Variational Image Inpainting

Cusuh Ham*, **Amit Raj***, **Vincent Cartillier***, **Irfan Essa**
Georgia Institute of Technology
{cusuh, amit.raj, vcartillier3, irfan}@gatech.edu

Abstract

We present several variational methods for image generation and image completion. We explore attribute-free and attribute-based settings and demonstrate the effectiveness of variational models in one-shot generation. We also present models that are robust to partial observation in the input image and are better suited for the image completion task. We benchmark the performance of these models on the CelebA dataset and demonstrate state of the art results in the attribute-guided image completion task.

1 Introduction

Understanding the generative model for the space of images is an important requirement for many computer vision tasks such as image denoising, image inpainting, and image dataset augmentation. Generative adversarial networks [5, 10] and variational methods [3] have been successfully employed to generate complex real world images. Most of these methods attempt to find a low-dimensional manifold that the natural images exist on and a corresponding mechanism to sample from this manifold. We build on these frameworks by proposing a set of methods for attribute-free and attribute-based image generation and further extend these models to image in-painting. We use a variational autoencoder (VAE) [7] model and incorporate perceptual loss using a pretrained classification network and demonstrate its improvement over a vanilla VAE. We show that this is equivalent to the maximum likelihood estimate of the data under "neural" class of distributions. Additionally, introducing attributes presents a harder task than attribute-free generation since the attributes need to be de-correlated from the latent variables and from each other. However, this assumption rarely holds when we have a pre-specified set of attributes. We demonstrate multiple models that incorporate attribute information, either in the encoder, the decoder or both.

2 Related Works

Autoencoders are graphical models composed of a generative model of data given latent variables and a recognition model for the latent variables. Bengio et al. [2] present a generalized framework to use autoencoders as generative models for a wide class of distributions. Kingma et al. [7] introduced the first framework to formulate variational inference on neural networks. VAEs provide a mechanism for sampling from certain classes of distributions that can be approximated with a neural network. Furthermore, they can be trained with only gradient-based methods, making them conducive for learning distributions over high-dimensional data. Recent work by Yan et al. [15] and Vedantam et al. [13] have demonstrated the ability to generate images using conditional autoencoders. Yan et al. use an autoencoder to disentangle an image into its attributes and use this representation to generate images. Vedantam et al. use a product of experts models to hierarchically calculate the posterior and generate images given any subset of attributes. More recents methods like the VAE-GAN [8], CVAE-GAN [1] and ALI [4] have explored a hybrid approach to generation that combines aspects of both variational autoencoders and GANS.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

3 Approach

We will expand on using VAEs for image generation and completion using the CelebA dataset [9]. The work by [15] propose a model based on the CVAE framework. We propose to extend their work by changing the objective function to incorporate better image generation. The setup is similar to Yan et al. [13] in that we will have a modified ELBO loss that measures diversity and "success" of image completion based on the discovered attributes. Additionally, we will factor in the perceptual loss [6] to ensure that the generated images look perceptually accurate. We will evaluate the performances using the newly proposed perceptual distance [16].

3.1 Latent variable

The latent variable approach is an extension of the architecture used by Yan et al.[15] which uses a CVAE to learn a generative model of image and attributes.

We assume a set of latent factors that explain the image and serve as sufficient statistics for image generation. We build a recognition network to approximate the latent distribution given an image. More concretely, for a data point X and latent variable Z, we instantiate the generative model $p_\theta(X|Z)$ and we approximate $q_\phi(Z|X)$ as the variational lower bound of the posterior $p_\theta(Z|X)$. We assume a Gaussian prior on Z and maximize $\mathcal{L} = \mathbb{E}[ELBO(x;\theta,\phi)]$ where:

$$ELBO(x;\theta,\phi) = \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)||p_\theta(z)) \tag{1}$$

The MLE of θ with a Gaussian prior is the same as optimizing for the reconstruction error of x. However, we note that this MLE is not an unbiased estimate unless $p(z|x)$ and $q(z|x)$ match exactly. In addition to the reconstruction loss, we add perceptual loss to the objective as follows and learn the parameters of the generative model using the updated objective.

$$\mathcal{L}_{recon} = \left\|x_{gen} - x_{gt}\right\|^2 + \sum_l \lambda_l \left\|\eta_l(x_{gen}) - \eta_l(x_{gt})\right\|^2 \tag{2}$$

Where x_{gen} denotes the generated image and x_{gt} denotes the ground truth image. η_l refers to the activation of the l^{th} layer of a pre-trained VGG16 classification network. The λ_l terms decide the contribution of each layer towards the loss. The addition of these terms is akin to increasing the statistics to match an image. For a purely Gaussian assumption on the output, the first term is equivalent to matching the mean statistics of each pixel in the image. This leads to often blurry generated images. However, if we also assume that the features of images are Gaussian distributed for a class of images, the additional terms from VGG act as adding higher order non-linear relations between pixels. By matching the statistics of nonlinear interactions between pixels in the generated image and ground truth image, more information can be captured in the image, leading to better generation quality.

