Laplacian Pyramid of Conditional Variational Autoencoders

Garoe Dorta*† Sara Vicente† Lourdes Agapito‡ Neill Campbell* Ivor Simpson†

University of Bath* Anthropics Technology Ltd.† University College London‡

Introduction

Intuitive image editing requires good image models, which include information such as what objects are present in the picture and their pose. General purpose software requires extensive knowledge to successfully execute any non-trivial modification.

By finding a transformation that embeds some input data in a low-dimensional space, less general but more powerful image models can be built.

Recent advances in Deep Convolutional Neural Networks (DNN) has shown state-of-the-art results in image modeling using latent embeddings [1].

We present LapCVAE a new DNN architecture, which decomposes image generation process into smaller tractable steps.

Related work

Deep Variational Autoencoders (VAE) [1] are popular DNN generative models. The framework is derived from a directed probabilistic model, and learning the parameters is made tractable by employing a variational approximation for the marginal likelihood of the data. Coarse to fine approaches have been used to improve the image quality for other DNN models, such as Generative Adversarial Networks (GAN) [2] and PixelCNN [3]. Hybrids of GAN and VAE have also shown promising results as generative models for natural images [4].

Methodology

The objective is to estimate the network parameters ϕ and θ, such that the likelihood of the training data $p_{\boldsymbol{\theta}}(\mathbf{x})$ is maximized:

$$\log\left[p_{\boldsymbol{\theta}}(\mathbf{x})\right] = D_{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right] + L, \tag{1}$$

where the variational lower bound is

$$L = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - D_{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z})\right]. \tag{2}$$

where the $\mathbf{x}$ denotes the input data, $\mathbf{z}$ is the latent low-dimensional representation of $\mathbf{x}$, $p_{\boldsymbol{\theta}}(\cdot)$ is a decoder and $q_{\boldsymbol{\phi}}(\cdot)$ is an encoder, and $p_{\boldsymbol{\theta}}(\mathbf{z}))$ is a prior defined as a Gaussian with zero mean and unit variance.

A Laplacian pyramid is an invertible image decomposition, containing a set of images which encode high-frequency details at different scales. To generate a pyramid of K levels, given an input image $\mathbf{x}$, a blur and downsample operator $d(\cdot)$ is repeatedly applied to $\mathbf{x}$. This step creates a list of blurred images $[\mathbf{b}_0, \mathbf{b}_1, \cdots \mathbf{b}_K]$. The images in each level of the Laplacian pyramid are computed as

$$\mathbf{h}_k = \mathbf{b}_k - u(\mathbf{b}_{k+1}), \tag{3}$$

where $u(\cdot)$ is a blur and upsample operator and $\mathbf{h}_K = \mathbf{b}_K$.

A series of VAEs are used to generate the images in the pyramid

$$\mathbf{x}_k = p_k(\mathbf{h}_k|\mathbf{z}_k, u(\mathbf{x}_{k+1}); \boldsymbol{\theta}_k) + u(\mathbf{x}_{k+1}), \tag{4}$$

We add a tighter regularizer for all but the coarsest level of the pyramid, which encourages the latent space of a given level to be a transformation of the previous level latent distribution

$$p_k(\mathbf{z}; \boldsymbol{\theta}) = M_k\left[\mathcal{N}(\boldsymbol{\mu}_{k+1}, \boldsymbol{\sigma}_{k+1}); \boldsymbol{\psi}_k\right], \tag{5}$$

where M_k is a function parametrized by $\boldsymbol{\psi}_i$, λ_i is an user defined weight, and $\boldsymbol{\mu}_{k+1}$ and $\boldsymbol{\sigma}_{k+1}$ are the means and variances of the Gaussian distribution of the latent space at the coarser level.

Conclusions and Future Work

We presented a conditional multi-scale extension of VAE models. A possible avenue of future work to improve the quality of the generated images is to use a perceptual or adversarial loss for the reconstruction error, analogous to the work in Larsen et al. [4].

