Autoencoding beyond pixels using a learned similarity metric

Anders Boesen Lindbo Larsen[1] ABLL@DTU.DK
Søren Kaae Sønderby[2] SKAAESONDERBY@GMAIL.DK
Ole Winther[1,2] OLWI@DTU.DK
[1] Department for Applied Mathematics and Computer Science, Technical University of Denmark
[2] Bioinformatics Centre, Department of Biology, University of Copenhagen

Abstract

We present an autoencoder that leverages the power of learned representations to better measure similarities in data space. By combining a variational autoencoder (VAE) with a generative adversarial network (GAN) we can use learned feature representations in the GAN discriminator as basis for the VAE reconstruction objective. Thereby, we replace element-wise errors with feature-wise errors that better capture the data distribution while offering invariance towards e.g. translation. We apply our method to images of faces and show that our method outperforms VAEs with element-wise similarity measures in terms of visual fidelity. Moreover, we show that our method learns an embedding in which high-level abstract visual features (e.g. wearing glasses) can be modified using simple arithmetic.

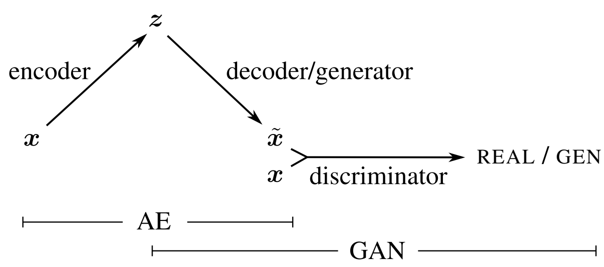

Figure 1. Overview of our network. We combine a VAE with a GAN by collapsing the decoder and the generator into one.

1. Introduction

Deep architectures have allowed a wide range of discriminative models to scale to large and diverse datasets. However, generative models still have problems with complex data distributions such as images and sound. In this work, we show that currently used similarity metrics impose a hurdle for learning good generative models and that we can improve a generative model by employing a learned similarity measure.

When learning models such as the variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014), the choice of similarity metric is central as it provides the main part of the training signal via the reconstruction error objective. For this task, the element-wise Euclidean distance is a commonly used metric. This metric is simple but not very suitable for image data, as it does not model the properties of human visual perception. E.g. a small image translation might result in a large pixel-wise error whereas a human would barely notice the change. Therefore, we argue in favor of measuring image similarity using a higher-level and *sufficiently invariant* representation of the images. Rather than hand-engineering a suitable measure to accommodate the problems of element-wise metrics, we want to learn a function for the task. The question is how to learn such a similairty measure? We find that by jointly training a VAE and a GAN we can use the GAN discriminator to measure sample similairty. We achieve this by combining a VAE with a GAN (Goodfellow et al., 2014) as shown in Fig. 1. We collapse the VAE decoder and the GAN generator into one by letting them share parameters and training them together. For the VAE training objective, we replace the typical element-wise reconstruction metric with a feature-wise metric expressed in the discriminator.

Preliminary work. Comments and suggestions are most welcome!

1.1. Contributions

- We combine VAEs and GANs into an unsupervised generative model that simultaneously learns to *encode*, *generate* and *compare* dataset samples.
- We show that a model trained with learned error measures generates better image samples.

- We demonstrate that unsupervised training result in a latent image representation with disentangled factors of variation (Bengio et al., 2013).

2. Autoencoding with learned similarity

In this section we provide background on VAEs and GANs and introduce a slight modification of GANs that makes combining VAEs and GANs more straightforward. Furthermore, we replace the expected log likelihood in the VAE with a reconstruction error expressed by the GAN.

Variational autoencoder A VAE consists of two networks that *encode* a data sample $\boldsymbol{x}$ to a latent representation $\boldsymbol{z}$ and *decode* the latent representation back to data space, respectively:

$$\boldsymbol{z} \sim \mathrm{Enc}(\boldsymbol{x}) = q(\boldsymbol{z}|\boldsymbol{x}) \,, \quad \tilde{\boldsymbol{x}} \sim \mathrm{Dec}(\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z}) \,. \quad (1)$$

The VAE regularizes the encoder by imposing a prior over the distribution of the encoding $p(\boldsymbol{z})$. Typically $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is chosen. The VAE loss is minus the sum of the expected log likelihood (the reconstruction error) and a prior regularization term:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right] = \mathcal{L}_{\text{llike}}^{\text{pixel}} + \mathcal{L}_{\text{prior}} \quad (2)$$

with

$$\mathcal{L}_{\text{llike}}^{\text{pixel}} = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z})\right] \quad (3)$$

$$\mathcal{L}_{\text{prior}} = D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})) \,, \quad (4)$$

where D_{KL} is the Kullback-Leibler divergence.