We consider a variation of this previous model on partially observed images where the generative model remains $p_\theta(x|z)$ whilst the recognition model is modified to incorporate partially observed data x' as $q_\phi(z|x')$. This makes the latent space more robust to partially observed data by learning to capture the structure in the data better. For both models, we can generate novel images by sampling $z \sim \mathcal{N}(0,1)$.

For image completion, we perform an iterative variational Gibbs sampling approach during test time as demonstrated by Rezende et al. [11], given by:

$$Z_t \sim q_\phi(Z|X'_t) \qquad X_t \sim p_\theta(X|Z_t) \qquad X'_{t+1} = m \odot X'_t + (1-m) \odot X_t$$

where m is a binary mask of the cropped region, X'_t is the cropped input image, and X_t is the reconstructed image from the latent representation of X'_t.

We can refine the model further to be better suited for the task of image completion at the cost of reduced generation performance. In the previous approach, we only used the latent representation, Z, to reconstruct the image. As before, we do not use the attribute

annotations, but now we utilize the partially observed input image, X', and its latent representation, Z, as input to the network. Thus, the generative model is described by $p_\theta(X, Z, X')$ and the variational approximation by $q_\phi(Z, X')$.

The key difference in this approach is that the iterative sampling step from the previous approach now occurs during training time. This allows us to perform one-shot image completion at test time.

3.2 Latent variables and attributes

In this section we include the attributes in the network architecture and develop the optimal generative and recognition network suited for attribute-based generation and image completion tasks.

First, we pass the full attribute label, Y, and the latent representation, Z, of the corresponding image to the generative model as $p_\theta(x|z, y)$. The recognition model encodes some non-interpretable latents as $q_\phi(z|x)$. Although this is a natural way of introducing attributes into the generation process, it often leads to redundant encoding.

Ideally, latents and attributes need to be de-correlated from each other. The feature attributes can be added into the network in three different places: the encoder, decoder or both.

When the attribute is passed to only the encoder, we sample $Z \sim q(Z|X', Y)$ and $X \sim p(X|Z)$. When the attribute is passed to only the decoder, we sample $Z \sim q(Z|X')$ and $X \sim p(X|Z, Y)$. When the attribute is passed to both the encoder and decoder, we sample $Z \sim q(Z|X', Y)$ and $X \sim p(X|Z, Y)$.

4 Experiments

We present our results on the CelebA dataset [9]. The recognition model is an approximation of the posterior defined by the generative model. Figure 1 demonstrates our best qualitative results on image completion achieved by including the attributes in both the encoder and decoder. As shown in the figure, this model performs image completions consistent with the attribute with high probability. Additional results are included in the Appendix.

Figure 1: Results of feeding attribute to both encoder and decoder.

5 Conclusion

In this paper we introduce several methods for image generation in increasing complexity. We incorporate techniques such as an iterative MCMC sampling process and injecting the attribute vector in different places of the model to show improvements on the image completion task. We evaluate our results on a modified version of the CelebA dataset with qualitative evaluations as well as quantitative metrics, including the Inception score, perceptual distance, and SSIM.

6 Appendix

6.1 Fully observed data without attributes

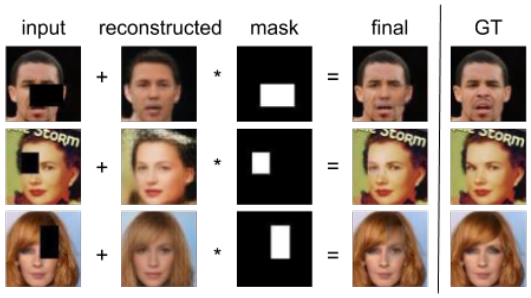

Figure 2: Results of training on complete images without attributes. This model is a VAE trained with an additional VGG loss term. Where $x \sim p_\theta(x|z)$ and $z \sim q_\phi(z|x)$. For image x and latent z. Inpainting is performed by iterative sampling.

6.2 Partially observed data without attributes

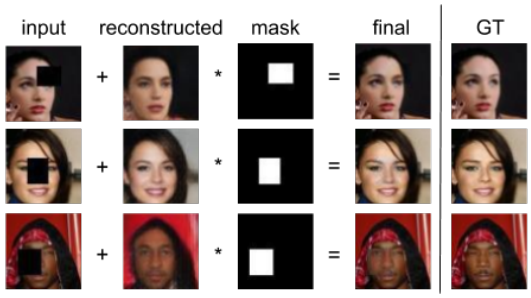

Figure 3: Results of training on cropped images without attributes. The encoder is trained with partially observed images, $x \sim p_\theta(x|z)$ and $z \sim q_\phi(z|x')$. For image x and latent z and partially observed image x'.