Results

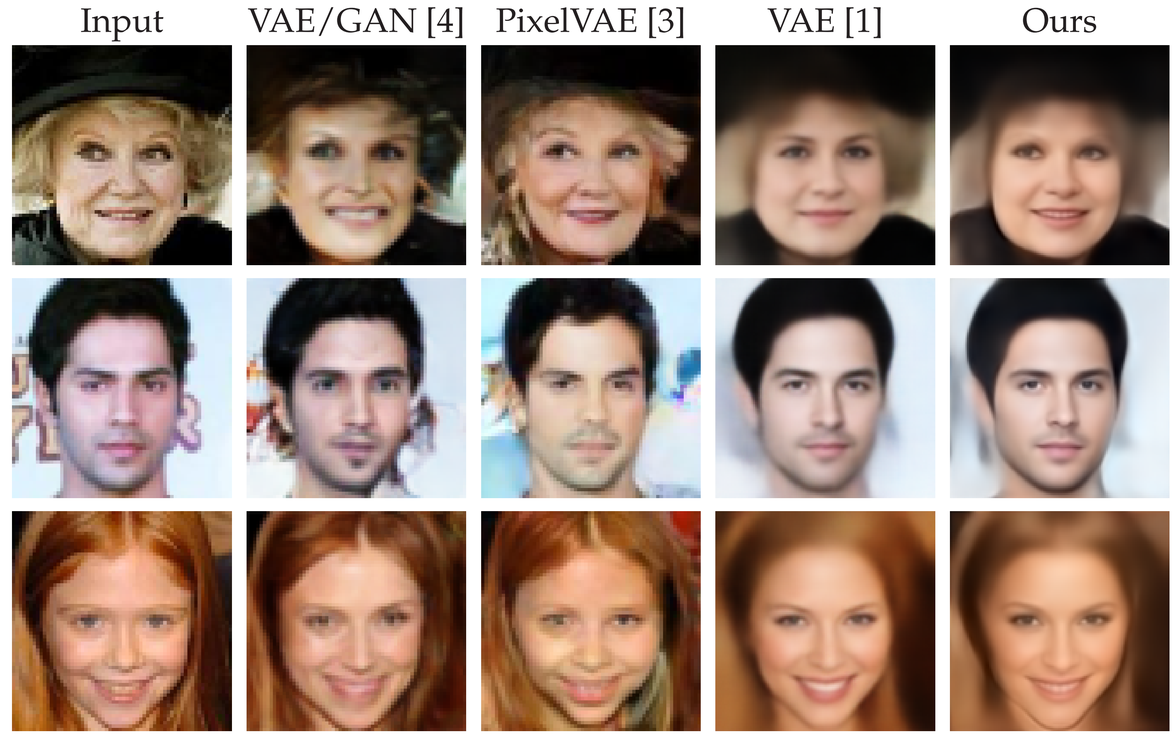

Figure 1: Comparison of image reconstruction results for different AutoEncoder architectures on the test data. VAE/GAN and PixelVAE add additional details in the images that do not necessarily correspond to the input. Our architecture produces sharper images than VAE.

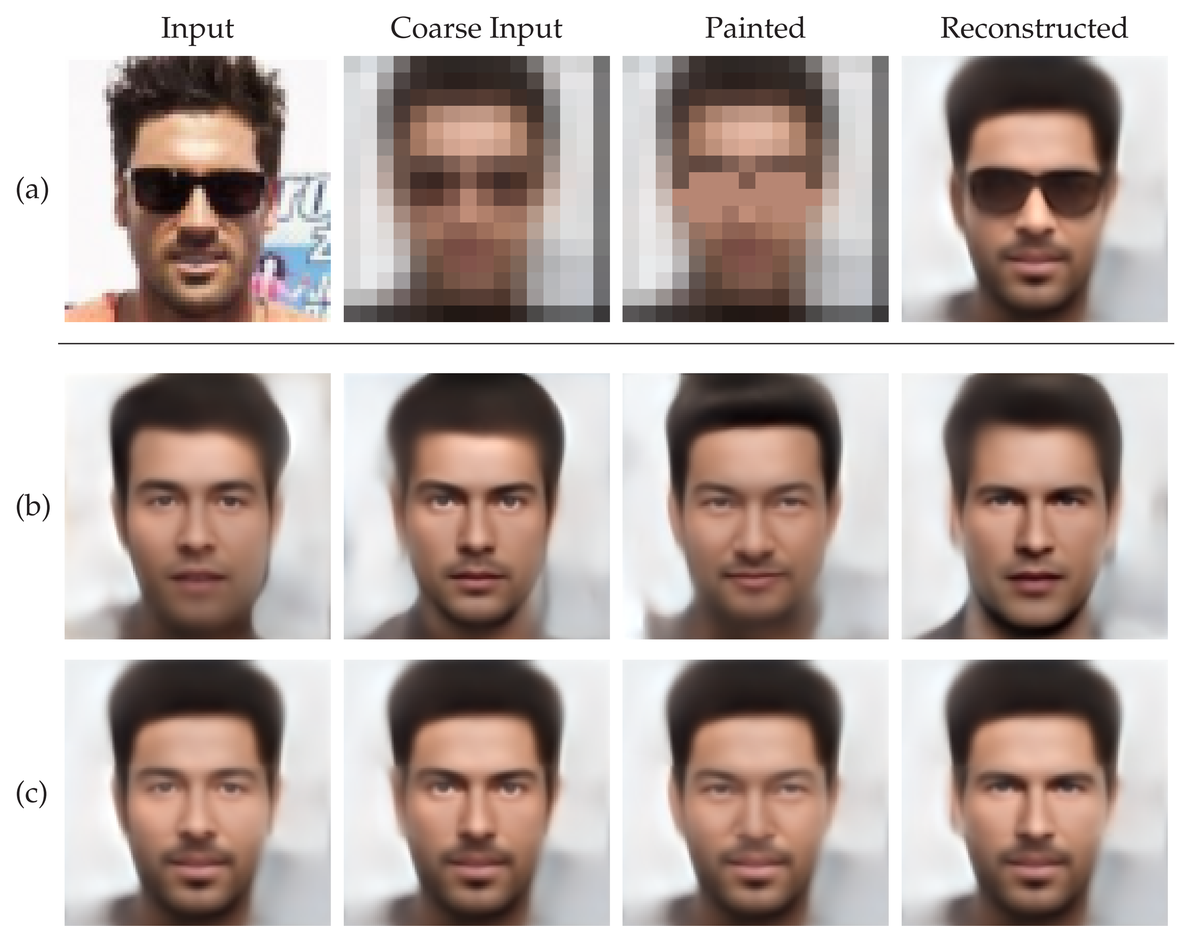

Figure 2: Image editing example. (a) An input image is decomposed using a Laplacian pyramid. The user paints into the coarse image. (b) Several samples are generated conditioned on this. (c) The selected regions are composed with the original reconstructed image.

Our model is evaluated on a faces dataset, using 128×128 images and four levels in the pyramid. Examples of reconstructions are shown in Figure 1, where our architecture achieves sharper results than previous work.

Moreover, our model can be used for novel images editing applications as shown in Figure 2, where the sunglasses in a face are removed by painting in a coarse level.

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[2] Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494. Curran Associates, Inc., 2015.

[3] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A Latent Variable Model for Natural Images. In *International Conference on Learning Representations (ICLR)*, 2017.

[4] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *International Conference on Machine Learning (ICML)*, 48, 2016.