataset. (Results marked with a star are taken from the paper directly).

		VAE		IWAE	
# stoch layers	k	NLL	active units	NLL	active units
1	1	86.72	19	86.72	19
	5	86.50	20	85.44	22
	50	86.47	20	84.78	25
2	1	85.72	16+5	85.72	16+5
	5	84.80	17+5	83.89*	21+5
	50	84.85	17+5	82.90*	26+7

Future work

Weighting the KL divergence term by a variable parameter β can dictate the extent to which gradients are driven by pure deterministic reconstruction error and the variational regularisation term given by the KL divergence. We intend to follow the results of [3] by implementing 'warm up'. A technique that gradually increases β from 0 to 1 over the course of training. Previous results with VAEs and Ladder VAEs show this reduces the number of inactive latent cells and improves performance over the regular VAEs.

References

1. Kingma, D. P., and Welling M., "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
2. Yuri B., Roger G. and Ruslan S., "Importance Weighted Autoencoders" arXiv:1509.00519 (2015)
3. Sønderby CK., Raiko T, Maaløe L, et al "Ladder Variational Autoencoders" arXiv:1602.02282 (2016)
4. Krakovna V. Highlights from the Deep Learning Summer School. Available: https://goo.gl/xBdxPu (2016)