Generative adversarial network A GAN consists of two networks: the *generator* network $\mathrm{Gen}(\boldsymbol{z})$ maps latents $\boldsymbol{z}$ to data space while the *discriminator* network assigns probability $y = \mathrm{Dis}(\boldsymbol{x}) \in [0, 1]$ that $\boldsymbol{x}$ is an actual training sample and probability $1 - y$ that $\boldsymbol{x}$ is generated by our model through $\boldsymbol{x} = \mathrm{Gen}(\boldsymbol{z})$ with $\boldsymbol{z} \sim p(\boldsymbol{z})$. The GAN objective is to find the binary classifier that gives the best possible discrimination between true and generated data and simultaneously encouraging Gen to fit the true data distribution. We thus aim to maximize/minimize the binary cross entropy:

$$\mathcal{L}_{\text{GAN}} = \log(\mathrm{Dis}(\boldsymbol{x})) + \log(1 - \mathrm{Dis}(\mathrm{Gen}(\boldsymbol{z}))) \,, \quad (5)$$

with respect to Dis / Gen with $\boldsymbol{x}$ being a training sample and $\boldsymbol{z} \sim p(\boldsymbol{z})$.

We generalize GANs by replacing the deterministic generator network with a sample from the VAE decoder:

$$\mathcal{L}_{\text{GAN}} = \log(\mathrm{Dis}(\boldsymbol{x})) + \log(1 - \mathrm{Dis}(\tilde{\boldsymbol{x}})) \quad (6)$$

with $\tilde{\boldsymbol{x}} \sim \mathrm{Dec}(\boldsymbol{z})$ and $\boldsymbol{z} \sim p(\boldsymbol{z})$. This parameter sharing step allows us to combine the VAE encoder/decoder framework with GAN.

2.1. Beyond element-wise reconstruction error

Since element-wise reconstruction errors are not adequate for images and other signals with invariances, we propose replacing the VAE reconstruction (expected log likelihood) error term from Eq. 3 with a reconstruction error expressed in the GAN discriminator. To achieve this let $\mathrm{Dis}_l(\boldsymbol{x})$ denote the hidden representation of the lth layer of the discriminator. We introduce a Gaussian observation model for $\mathrm{Dis}_l(\boldsymbol{x})$ with mean $\mathrm{Dis}_l(\tilde{\boldsymbol{x}})$ and identity covariance:

$$p(\mathrm{Dis}_l(\boldsymbol{x})|\boldsymbol{z}) = \mathcal{N}(\mathrm{Dis}_l(\boldsymbol{x})|\,\mathrm{Dis}_l(\tilde{\boldsymbol{x}}), \mathbf{I}) \,, \quad (7)$$

where $\tilde{\boldsymbol{x}} \sim \mathrm{Dec}(\boldsymbol{z})$ is the the sample from the decoder of $\boldsymbol{x}$. We can now the replace the VAE error Eq. 3 with

$$\mathcal{L}_{\text{llike}}^{\mathrm{Dis}_l} = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\mathrm{Dis}_l(\boldsymbol{x})|\boldsymbol{z})\right] \quad (8)$$

We train our combined model with the triple criterion

$$\mathcal{L} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{llike}}^{\mathrm{Dis}_l} + \mathcal{L}_{\text{GAN}} \,. \quad (9)$$

Notably, we optimize the VAE wrt. $\mathcal{L}_{\text{GAN}}$ which we regard as a *style* error in addition to the reconstruction error which can be interpreted as a *content* error using the terminology from Gatys et al. (2015).

In practice, we have observed the devil in the details during development and training of this model. We therefore provide a list of practical considerations in this section. We refer to Fig. 2 and Alg. 1 for overviews of the training procedure.

Limiting error signals to relevant networks Using the loss function in Eq. 9, we train both a VAE and a GAN simultaneously. This is possible because we do not update all network parameters wrt. the combined loss. In particular, Dis should not try to minimize $\mathcal{L}_{\text{llike}}^{\mathrm{Dis}_l}$ as this would collapse the discriminator to 0. We also observe better results by not backpropagating the error signal from $\mathcal{L}_{\text{GAN}}$ to Enc.

