IMPROVING SAMPLING FROM GENERATIVE AUTOENCODERS WITH MARKOV CHAINS

Kai Arulkumaran,* Antonia Creswell* & Anil A. Bharath
Department of Bioengineering
Imperial College London
London SW7 2BP, UK
{ka709,ac2211,aab01}@imperial.ac.uk

ABSTRACT

We focus on generative autoencoders, such as variational or adversarial autoencoders, which jointly learn a generative model alongside an inference model. We define generative autoencoders as autoencoders which are trained to softly enforce a prior on the latent distribution learned by the model. However, the model does not necessarily learn to match the prior. We formulate a Markov chain Monte Carlo (MCMC) sampling process, equivalent to iteratively encoding and decoding, which allows us to sample from the learned latent distribution. Using this we can improve the quality of samples drawn from the model, especially when the learned distribution is far from the prior. Using MCMC sampling, we also reveal previously unseen differences between generative autoencoders trained either with or without the denoising criterion.

1 INTRODUCTION

Unsupervised learning has benefited greatly from the introduction of deep generative models. In particular, the introduction of generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Kingma & Welling; Rezende et al., 2014) has led to a plethora of research into learning latent variable models that are capable of generating data from complex distributions, including the space of natural images (Radford et al., 2015). Both of these models, and their extensions, operate by placing a prior distribution over a space Z, and learn mappings from the latent space to the space of the observed data: $Z \mapsto X$.

We are interested in autoencoding generative models, that is, models which learn not just the generative mapping $Z \mapsto X$, but also the inferential mapping $X \mapsto Z$. Specifically, we define *generative autoencoders* as autoencoders which softly constrain their latent distribution, $Q(Z)$, to match a specified prior distribution, $P(Z)$. This includes VAEs, extensions of VAEs (Kingma et al., 2016), and also adversarial autoencoders (AAEs) (Makhzani et al., 2015). Whilst other autoencoders also learn an encoding function, $e : X \to Z$, together with a decoding function, $d : Z \to X$, $Q(Z)$ is not necessarily constrained to conform to a specified probability distribution. This is the key distinction for generative autoencoders; both e and d can still be deterministic functions (Makhzani et al., 2015).

The process of decoding can be interpreted as sampling the conditional probability, $P(X|Z)$, approximated as $Q(X|Z)$, by the parameterised function d. Likewise, encoding can be interpreted as sampling the conditional probability, $P(Z|X)$, approximated as $Q(Z|X)$, by the parameterised function e. The combination of these two functions enables interpolating between two samples $\mathbf{x} \in X$ by encoding each of them, interpolating between their encodings in Z, and generating new samples from these encodings. However, if the two original samples differ greatly, the interpolated samples often fail to be consistent with $P(X)$ (see Figure 1).

The decoder of a generative autoencoder can also be used to be generate novel samples of $\mathbf{x}$. d, representing $Q(X|Z)$, should be sampled conditioned on samples $\mathbf{z} \sim Q(Z)$. However, during training, $Q(Z)$ is constrained to approximate a known prior distribution, $P(Z)$, and so the common practice is to instead to draw $\mathbf{x} \sim Q(X|Z)$, conditioning on $\mathbf{z} \sim P(Z)$. This can produce poor

*Equal contributions

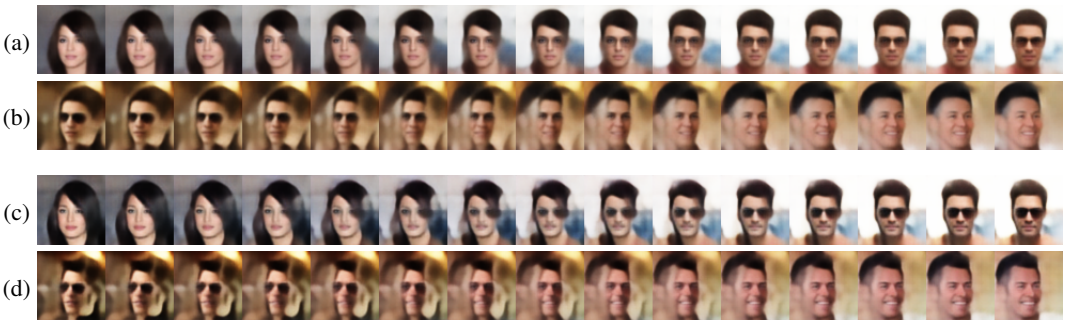


Figure 1: Spherically interpolating (White, 2016) between two faces using (a-b) a VAE and (c-d) an AAE. In (a) and (c), the attempt to gradually generate sunglasses results in visual artefacts around the eyes. In (b) and (d), the models fail to properly capture the desired change in orientation of the face, resulting in three partial faces in the middle of the interpolation.

samples of $\mathbf{x}$ because the $\mathbf{z}$'s are not drawn from the learned distribution, $Q(Z)$, and realistically $Q(Z) \neq P(Z)$. Unfortunately, it is impossible to directly sample $Q(Z)$.

Our main contribution is the formulation of a Markov chain Monte Carlo (MCMC) sampling process for generative autoencoders, which allows us to sample $Q(Z)$. By iteratively sampling the chain, starting from an arbitrary $\mathbf{z}_{t=0} \in \mathbb{R}$, the chain converges to $\mathbf{z}_{t\to\infty} \sim Q(Z)$, allowing us to draw latent samples from $Q(Z)$ after several steps of MCMC sampling. From a practical perspective, this is achieved by iteratively encoding and decoding, which allows it to be trivially applied to existing generative autoencoders. Because $Q(Z)$ is optimised to be close to $P(Z)$, the initial sample, $\mathbf{z}_{t=0}$ can be drawn from $P(Z)$, improving the quality of the samples within a few iterations.

When interpolating between latent encodings, there is no guarantee that $\mathbf{z}$ stays within high density regions of $Q(Z)$. Previously, this has been addressed by spherical, rather than linear interpolation, due to the typically high dimensionality of Z (White, 2016). However, this approach attempts to keep $\mathbf{z}$ within $P(Z)$, which is not the same as $Q(Z)$. If the model can instead sample from $Q(X|Z)$ with $\mathbf{z} \sim Q(Z)$, it is able to correct the unrealistic artefacts observed in Figure 1 (see Figure 3). Whilst most methods that aim to generate realistic samples from X rely on adjusting encodings of the observed data (White, 2016), our use of MCMC allows us to walk any sample to more probable regions on the learned manifold, resulting in more convincing generations. We demonstrate that the use of MCMC sampling improves generations from both VAEs and AAEs with high-dimensional Z; this is important as previous studies have shown that the dimensionality of Z should be scaled with the intrinsic latent dimensionality of the observed data.

