Intelligent Information Management, 2018, 10, 108-113
http://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

Generate Faces Using Ladder Variational Autoencoder with Maximum Mean Discrepancy (MMD)

Haoji Xu

Tianjin Nankai High School, Tianjin, China
Email: haoji.xu@outlook.com

How to cite this paper: Xu, H.J. (2018) Generate Faces Using Ladder Variational Autoencoder with Maximum Mean Discrepancy (MMD). *Intelligent Information Management*, **10**, 108-113.
https://doi.org/10.4236/iim.2018.104009

Received: November 22, 2017
Accepted: July 14, 2018
Published: July 17, 2018

Abstract

Generative Models have been shown to be extremely useful in learning features from unlabeled data. In particular, variational autoencoders are capable of modeling highly complex natural distributions such as images, while extracting natural and human-understandable features without labels. In this paper we combine two highly useful classes of models, variational ladder autoencoders, and MMD variational autoencoders, to model face images. In particular, we show that we can disentangle highly meaningful and interpretable features. Furthermore, we are able to perform arithmetic operations on faces and modify faces to add or remove high level features.

Keywords

Generative Models, Ladder Variational Autoencoders, Facial Recognition

1. Introduction

Generative Models have been highly successful in a wide variety of tasks by generating new observations from an existing probability density function. These models have been highly successful in various tasks such as semi-supervised learning, missing data imputation, and generation of novel data samples.

Variational Autoencoder is a very important class of models in Generative Models [1] [2]. These models map a prior on latent variables to conditional distributions on the input space. Training by maximum likelihood is intractable, so a parametric approximate inference distribution is jointly trained, and surprisingly, jointly training the generative model for maximum likelihood, and the inference distribution to approximate the true posterior is tractable, through a "reparameterization trick" [1]. These models have been highly successful in

modeling complex natural distributions such as natural images. In addition it has been observed that these models can make use of the latent space in a meaningful manner. For example, it can learn to map different regions of the latent variable space into different object classes.

Ladder Variational Autoencoders [3] have been recently proposed to further augment this ability. In particular, it is able to disentangle high level and low level features. It utilizes the assumption that high level features require deeper networks to model, so that latent variables that are connected with the input with deep neural networks learn complicated, high level features while low level features are represented by low level variables.

It has also been observed that the evidence lower bound (ELBO) used in traditional variational autoencoders suffers from uninformative latent feature problem [4] where these models tend to under-use the latent variables. Multiple methods have been proposed to alleviate this [4] [5]. In particular, [5] showed that this problem can be avoided altogether if an MMD loss is used instead of the KL divergence in the original ELBO variational autoencoders.

In this paper we combine these ideas to build a variational ladder autoencoder with MMD loss instead of KL divergence, and utilize this model to analyze of structure and hidden features of human faces. As an application we use this model to perform "arithmetic" operations on faces. For example, we can perform arithmetic operations such as: men with pale skin − men with dark skin + women with dark skin = women with pale skin. The way we do this is by performing arithmetic operations in the feature space, and transform the results back into image space. This can be potentially useful in games and virtual reality where arbitrary features can be added to a face through the above process of analogy. This further demonstrates the effectiveness of our model in learning highly meaningful latent features.

2. Model Definition

2.1. Generative Modeling and Variational Autoencoders

Generative models seek to model a distribution pdata (x) in some input space X. The model is usually a parameterized family of distribution $p_\theta(x)$ trained by maximum likelihood

$$\max \text{Epdata}(x)\left[\log p_\theta(x)\right]\theta$$

Intuitively this encourages the model distribution to place probability mass where pdata is more likely.

Variational autoencoder (Kingma & Welling, 2013; Jimenez Rezende *et al.*, 2014) is an important class of generative models. It models a probability distribution by a prior $p(z)$ on a latent space Z, and a conditional distribution $p(x|z)$ on. Usually $p(z)$ is a fixed simple distribution such as white Gaussian $N(0, I)$, and $p(x|z)$ is parameterized by a deep network with parameters θ, so we denote it as $p_\theta(x|z)$. The model distribution is defined by

$$p_\theta(x) = \int_z p_\theta(x|z)p(z)dz$$

However maximum likelihood training is intractable because $p_\theta(x)$ requires an integration which is very difficult to compute. The solution is by jointly defining an inference distribution $q_\varphi(z|x)$ parameterized by φ to approximate $p_\theta(z|x)$. Jointly training both criteria give the following optimization function, called the evidence lower bound (ELBO)

$$\begin{aligned}\text{LELBO} &= \left(-\text{KL}q_\varphi(z|x)\middle\| p_\theta(z|x)\right) - \text{KL}\left(\text{pdata}(x)\middle\| p_\theta(x)\right) \\ &= -\text{KL}\left(q_\varphi(z|x)\middle\| p(z)\right) + Eq_\varphi(z|x)\left[\log p_\theta(z|x)\right]\end{aligned}$$

where KL denotes the Kullback-Leibler divergence. Intuitively this model achieves its goal by first applying an "encoder" $q_\varphi(z|x)$ to the input, then "decode" the generated latent code by $p_\theta(x|z)$ and compare the generated results with the original data x using the cost function $\log p_\theta(z|x)$.