Weighting VAE vs. GAN As Dec receives an error signal from both $\mathcal{L}_{\text{llike}}^{\mathrm{Dis}_l}$ and $\mathcal{L}_{\text{GAN}}$, we use a parameter γ to weight the ability to reconstruct vs. fooling the discriminator. This can also be interpreted as weighting *style* and *content*. Rather than applying γ to the entire model (Eq. 9), we perform the weighting only when updating the parameters of Dec:

$$\boldsymbol{\theta}_{\text{Dec}} \stackrel{+}{\leftarrow} -\nabla_{\boldsymbol{\theta}_{\text{Dec}}}(\gamma \mathcal{L}_{\text{llike}}^{\mathrm{Dis}_l} - \mathcal{L}_{\text{GAN}}) \quad (10)$$

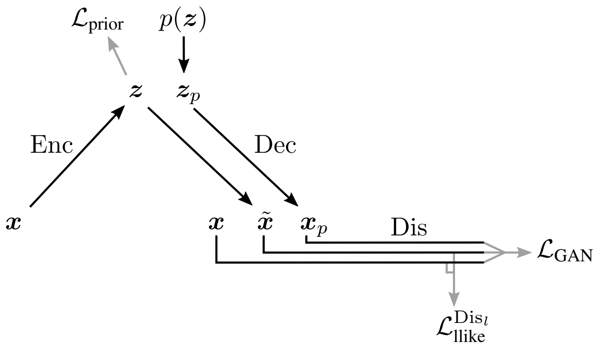

Figure 2. Flow through the combined VAE/GAN model during training. Gray lines represent terms in the training objective.

Algorithm 1 Training the VAE/GAN model

$\boldsymbol{\theta}_{\text{Enc}}, \boldsymbol{\theta}_{\text{Dec}}, \boldsymbol{\theta}_{\text{Dis}} \leftarrow$ initialize network parameters
repeat
  $\boldsymbol{X} \leftarrow$ random mini-batch from dataset
  $\boldsymbol{Z} \leftarrow \text{Enc}(\boldsymbol{X})$
  $\mathcal{L}_{\text{prior}} \leftarrow D_{\text{KL}}(q(\boldsymbol{Z}|\boldsymbol{X}) \| p(\boldsymbol{Z}))$
  $\tilde{\boldsymbol{X}} \leftarrow \text{Dec}(\boldsymbol{Z})$
  $\mathcal{L}_{\text{llike}}^{\text{Dis}_l} \leftarrow -\mathbb{E}_{q(\boldsymbol{Z}|\boldsymbol{X})}\left[p(\text{Dis}_l(\boldsymbol{X})|\boldsymbol{Z})\right]$
  $\boldsymbol{Z}_p \leftarrow$ samples from prior $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
  $\boldsymbol{X}_p \leftarrow \text{Dec}(\boldsymbol{Z}_p)$
  $\mathcal{L}_{\text{GAN}} \leftarrow \| \log(\text{Dis}(\boldsymbol{X})) + \log(1 - \text{Dis}(\tilde{\boldsymbol{X}})) + \log(1 - \text{Dis}(\boldsymbol{X}_p)) \|_1$
  // Update parameters according to gradients
  $\boldsymbol{\theta}_{\text{Enc}} \overset{+}{\leftarrow} -\nabla_{\boldsymbol{\theta}_{\text{Enc}}}(\mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{llike}}^{\text{Dis}_l})$
  $\boldsymbol{\theta}_{\text{Dec}} \overset{+}{\leftarrow} -\nabla_{\boldsymbol{\theta}_{\text{Dec}}}(\gamma \mathcal{L}_{\text{llike}}^{\text{Dis}_l} - \mathcal{L}_{\text{GAN}})$
  $\boldsymbol{\theta}_{\text{Dis}} \overset{+}{\leftarrow} -\nabla_{\boldsymbol{\theta}_{\text{Dis}}} \mathcal{L}_{\text{GAN}}$
until deadline

Sampling from $p(\boldsymbol{z})$ We observe better results by sampling directly from our prior $p(\boldsymbol{z})$ and using these in addition to $\text{Enc}(\boldsymbol{x})$ when training the GAN. We suspect that our regularization of the latent space does not quite achieve $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ which is problematic when we draw samples from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ at test time.