Our secondary contribution is the introduction of denoising generative autoencoders, where we add the denoising criterion (Seung, 1997; Vincent et al., 2008) to generative autoencoders, constructing denoising VAEs (DVAEs) and denoising AAEs (DAAEs). Unlike previous work on DVAEs (Im et al., 2015), we simply use the denoising criterion as-is, without deriving a new variational lower bound. We reformulate our original MCMC sampling process to incorporate the noising and denoising processes, allowing us to use MCMC sampling on denoising generative autoencoders. The initial generations and interpolations from these models appear similar to those from their non-denoising counterparts, but by applying MCMC sampling we show that these models do indeed exhibit denoising properties (see Figure 3).

2 Background

One of the main tasks in machine learning is to learn explanatory factors for observed data, commonly known as inference. That is, given a data sample $\mathbf{x} \in X \subseteq \mathbb{R}^a$, we would like to find a corresponding latent encoding $\mathbf{z} \in Z \subseteq \mathbb{R}^b$. Another task is to learn the inverse, generative mapping from a given $\mathbf{z}$ to a corresponding $\mathbf{x}$. In general, coming up with a suitable criterion for learning these mappings is a difficult task. Autoencoders solve both tasks efficiently by jointly learning an inferential mapping e and generative mapping d, using unlabelled data from X in a self-supervised fashion. The basic objective of all autoencoders is to minimise a reconstruction

cost, $\mathcal{L}_{reconstruct}$, between the original data, X, and its reconstruction, $d(e(X))$. Examples of $\mathcal{L}_{reconstruct}$ include the squared error loss, $\frac{1}{2}\sum_{n=1}^{N}\|d(e(\mathbf{x}_n)) - \mathbf{x}_n\|^2$, and the cross-entropy loss, $\mathcal{H}[P(X)\|P(d(e(X)))] = -\sum_{n=1}^{N}\mathbf{x}_n \log(d(e(\mathbf{x}_n))) + (1-\mathbf{x}_n)\log(1-d(e(\mathbf{x}_n)))$.

Autoencoders may be cast into a probablistic framework, by considering samples $\mathbf{x} \sim P(X)$ and $\mathbf{z} \sim P(Z)$, and attempting to learn the conditional distributions $P(Z|X)$ and $P(X|Z)$ as e and d respectively, with $\mathcal{L}_{reconstruct}$ representing the negative log-likelihood of the reconstruction given the encoding (Bengio, 2009). With any autoencoder it is possible to create novel $\mathbf{x} \in X$ by passing a $\mathbf{z} \in Z$ through d, but we have no knowledge of appropriate choices of $\mathbf{z}$ beyond those obtained via $e(X)$. One solution is to constrain e such that the learned marginal of the model, $Q(Z)$, corresponds to a probability distribution that we can sample from. This can be achieved by an additional loss, $\mathcal{L}_{prior}$, that penalises encodings far away from a specified prior distribution, $P(Z)$. We now review two types of generative autoencoders, VAES (Kingma & Welling; Rezende et al., 2014) and AAEs (Makhzani et al., 2015), which each take different approaches to formulating $\mathcal{L}_{prior}$.

2.1 GENERATIVE AUTOENCODERS

Consider the case where e is constructed with stochastic neurons that can produce outputs from a specified probability distribution, and $\mathcal{L}_{prior}$ is used to constrain the distribution of outputs to $P(Z)$. This leaves the problem of estimating the gradient of the autoencoder over the expectation $\mathbb{E}_{Q(Z|X)}$, which would typically be addressed with a Monte Carlo method. VAEs sidestep this issue by constructing random samples using a deterministic function and a source of noise, moving the source of stochasticity to an input, and leaving the network itself deterministic for standard gradient calculations—a technique commonly known as the reparameterisation trick (Kingma & Welling). e then consists of a deterministic function, e_{rep}, that outputs parameters for a probability distribution, plus a source of noise. In the case where $P(Z)$ is a diagonal covariance Gaussian, e_{rep} maps $\mathbf{x}$ to a vector of means, $\boldsymbol{\mu} \in \mathbb{R}^b$, and a vector of standard deviations, $\boldsymbol{\sigma} \in \mathbb{R}^b_+$, with the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Put together, the encoder outputs samples $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}$, where $\odot$ is the Hadamard product. VAEs attempt to make these samples from the encoder match up with $P(Z)$ by using the KL divergence between the parameters for a probability distribution outputted by e_{rep}, and the parameters for the prior distribution, giving $\mathcal{L}_{prior} = D_{KL}[Q(Z|X)\|P(Z)]$. A multivariate Gaussian has an analytical KL divergence that can be further simplified when considering the unit Gaussian, resulting in $\mathcal{L}_{prior} = \frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\mu}^2 + \boldsymbol{\sigma}^2 - \log(\boldsymbol{\sigma}^2) - \mathbf{1}$.

Another approach is to forgo stochastic neurons, and deterministically output the encodings $\mathbf{z}$. Rather than minimising a metric between probability distributions using their parameters, we can turn this into a density ratio estimation problem where the goal is to learn a parametric $Q(Z|X)$ such that the probability of a sample $\mathbf{z}$ from the encoder is the same as the probability of $\mathbf{z}$ under the prior: $\frac{Q(\mathbf{z}|\mathbf{x})}{P(\mathbf{z})} = 1$. The GAN framework solves the density ratio estimation problem by transforming it into a class estimation problem (Goodfellow et al., 2014). The first network in GAN training is the discriminator network, D_ϕ, which is trained to maximise the log probability of samples from the "real" distribution, $\mathbf{z} \sim P(Z)$, and minimise the log probability of samples from the "fake" distribution, $\mathbf{z} \sim Q(Z|X)$. In our case e plays the role of the generator network, G_θ, which generates the "fake" samples.[1] The two networks compete in a minimax game, where G_θ receives gradients from D_ϕ such that it learns to better fool D_ϕ. The training objective for both networks is therefore given by $\mathcal{L}_{prior} = \min_\theta \max_\phi \mathbb{E}_{P(Z)}[\log(D_\phi(\mathbf{z}))] + \mathbb{E}_{Q(Z|X)}[\log(1 - D_\phi(G_\theta(\mathbf{x})))]$. This formulation can create problems during training, so instead G_θ is trained to minimise $-\log(D_\phi(G_\theta(x)))$, which provides the same fixed point of the dynamics of G_θ and D_ϕ. The result of applying the GAN framework to the encoder of an autoencoder is the deterministic AAE (Makhzani et al., 2015). AAEs can be further extended to incorporate probabilistic posteriors, which we do not discuss here.