6.3 Comparison of fully observed and partially observed data

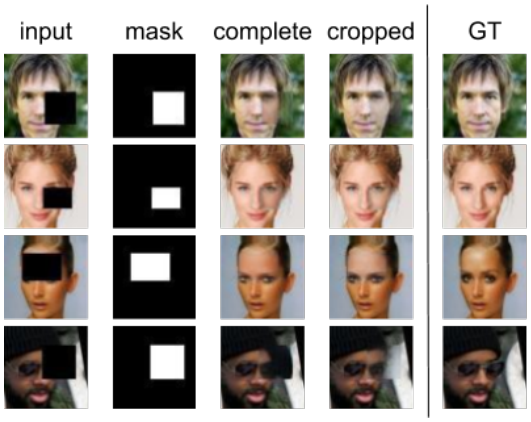

Figure 4: Side-by-side comparison of training on complete vs. cropped images.

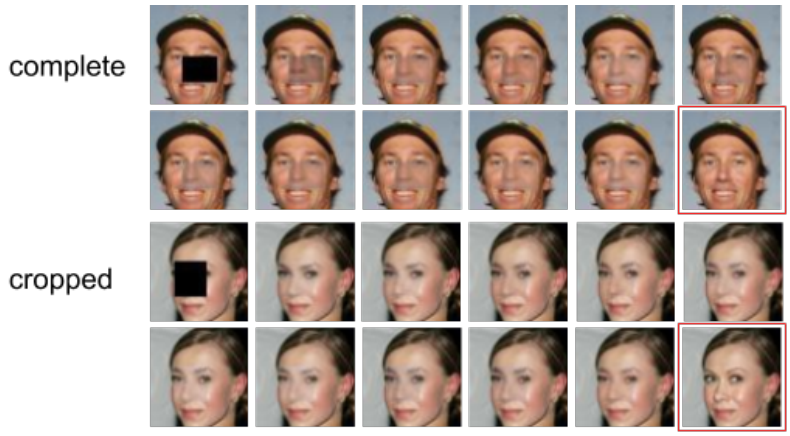

Figure 5: Comparison of the iterative sampling approach using the latent variable on complete and cropped images. The red box indicates the ground truth image.

6.4 Fully observed data with 1 attribute

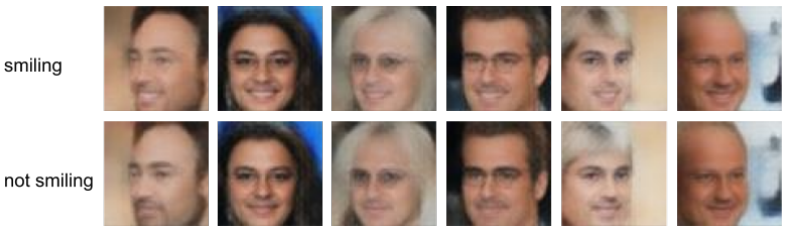

Figure 6: Attribute-based generation. We sample $z \sim N(0, I)$ and use the decoder trained on fully observed images and attributes, to generate $x \sim p_\theta(x|z, y)$,with different attributes for the same latent vector, demonstrating that the latents are disentangles from the attributes
.

6.5 Comparison of conditioning only the encoder or the decoder on attributes

(a) Results of conditioning only the encoder on attributes.

(b) Results of conditioning only the decoder on attributes.

Figure 7: Side-by-side comparison of the two methods.

6.6 Quantitative results

We evaluate the results of our various approaches using the Inception score [12], perceptual distance [17], and the Structural Similarity (SSIM) metric [14].

Model	Inception	Perceptual Distance	SSIM
Complete + no attributes	3.096±0.172	0.175±0.002	0.451±0.001
Cropped + no attributes	2.934±0.044	0.169±0.002	0.446±0.001
Cropped + no attr, unmasked recon	2.217±0.052	0.165±0.002	0.437±0.002
Complete + all attributes	2.904±0.078	0.167±0.002	0.450±0.001
Complete + all attr, unmasked recon	2.017±0.061	0.180±0.002	0.435±0.002
Cropped + 1 attribute (E)	2.522±0.100	0.167±0.002	0.417±0.001
Cropped + 1 attribute (D)	2.468±0.093	0.166±0.002	0.417±0.001
Cropped + 1 attribute (E + D)	2.496±0.065	0.167±0.002	0.416±0.001
Real images	2.744±0.140	-	-

Table 1: Quantitative evaluations of our proposed methods.

References

[1] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Cvae-gan: fine-grained image generation through asymmetric training. *arXiv preprint arXiv:1703.10155*, 2017. 1

[2] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013. 1

[3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. 1

[4] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 1

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[6] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 2

[7] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[8] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 1

[9] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 2, 3

[10] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[11] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 2

[12] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 5

[13] R. Vedantam, I. Fischer, J. Huang, and K. Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017. 1, 2

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[15] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016. 1, 2

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*, 2018. 2

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018. 5