3. Experiments

Measuring the quality of generative models is challenging as current evaluation methods do not scale to natural images. Because the log-likelihood is intractable, it is often estimated using a Parzen window approach (Breuleux et al., 2011), but even for small images of 6×6 pixels this is shown to be biased in favor of methods that overfit (Theis et al., 2015). In this work, we use images of size 64x64 and perform a qualitative assessment only since current evaluation methods are not realiable for our problem size. We leave it as future work to evaluate the method with more comparable measures.

We investigate the performance of different generative models:

- Plain VAE with an element-wise Gaussian recognition model. This corresponds to a reconstruction error with an Euclidean distance.
- VAE with a learned distance ($\text{VAE}_{\text{Dis}_l}$). We first train a GAN and use the discriminator network as a learned similarity measure. We select a single layer l at which we measure the similariy according to Dis_l. l is chosen such that the comparison is performed after 3 downsamplings of each a factor of 2.
- The combined VAE/GAN model. This model is similar to $\text{VAE}_{\text{Dis}_l}$ but we also optimize Dec wrt. $\mathcal{L}_{\text{GAN}}$.
- GAN. This modes has recently been shown capable of generating high-quality images (Denton et al., 2015; Radford et al., 2015).

All models share the same architectures for Enc, Dec and Dis respectively. For all our experiments, we use convolutional architectures and use *backward convolution* (aka. *fractional striding*) with stride 2 to upscale images in Dec. Backward convolution is achieved by flipping the convolution direction such that striding causes upsampling. We omit further details concerning network construction and training and refer to our implementation available online[1].

3.1. Face images

We apply our methods to face images from the *CelebA* dataset[2] (Liu et al., 2015). This dataset consists of 202,599 images annotated with 40 binary attributes such as *eyeglasses*, *bangs*, *pale skin* etc. We scale and crop the images to 64×64 pixels and use only the images (not the attributes) for unsupervised training.

After training, we draw samples from $p(\boldsymbol{z})$ and propagate these through Dec to generate new images which are shown in Fig. 3. The plain VAE is able draw the frontal part of the face sharply, but off-center the images get blurry. This is because the dataset aligns faces using frontal landmarks. When we move too far away from the aligned parts, the recognition model breaks down because pixel correspondence cannot be assumed. $\text{VAE}_{\text{Dis}_l}$ produces sharper images even off-center because the reconstruction error is lifted beyond pixels. However, we see severe noisy artefacts which we believe are caused by the harsh downsampling scheme. In comparison, VAE/GAN and pure GAN produce sharper images with more natural textures and face

[1] http://github.com/andersbll/autoencoding_beyond_pixels

[2] We use the aligned and cropped images.

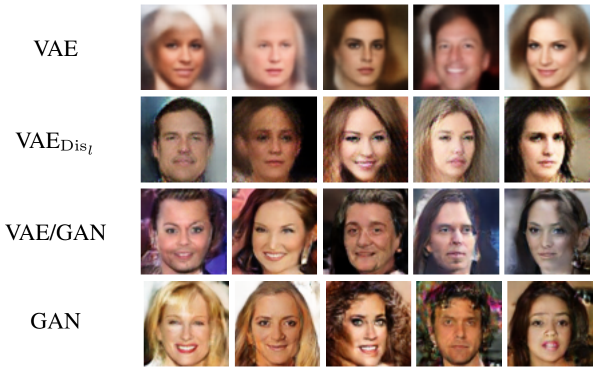

Figure 3. Samples from different generative models.

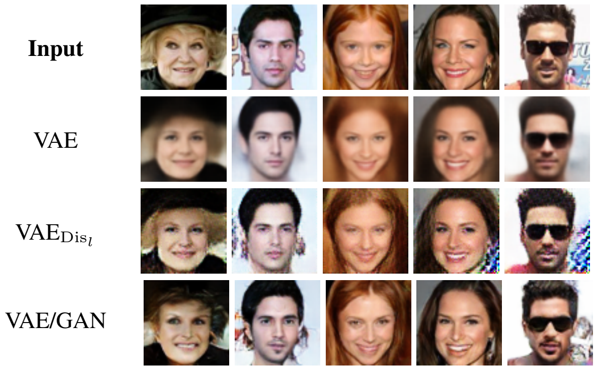

Figure 4. Reconstructions from different autoencoders.

parts.

Additionally, we make the VAEs reconstruct images taken from a separate test set. Reconstruction is not possible with the GAN model as it lacks an encoder network. The results are shown in Fig. 4 and our conclusions are similar to what we observed for the random samples. Note that VAE$_{\mathrm{Dis}_l}$ generates noisy blue patterns in some of the reconstructions. We suspect the GAN-based similarity measure can collapse to 0 in certain cases (such as the pattern we observe), which encourages Dec to generate such patterns.