2.2 DENOISING AUTOENCODERS

In a more general viewpoint, generative autoencoders fulfill the purpose of learning useful representations of the observed data. Another widely used class of autoencoders that achieve this are denoising autoencoders (DAEs), which are motivated by the idea that learned features should be

[1] We adapt the variables to better fit the conventions used in the context of autoencoders.

robust to "partial destruction of the input" (Vincent et al., 2008). Not only does this require encoding the inputs, but capturing the statistical dependencies between the inputs so that corrupted data can be recovered (see Figure 2). DAEs are presented with a corrupted version of the input, $\tilde{\mathbf{x}}$, but must still reconstruct the original input, $\mathbf{x}$, where the noisy inputs are created through sampling $\tilde{\mathbf{x}} \sim Q(\tilde{\mathbf{x}}|\mathbf{x})$. The denoising criterion, $\mathcal{L}_{denoise}$, can be applied to any type of autoencoder by replacing the straightforward reconstruction criterion, $\mathcal{L}_{reconstruct}(X, d(e(X))$, with the reconstruction criterion applied to noisy inputs: $\mathcal{L}_{reconstruct}(X, d(e(\tilde{X}))$. As such we can construct *denoising generative autoencoders* by training autoencoders to minimise $\mathcal{L}_{denoise} + \mathcal{L}_{prior}$.

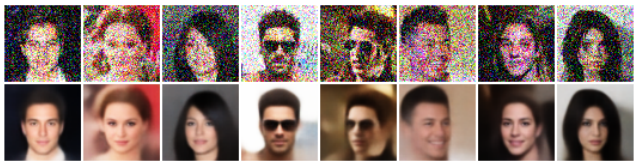


Figure 2: Reconstructions of faces from a DVAE trained with additive Gaussian noise: $Q(\tilde{X}|X) = \mathcal{N}(X, 0.25\mathbf{I})$. The model successfully recovers much of the detail from the noise-corrupted images.

One might expect to see differences in samples drawn from denoising generative autoencoders and their non-denoising counterparts. However, Figures 3 and 4 show that this is not the case. Im et al. (2015) address the case of DVAEs, claiming that the noise mapping requires adjusting the original VAE objective function. Our work is orthogonal to theirs, and others which adjust the training or model (Kingma et al., 2016), as we focus purely on sampling from generative autoencoders after training. We claim that the existing practice of drawing samples from generative autoencoders conditioned on $\mathbf{z} \sim P(Z)$ is suboptimal, and the quality of samples can be improved by instead conditioning on $\mathbf{z} \sim Q(Z)$ via MCMC sampling.

3 Markov sampling

We now consider the case of sampling from generative autoencoders, where $d = Q(X|Z)$. To generate new samples from X via $Q(X|Z)$, we need to know the distribution $Q(Z)$. In practice we draw samples from $P(Z)$ to sample $Q(X|Z)$, since $Q(Z)$ approaches $P(Z)$ through the learning process. However, we now show that for any initial $\mathbf{z}_0 \in Z_0 = \mathbb{R}^b$, Markov sampling can be used produce a chain of samples $\mathbf{z}_t$, which as $t \to \infty$, produces samples $\mathbf{z}_t$ that are from the distribution $Q(Z)$, thereby allowing us to draw samples from $Q(X|Z)$, conditioned on $\mathbf{z} \sim Q(Z)$. To speed up convergence we can initialise $\mathbf{z}_0$ from a distribution close to $Q(Z)$, by drawing $\mathbf{z}_0 \sim P(Z)$.

3.1 Markov sampling process

A generative autoencoder can be sampled by the following process:

$$\mathbf{z}_0 \in Z_0 = \mathbb{R}^b$$

$$\mathbf{x}_{t+1} \sim Q(X|Z_t)$$

$$\mathbf{z}_{t+1} \sim Q(Z|X_{t+1})$$

This allows us to define a Markov chain with the transition operator

$$T(Z_{t+1}|Z_t) = \int Q(Z_{t+1}|X)Q(X|Z_t)dX \tag{1}$$

for $t \geq 0$.

Drawing samples according to the transition operator $T(Z_{t+1}|Z_t)$ produces a Markov chain. For the transition operator to be homogeneous, the parameters of the parameterised distributions $Q(X|Z)$ and $Q(Z|X)$ are fixed during sampling.

3.2 CONVERGENCE PROPERTIES

We now show that the stationary distribution of sampling from the Markov chain is $Q(Z)$.

Theorem 1. *If $T(Z_{t+1}|Z_t)$ defines an ergodic Markov chain, $\{Z_1, Z_2 ... Z_t\}$, then the chain will converge to a stationary distribution $\Pi(Z)$ from any arbitrary initial distribution. The stationary distribution $\Pi(Z) = Q(Z)$.*

The proof of Theorem 1 can be found in (Rosenthal, 2001).

Lemma 1. *$T(Z_{t+1}|Z_t)$ defines an ergodic Markov chain.*

For a Markov chain to be ergodic it must be both irreducible (it is possible to get from any state to any other state in a finite number of steps) and aperiodic (it is possible to get from any state to any other state without having to pass through a cycle). To satisfy these requirements it is more than sufficient to show that $T(Z_{t+1}|Z_t) > 0$, since every $\mathbf{z} \in Z$ would be reachable from every other $\mathbf{z} \in Z$. We show that $Q(X|Z) > 0$ and $Q(Z|X) > 0$, giving $T(Z_{t+1}|Z_t) > 0$, providing the proof of this in Section A of the supplementary material.

Lemma 2. *The stationary distribution of the chain defined by $T(Z_{t+1}|Z_t)$ is $\Pi(Z) = Q(Z)$.*

For the transition operator defined in Equation 1, the asymptotic distribution to which $T(Z_{t+1}|Z_t)$ converges to is $Q(Z)$, because $Q(Z)$ is the marginal of the joint distribution, $Q(X, Z)$, defined by the sampling process. Using Lemmas 1 and 2 with Theorem 1, we can say that the Markov chain defined by the transition operator in Equation 1 will produce a Markov chain that converges to the stationary distribution $\Pi(Z) = Q(Z)$.

3.3 EXTENSION TO DENOISING GENERATIVE AUTOENCODERS

A denoising generative autoencoder can be sampled by the following process:

$$\mathbf{z}_0 \in Z_0 = \mathbb{R}^b$$
$$\mathbf{x}_{t+1} \sim Q(X|Z_t)$$
$$\tilde{\mathbf{x}}_{t+1} \sim Q(\tilde{X}|X_{t+1})$$
$$\mathbf{z}_{t+1} \sim Q(Z|\tilde{X}_{t+1})$$

This allows us to define a Markov chain with the transition operator

$$T(Z_{t+1}|Z_t) = \int Q(Z_{t+1}|\tilde{X})Q(\tilde{X}|X)Q(X|Z_t)dXd\tilde{X} \tag{2}$$

for $t \geq 0$.

The same arguments for the proof of convergence of Equation 1 can be applied to Equation 2.

3.4 RELATED WORK

Our work is similar to several approaches proposed by Bengio et al. (2013; 2014). Bengio et al. define the transition operator in terms of X_t and X_{t-1}, and initially generate samples with an X_0 from the observed data. However, we define the transition operator in terms of Z_{t+1} and Z_t, initialise samples with a Z_0 that is drawn from a prior distribution we can directly sample from, and then sample X_1 conditioned on Z_0. Although the initial samples may be poor, we are likely to generate a novel X_1 on the first step of MCMC sampling, which would not be achieved using Bengio et al.'s approach. We are able to use this approach because we constrain $Q(Z)$ to be close to a prior distribution $P(Z)$.

3.5 EFFECT OF REGULARISATION METHOD

The choice of $\mathcal{L}_{prior}$ may effect how much improvement can be gained when using MCMC sampling, assuming that the optimisation process converges to a reasonable solution. We first consider the case of VAEs, which minimise $D_{KL}[Q(Z|X)\|P(Z)]$. Minimising this KL divergence penalises

the model $Q(Z)$ if it contains samples that are outside the support of the true distribution $P(Z)$, which might mean that $Q(Z)$ captures only a part of $P(Z)$. This means that when sampling $P(Z)$, we may sample a region that is not captured by $Q(Z)$. This suggests that MCMC sampling can improve samples from VAEs by walking them towards denser regions in $Q(Z)$.

The reverse KL divergence, $D_{KL}[P(Z)\|Q(Z|X)]$, penalises the model $Q(Z)$ if $P(Z)$ produces samples that are outside of the support of $Q(Z)$. By minimising this KL divergence, most samples in $P(Z)$ will likely be in $Q(Z)$ as well. AAEs are regularised using the JS entropy, given by $\frac{1}{2}D_{KL}[P(Z)\|\frac{1}{2}(P(Z)+Q(Z|X))]+\frac{1}{2}D_{KL}[Q(Z|X)\|\frac{1}{2}(P(Z)+Q(Z|X))]$. Minimising this cost function attempts to find a compromise between the aforementioned extremes. However, this still suggests that some samples from $P(Z)$ may lie outside $Q(Z)$, and so we expect AAEs to also benefit from MCMC sampling.

4 EXPERIMENTS

4.1 MODELS

We utilise the deep convolutional GAN (DCGAN) (Radford et al., 2015) as a basis for our autoencoder models. Although the recommendations from Radford et al. (2015) are for standard GAN architectures, we adopt them as sensible defaults for an autoencoder, with our encoder mimicking the DCGAN's discriminator, and our decoder mimicking the generator. The encoder uses strided convolutions rather than max-pooling, and the decoder uses fractionally-strided convolutions rather than a fixed upsampling. Each convolutional layer is succeeded by spatial batch normalization and ReLU nonlinearities, except for the top of the decoder which utilises a sigmoid function to constrain the output values between 0 and 1. We minimise the cross-entropy between the original and reconstructed images. Although this results in blurry images in regions which are ambiguous, such as hair detail, we opt not to use extra loss functions that improve the visual quality of generations (Larsen et al., 2015; Dosovitskiy & Brox, 2016) to avoid confounding our results.

Although the AAE is capable of approximating complex probabilistic posteriors (Makhzani et al., 2015), we construct ours to output a deterministic $Q(Z|X)$. As such, the final layer of the encoder part of our DAAE is a convolutional layer that deterministically outputs a latent sample $\mathbf{z}$. The adversary is a fully-connected network with dropout and leaky ReLU nonlinearities. e_{rep} of our DVAE has an output of twice the size, which corresponds to the means, $\boldsymbol{\mu}$, and standard deviations, $\boldsymbol{\sigma}$, of a diagonal covariance Gaussian distribution. For all models our prior, $P(Z)$, is a 200D isotropic Gaussian with zero mean and unit variance: $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

4.2 DATASETS

Our primary dataset is the (aligned and cropped) CelebA dataset, which consists of 200,000 images of celebrities (Liu et al., 2015). The DCGAN (Radford et al., 2015) was the first generative neural network model to show convincing novel samples from this dataset, and it has been used ever since as a qualitative benchmark due to the amount and quality of samples. In Figures 7 and 8 of the supplementary material we also include results on the SVHN dataset, which consists of 100,000 images of house numbers extracted from Google Street view images (Netzer et al., 2011).

4.3 TRAINING & EVALUATION

For all datasets we perform the same preprocessing: cropping the centre to create a square image, then resizing to 64×64px. We train our generative autoencoders for 20 epochs on the training split of the datasets, using Adam (Kingma & Ba, 2014) with $\alpha = 0.0002$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The denoising generative autoencoders use the additive Gaussian noise mapping $Q(\tilde{X}|X) = \mathcal{N}(X, 0.25\mathbf{I})$. All of our experiments were run using the Torch library (Collobert et al., 2011).[2]

For evaluation, we generate novel samples from the decoder using $\mathbf{z}$ initially sampled from $P(Z)$; we also show spherical interpolations (White, 2016) between four images of the testing split, as depicted in Figure 1. We then perform several steps of MCMC sampling on the novel samples and

[2]Example code is available at `https://github.com/Kaixhin/Autoencoders`.

interpolations. We compare results between VAEs and DVAEs below, and leave results from AAEs and DAAEs to Figures 5 and 6 of the supplementary material.