3.1.1. Visual attribute vectors

Inspired by attempts at learning embeddings in which semantic concepts can be expressed using simple arithmetic (Mikolov et al., 2013), we inspect the latent space of a trained AE/GAN model. The idea is to find directions in the latent space corresponding to specific visual features in image space.

We use the binary attributes of the dataset to extract *visual attribute vectors*. For all images we use the encoder to calculate latent vector representation. For each attribute, we compute the mean vector for images with the attribute and the mean vector for images without the attribute. We then compute the visual attribute vector as the difference between the two mean vectors. This is a very simple method for calculating visual attribute vectors that will have problems with highly correlated visual attributes such as *heavy makeup* and *wearing lipstick*. In Fig. 5, we show face images as well as the reconstructions after adding different visual attribute vectors to the latent representations. Though not perfect, we clearly see that the attribute vectors capture semantic concept like *eyeglasses*, *bangs*, etc. E.g. when bangs are added to the faces, both the hair color and the hair texture matches the original face. We also see that being a man is highly correlated with having a mustache which caused by attribute correlations in the dataset.

4. Related work

Element-wise distance measures are notoriously inadequate for complex data distributions like images. In the computer vision community, preprocessing images is a prevalent solution to improve robustness to certain perturbations. Examples of preprocessing are contrast normalization, working with gradient images or pixel statistics gathered in histograms. We view these operations as a form of metric engineering to account for the shortcomings of simple element-wise distance measures. A more detailed discussion on the subject is provided by Wang & Bovik (2009).

Neural networks have been applied to metric learning in form of the Siamese architecture (Bromley et al., 1993; Chopra et al., 2005). The learned distance metric is minimized for similar samples and maximized for dissimilar samples using a max margin cost. However, since Siamese networks are trained in a supervised setup, we cannot apply them directly to our problem.

For autoencoders (AE), several attempts at improving on element-wise distances have been proposed within the last year. Ridgeway et al. (2015) apply the structural similarity index as reconstruction metric for grey-scale images. Yan et al. (2015) let the AE output two additional images to learn shape and edge structures more explicitly. Mansimov et al. (2015) append a GAN-based sharpening step to their decoder network. While these AE extensions yield sharper images, they are not built capture high-level structure compared to a deep learning approach.

In contrast to AEs that model the relationship between a dataset sample and a latent representation directly, GANs learn to generate samples indirectly. By optimizing the GAN generator to produce samples that imitate the dataset according to the GAN discriminator, GANs avoid element-wise similarity measures by construction. This is a likely explanation for their ability to produce high-quality images

Figure 5. Using the VAE/GAN model to reconstruct dataset samples with visual attribute vectors added to their latent representations.

as demonstrated by Denton et al. (2015); Radford et al. (2015).

Lately, convolutional networks with upsampling have shown useful for generating images from a latent representation. This has sparked interest in learning image embeddings where semantic relationships can be expressed using simple arithmetic – similar to the suprising results of the *word2vec* model by Mikolov et al. (2013). First, Dosovitskiy et al. (2015) used supervised training to train convolutional network to generate chairs given high-level information about the desired chair. Later, Kulkarni et al. (2015); Yan et al. (2015); Reed et al. (2015) have demonstrated encoder-decoder architectures with disentangled feature representations, but their training schemes rely on supervised information. Radford et al. (2015) inspect the latent space of a GAN after training and find directions corresponding to eyeglasses and smiles. As they rely on pure GANs, however, they cannot encode images making it challenging to explore the latent space.

Our idea of a learned similarity metric is partly motivated by the neural artistic style network of Gatys et al. (2015) who demonstrate the representational power of deep convolutional features. They obtain impressive results by optimizing an image to have similar features as a subject image and similar feature correlations as a style image in a pre-trained convolutional network. In our VAE/GAN model, one could view $\mathcal{L}_{\text{llike}}^{\text{Dis}_l}$ as content and $\mathcal{L}_{\text{GAN}}$ as style. Our style term, though, is not computed from feature correlations but is the error signal from trying to fool the GAN discriminator.