4.4 INTERPOLATIONS

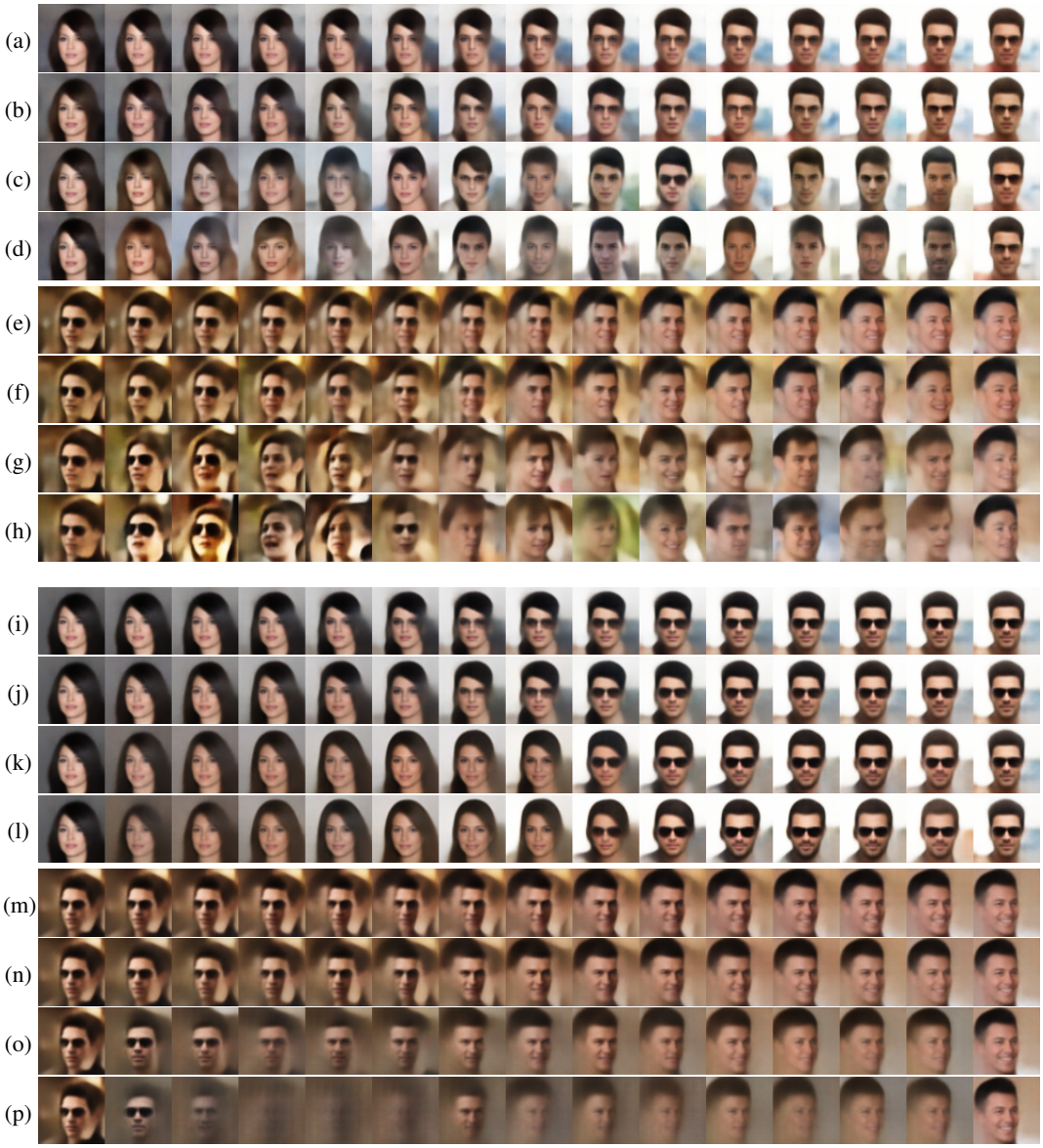
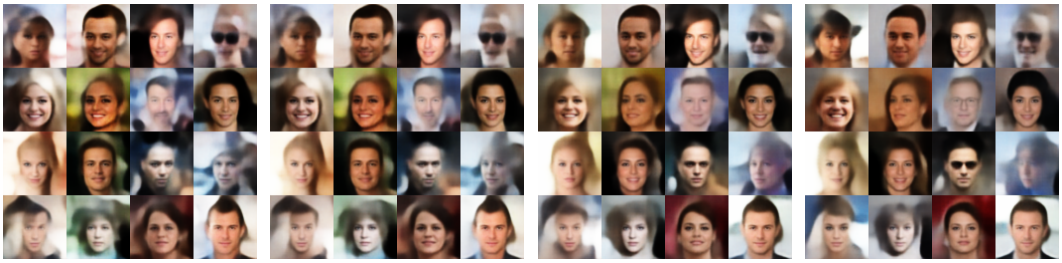
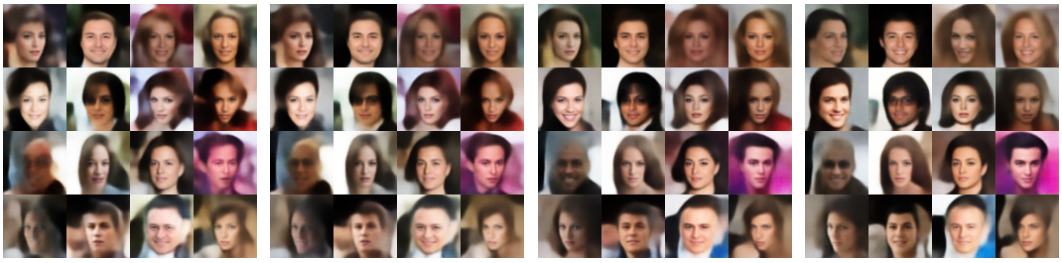


Figure 3: Interpolating between two faces using (a-h) a VAE and (i-p) a DVAE. The top row for each face is the original interpolation, whilst the second, third and fourth rows are the result of 1, 5 and 10 steps of MCMC sampling respectively. In (a-d) and (i-l), the discolouration around the eyes disappears, with the models settling on either generating or not generating sunglasses. In (e-h) the VAE moves away from multiple faces in the interpolation by producing new faces with appropriate orientations. The DVAE, having been trained to denoise, instead blurs away the improbable regions in (m-p). (p) demonstrates an effect that occasionally occurs in denoising models—samples with low contrast occasionally revert to a "mean face".

4.5 SAMPLES



(a) VAE (initial) (b) VAE (1 step) (c) VAE (5 steps) (d) VAE (10 steps)



(e) DVAE (initial) (f) DVAE (1 step) (g) DVAE (5 steps) (h) DVAE (10 steps)

Figure 4: Samples from a VAE (a-d) and DVAE (e-h) trained on the CelebA dataset. (a) and (e) show initial samples conditioned on $\mathbf{z} \sim P(Z)$, which mainly result in recognisable faces emerging from noisy backgrounds. After 1 step of MCMC sampling, the more unrealistic generations change noticeably, and continue to do so with further steps. On the other hand, realistic generations, i.e. samples from a region with high probability, do not change as much.

5 DISCUSSION

In this paper we derive a MCMC sampling process that allows us to directly draw samples from $Q(Z)$, the latent distribution learned by generative autoencoders, where $Q(Z)$ approximates a specified prior distribution, $P(Z)$. This allows us to improve samples $\mathbf{x} \sim Q(X|Z)$, as MCMC sampling allows us to condition these on $\mathbf{z} \sim Q(Z)$. The process is simple, where each step requires iteratively encoding and decoding. Furthermore, as $Q(Z)$ is constrained to be close to $P(Z)$, the initial sample $\mathbf{z}_0$ can be drawn from $P(Z)$ to quickly reach regions of high probability under $Q(Z)$. We show that not only do initially poor samples improve, but unusual artefacts from performing interpolations across the latent space can be corrected through the use of further sampling. We further validate our work by showing that the denoising properties of denoising generative autoencoders are best revealed the use of MCMC sampling.

Our MCMC sampling process is straightforward, and can be applied easily to existing generative autoencoders. This technique is orthogonal to the use of more powerful posteriors in AAEs (Makhzani et al., 2015) and VAEs (Kingma et al., 2016), and the combination of both could result in further improvements in generative modelling. Finally, our basic MCMC proces opens the doors to apply a large existing body of research on sampling methods to generative autoencoders.

REFERENCES

Yoshua Bengio. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1): 1–127, 2009.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pp. 899–907, 2013.

Yoshua Bengio, Eric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *Journal of Machine Learning Research: Proceedings of the 31st International Conference on Machine Learning*, volume 32, 2014.

Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A *matlab*-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv:1602.02644*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. *arXiv preprint arXiv:1511.06406*, 2015.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR-2015), arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/pdf/1412.6980v8.pdf.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR-2015), arXiv preprint arXiv:1312.6114*. URL https://arxiv.org/abs/1312.6114.

Diederik P Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning, arXiv preprint arXiv:1512.09300*, pp. 1558–1566, 2015. URL http://jmlr.org/proceedings/papers/v48/larsen16.pdf.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/37648.pdf.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR) 2016, arXiv preprint arXiv:1511.06434*, 2015. URL https://arxiv.org/pdf/1511.06434.pdf.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning, arXiv preprint arXiv:1401.4082*, 2014. URL https://arxiv.org/pdf/1401.4082.pdf.

Jeffrey S Rosenthal. A review of asymptotic convergence for general state space markov chains. *Far East J. Theor. Stat*, 5(1):37–50, 2001.

H Sebastian Seung. Learning continuous attractors in recurrent networks. In *NIPS Proceedings*, volume 97, pp. 654–660, 1997.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103. ACM, 2008.

Tom White. Sampling generative networks: Notes on a few effective techniques. *arXiv preprint arXiv:1609.04468*, 2016.

Supplementary Material

A PROOF THAT $T(Z_{t+1}|Z_t) > 0$

For $Q(X|Z) > 0$ we require that all possible $\mathbf{x} \in X \subseteq \mathbb{R}^a$ may be generated by the network. Assuming that the model $Q(X|Z)$ is trained using a sufficient number of training samples, $\mathbf{x} \in X_{train} = X$, and that the model has infinite capacity to model $X_{train} = X$, then we should be able to draw any sample $\mathbf{x} \in X_{train} = X$ from $Q(X|Z)$. In reality $X_{train} \subseteq X$ and it is not possible to have a model with infinite capacity. However, $Q(X|Z)$ is modeled using a deep neural network, which we assume has sufficient capacity to capture the training data well. Further, deep neural networks are able to interpolate between samples in very high dimensional spaces (Radford et al., 2015); we therefore further assume that if we have a large number of training samples (as well as large model capacity), that almost any $\mathbf{x} \in X$ can be drawn from $Q(X|Z)$.

Note that if we wish to generate human faces, we define X_{all} to be the space of all possible faces, with distribution $P(X)$, while X_{train} is the space of faces made up by the training data. Then, practically our model only learns to capture X, with distribution $Q(X)$, where $X_{train} \subseteq X \subseteq X_{all}$, because X additionally contains examples of interpolated versions of $\mathbf{x} \sim X_{train}$.

For $Q(Z|X) > 0$ it must be possible to generate all possible $\mathbf{z} \in Z \subseteq \mathbb{R}^b$. $Q(X|Z)$ is described by the function $e : X \to Z$. To ensure that $Q(Z|X) > 0$, we want to show that the function e allows us to represent all samples of $z \in Z$. VAEs and AAEs each construct e to produce $z \in Z$ in different ways.

The output of the encoder of a VAE, e_{VAE} is $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The output of a VAE is then always Gaussian, and hence there is no limitation on the $\mathbf{z}$'s that e_{VAE} can produce. This ensures that $Q(Z|X) > 0$, provided that $\boldsymbol{\sigma} \neq \mathbf{0}$.

The encoder of our AAE, e_{AAE}, is a deep neural network consisting of multiple convolutional and batch normalisation layers. The final layer of the e_{AAE} is a fully connected layer without an activation function. The input to each of the M nodes in the fully connected layer is a function $f_{i=1...M}(\mathbf{x})$. This means that $\mathbf{z}$ is given by: $\mathbf{z} = \mathbf{a}_1 f_1(\mathbf{x}) + \mathbf{a}_2 f_2(\mathbf{x}) + ... + \mathbf{a}_M f_M(\mathbf{x})$, where $\mathbf{a}_{i=1...M}$ are the learned weights of the fully connected layer. We now consider three cases:

Case 1: If $\mathbf{a}_i$ are a complete set of bases for Z then it is possible to generate any $\mathbf{z} \in Z$ from an $\mathbf{x} \in X$ with a one-to-one mapping, provided that $f_i(\mathbf{x})$ is not restricted in the values that it can take.

Case 2: If $\mathbf{a}_i$ are an overcomplete set of bases for Z, then the same holds, provided that $f_i(\mathbf{x})$ is not restricted in the values that it can take.

Case 3: If $\mathbf{a}_i$ are an undercomplete set of bases for Z then it is not possible to generate all $\mathbf{z} \in Z$ from $\mathbf{x} \in X$. Instead there is a many (X) to one (Z) mapping.

For $Q(Z|X) > 0$ our network must learn a complete or overcomplete set of bases and $f_i(x)$ must be unconstrained $\forall i$. The network is encouraged to learn an overcomplete set of bases by learning a large number of $\mathbf{a}_i$'s—specifically $M = 8192$ when basing our network on the DCGAN architecture (Radford et al., 2015)—more that 40 times the dimensionality of Z. By using batch normalisation layers throughout the network, we ensure that values of $f_i(x)$ are spread out, capturing a *close-to-Gaussian* distribution, encouraging infinite support.