5. Discussion

The problems with element-wise distance metrics are well known in the literature and many attempts have been made at going beyond pixels – typically using hand-engineered measures. Much in the spirit of deep learning, we argue that the similarity measure is yet another component which can be replaced by a learned model capable of capturing high-level structure relevant to the data distribution. In this work, our main contribution is an unsupervised scheme for learning and applying such a distance measure. With the learned distance measure we are able to train an image encoder-decoder network generating images of unprecedented visual fidelity as shown by our experiments. Moreover, we show that our network is able to disentangle factors of variation in the input data distribution and discover visual attributes in the high-level representation of the latent space. In principle, this lets us employ a large set of unlabeled images for training and use a small set of labeled images to discover features in latent space.

We regard our method as an extension of the VAE framework. Though, it must be noted that the high quality of our generated images is due to the combined training of Dec as a both a VAE decoder and a GAN generator. This makes our method more of a hybrid between VAE and GAN, and alternatively, one could view our method more as an extension of GAN where $p(\boldsymbol{z})$ is constrained by an additional network.

It is not obvious that the discriminator network of a GAN provides a useful similarity measure as it is trained for a different task, namely being able to tell generated faces from real faces. However, convolutional features are often surprisingly good for transfer learning, and as we show, good enough in our case to improve on element-wise distances for images. It would be interesting to see if better features in the distance measure would improve the model, e.g. by employing a similarity measure provided by a Siamese network trained on faces. Though in practice, Siamese networks are not a good fit with our method as they require labeled data. Alternatively one could investigate the effect of using a pretrained feedforward network for measuring similarity.

While we believe our preliminary results are convincing, we acknowledge the problem of not having a good measure for evaluating our generative model comparatively. Further work is needed in this direction to verify the applicability of our method beyond generating visually pleasing images.

Acknowledgements

We would like to thank Søren Hauberg, Casper Kaae Sønderby and Lars Maaløe for insightful discussions, Nvidia for donating GPUs used in experiments, and the authors of DeepPy[3] and CUDArray (Larsen, 2014) for the software frameworks used to implement our model.

References

Bengio, Yoshua, Courville, Aaron, and Vincent, Pierre. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

Breuleux, Olivier, Bengio, Yoshua, and Vincent, Pascal. Quickly generating representative samples from an rbm-derived process. *Neural Computation*, 23(8):2058–2073, 2011.

Bromley, Jane, Bentz, James W., Bottou, Léon, Guyon, Isabelle, LeCun, Yann, Moore, Cliff, Säckinger, Eduard, and Shah, Roopak. Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 07 (04):669–688, 1993.

Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 539–546 vol. 1, June 2005.

Denton, Emily L, Chintala, Soumith, Szlam, Arthur, and Fergus, Rob. Deep generative image models using a laplacian pyramid of adversarial networks. In Cortes,

[3]http://github.com/andersbll/deeppy

C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1486–1494. Curran Associates, Inc., 2015.

Dosovitskiy, Alexey, Springenberg, Jost Tobias, and Brox, Thomas. Learning to generate chairs with convolutional neural networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1538–1546, 2015.

Gatys, Leon A., Ecker, Alexander S., and Bethge, Matthias. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Kingma, Diederik P. and Welling, Max. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

Kulkarni, Tejas D., Whitney, Will, Kohli, Pushmeet, and Tenenbaum, Joshua B. Deep convolutional inverse graphics network. *CoRR*, abs/1503.03167, 2015.

Larsen, Anders Boesen Lindbo. CUDArray: CUDA-based NumPy. Technical Report DTU Compute 2014-21, Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2014.

Liu, Ziwei, Luo, Ping, Wang, Xiaogang, and Tang, Xiaoou. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Mansimov, Elman, Parisotto, Emilio, Ba, Lei Jimmy, and Salakhutdinov, Ruslan. Generating images from captions with attention. *CoRR*, abs/1511.02793, 2015.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

Reed, Scott E, Zhang, Yi, Zhang, Yuting, and Lee, Honglak. Deep visual analogy-making. In Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1252–1260. Curran Associates, Inc., 2015.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 1278–1286, 2014.

Ridgeway, Karl, Snell, Jake, Roads, Brett, Zemel, Richard S., and Mozer, Michael C. Learning to generate images with perceptual similarity metrics. *CoRR*, abs/1511.06409, 2015.

Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. *CoRR*, abs/1511.01844, 2015.

Wang, Zhou and Bovik, A.C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1):98–117, Jan 2009. doi: 10.1109/MSP.2008.930649.

Yan, X., Yang, J., Sohn, K., and Lee, H. Attribute2Image: Conditional Image Generation from Visual Attributes. *CoRR*, abs/1512.00570, 2015.