We have now shown that, under certain reasonable assumptions, $Q(X|Z) > 0$ and $Q(Z|X) > 0$, which means that $T(Z_{t+1}|Z_t) > 0$, and hence we can get from any Z to any another Z in only one step. Therefore the Markov chain described by the transition operator $T(Z_{t+1}|Z_t)$ defined in Equation 1 is both irreducible and aperiodic, which are the necessary conditions for ergodicity.

B ADVERSARIAL AUTOENCODERS

B.1 SAMPLES

(a) AAE (initial) (b) AAE (1 step) (c) AAE (5 steps) (d) AAE (10 steps)

(e) DAAE (initial) (f) DAAE (1 step) (g) DAAE (5 steps) (h) DAAE (10 steps)

Figure 5: Samples from an AAE (a-d) and DAAE (e-h) trained on the CelebA dataset. Even though AAEs can theoretically approximate $P(Z)$ better than VAEs, which are limited by their optimisation of $D_{KL}[Q(Z|X)\|P(Z)]$ (Makhzani et al., 2015), the adversarial criterion for deterministic AAEs is difficult to optimise when the dimensionality of Z is high. We observe that during training our AAEs and DAAEs, the empirical standard deviation of $\mathbf{z} \sim Q(Z|X)$ is less than 1, which means that $Q(Z)$ fails to approximate $P(Z)$ as closely as was achieved with the VAE and DVAE. However, this means that the effect of MCMC sampling is more pronounced, with the quality of all samples noticeably improving after a few steps. As a side-effect of the suboptimal solution learned by the networks, the denoising properties of the DAAE are more noticeable with even the novel samples.

B.2 INTERPOLATIONS

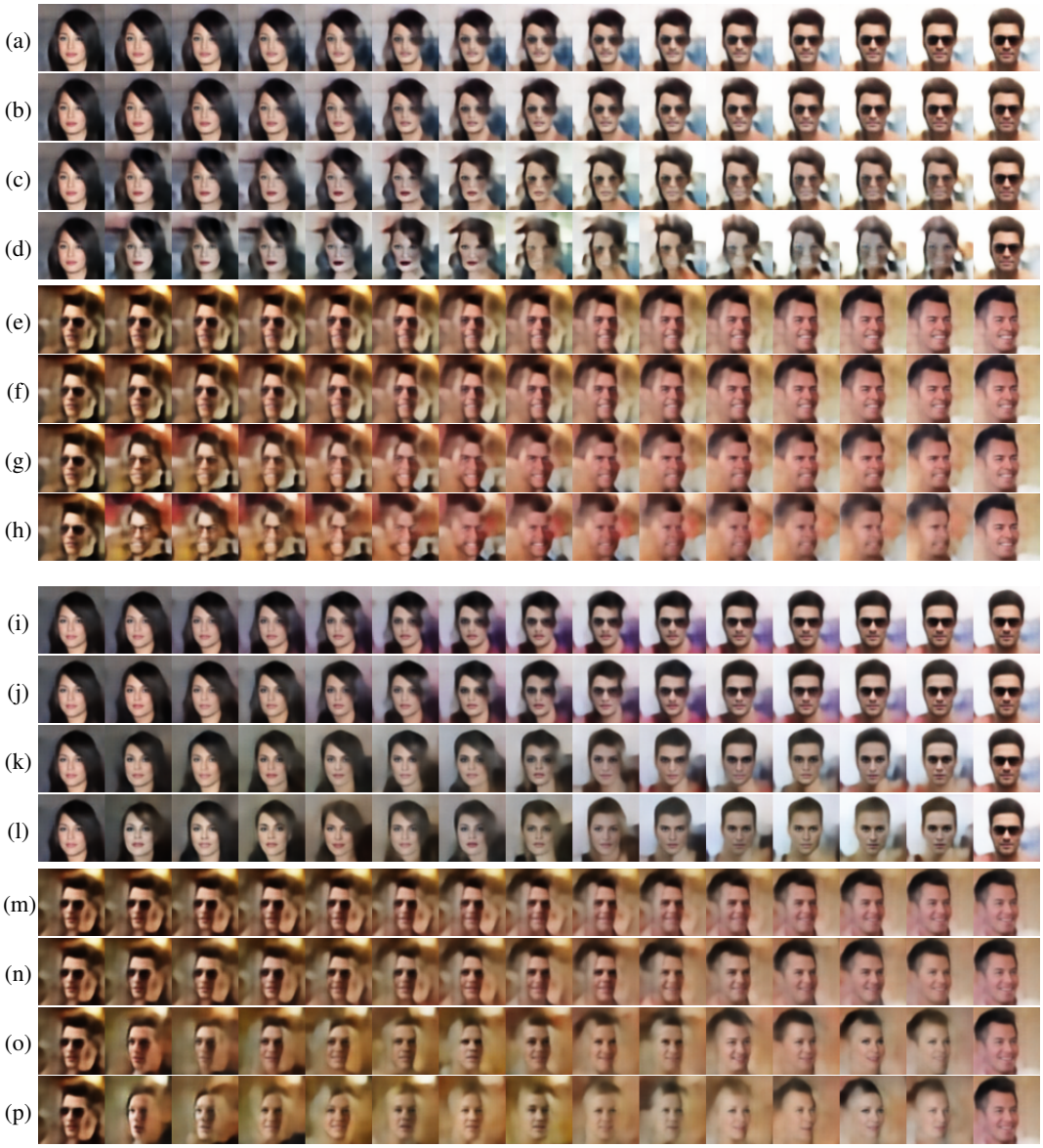


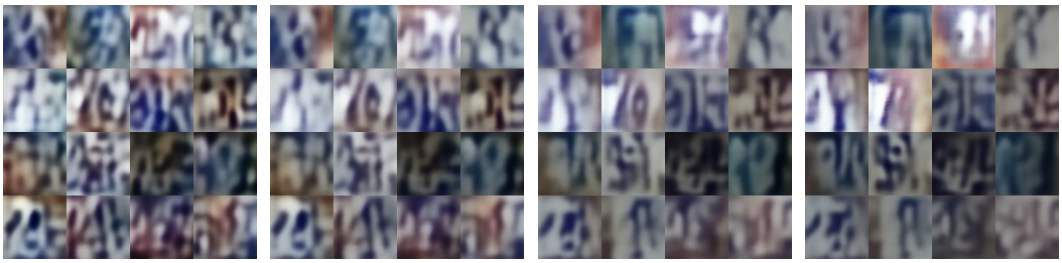
Figure 6: Interpolating between two faces using (a-h) an AAE and (i-p) a DAAE. The top row for each face is the original interpolation, whilst the second, third and fourth rows are the result of 1, 5 and 10 steps of MCMC sampling respectively. Although the AAE performs poorly, the regularisation effect of denoising can be clearly seen with the DAAE after applying MCMC sampling.

C STREET VIEW HOUSE NUMBERS

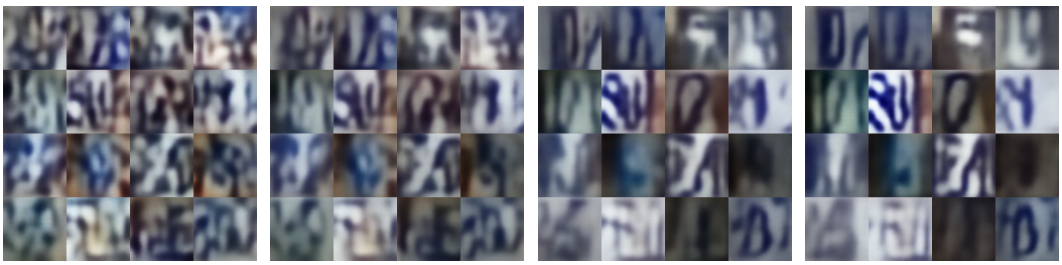
C.1 SAMPLES

(a) VAE (initial) (b) VAE (1 step) (c) VAE (5 steps) (d) VAE (10 steps)

(e) DVAE (initial) (f) DVAE (1 step) (g) DVAE (5 steps) (h) DVAE (10 steps)



(i) AAE (initial) (j) AAE (1 step) (k) AAE (5 steps) (l) AAE (10 steps)



(m) DAAE (initial) (n) DAAE (1 step) (o) DAAE (5 steps) (p) DAAE (10 steps)

Figure 7: Samples from a VAE (a-d), DVAE (e-h), AAE (i-l) and DAAE (m-p) trained on the SVHN dataset. The samples from the models imitate the blurriness present in the dataset. Although very few numbers are visible in the initial sample, the VAE and DVAE produce recognisable numbers from most of the initial samples after a few steps of MCMC sampling. Although the AAE and DAAE fail to produce recognisable numbers, the final samples are still a clear improvement over the initial samples.

C.2 INTERPOLATIONS

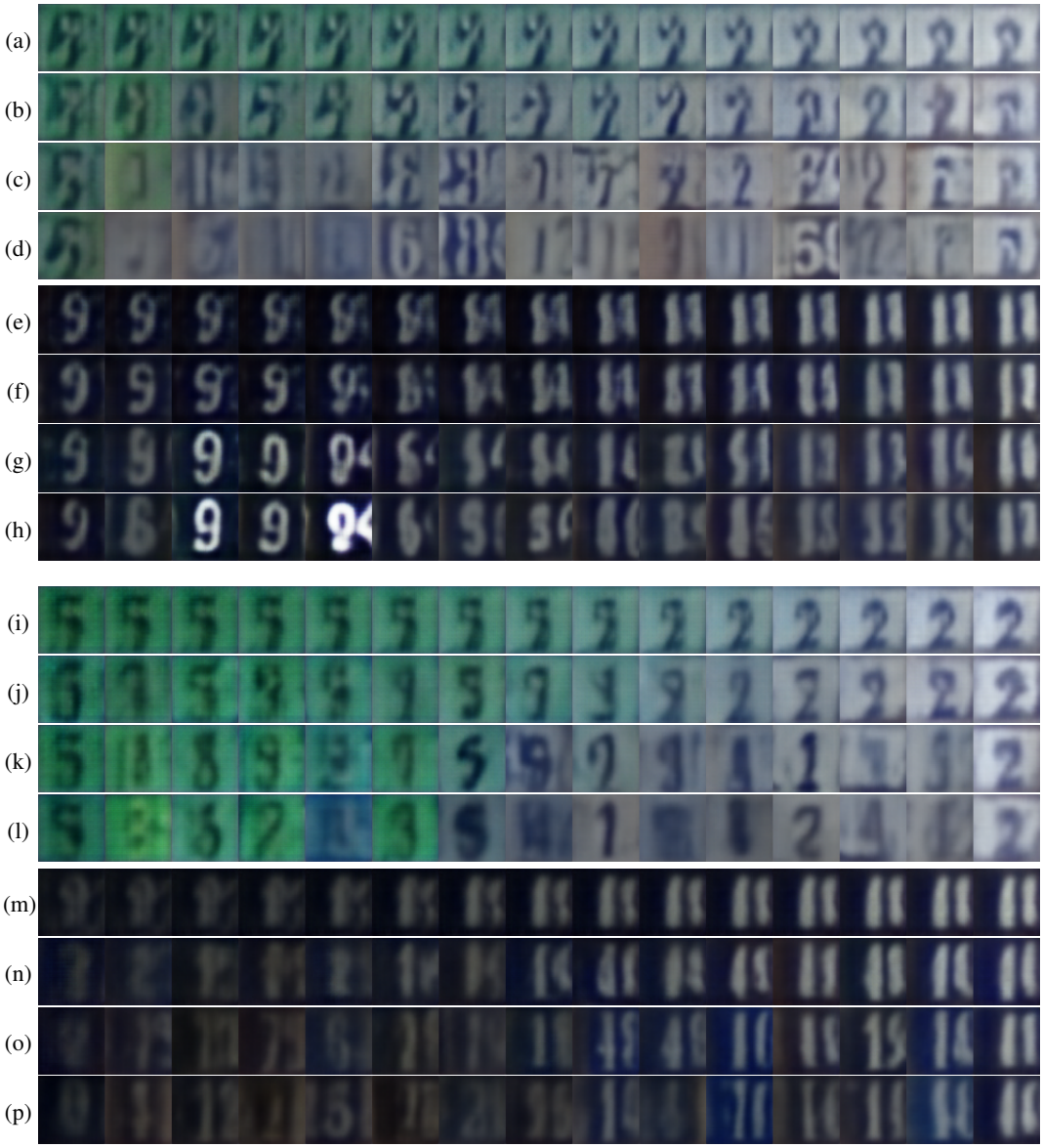


Figure 8: Interpolating between Google Street View house numbers using (a-h) a VAE and (i-p) a DVAE. The top row for each face is the original interpolation, whilst the second, third and fourth rows are the result of 1, 5 and 10 steps of MCMC sampling respectively. If the original interpolation is poor, as observed in (a), the models will attempt to move the samples towards a more realistic number (c-d). Interpolation between 1- and 2-digit numbers in an image (e, m) results in a meaningless blur in the middle of the interpolation. After a few steps of MCMC sampling the models instead produce recognisable 1- or 2-digit numbers (h, p).