2.2. Ladder Variational Autoencoder

Ladder variational autoencoders [3] add additional structure into the latent code by adding multiple layers to the model. The model is shown in **Figure 1**. High level latent features are connected with the input through a deep network while low level features are connected through a shallow network. The intuition is that complicated features require deeper networks to model, so that high level latent variables will be used to model the high-level features, and vice versa. This makes it possible to disentangle simple and sophisticated features.

2.3. MMD Regularization

It has been observed that the $\text{KL}\left(q_\varphi(z|x)\middle\| p(z)\right)$ term in ELBO criteria result in under-used latent features (Chen *et al.*, 2016; Zhao *et al.*, 2017a). A solution is to use the MMD ($q(z)$, $p(z)$) instead, which is defined by

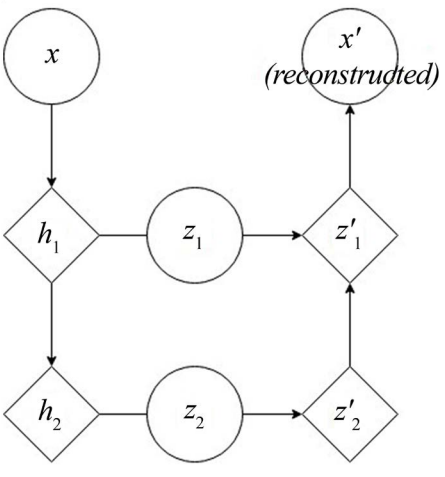

Figure 1. Structure of VLAE (variational ladder autoencoder). Here circles are stochastic variables and diamonds are deterministic variables.

$$\mathrm{MMD}\left(q(z),p(z)\right)=Eq(z),q(z_0)\left[k(z,z_0)\right]+Ep(z),p(z_0)\left[k(z,z_0)\right]$$
$$-2Ep(z),q(z_0)\left[k(z,z0)\right]$$

where $k(z, z_0)$ is a kernel function such as Gaussian. $k(z,z')=e^{-\|z-z'\|_2^2/\sigma^2}$ Intuitively $k(z, z_0)$ measures the distance between z and z_0, and $Ep(z)$, $q(z_0)$ $[k(z, z_0)]$ measures the average distance between samples from distributions $p(z)$ and $q(z_0)$. If two distributions are identical, then the average distance between samples from p, samples from q, and samples from p, q respectively, should all be identical, so MMD distance should be zero. This can be used to replace $\mathrm{KL}\left(q_\varphi(z|x)\|p(z)\right)$ in ELBO VAE to achieve better properties.

2.4. MMD Variational Ladder Autoencoder

We apply MMD regularization to Variational Ladder Autoencoders. In particular, we regularize all the latent features respectively

$$\mathrm{LMMD}-\mathrm{VLAE}=Eq_\varphi(z|x)\left[\log p_\theta(x|z)\right]-\mathrm{MMD}\left(p(z_0),q_\varphi(z_0)\right)$$
$$-MMD\left(p(z_1),q_\varphi(z_1)\right)$$

This combines the advantage of both models and learns meaningful hierarchical features.

3. Experiments

To verify the effective of our method we performed experiments on MNIST and CelebA [6]. We visualize the manifold learned for each dataset, and observe extremely rich disentangled features.

Samples from MNIST are shown in **Figure 2**. We are able to disentangle visual features such as digit width, inclination, digit identity, etc. For example, bottom layer represents style of the stroke, such as the width. Middle layer represents inclination while top layers mostly represent digit identity.

Samples from CelebA are shown in **Figure 3**. We are able to disentangle features such as lighting, hair style, face identity and pose.

4. Arithmetic Operations on Faces

We observed that by adding or subtracting values from latent code, we can modify

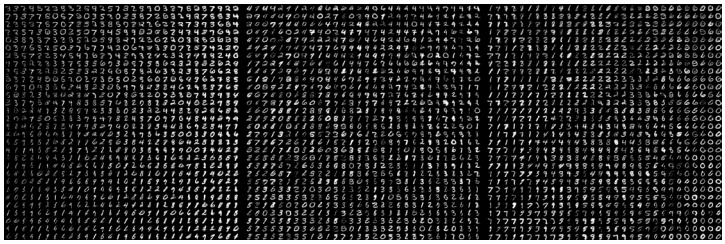

Figure 2. Training results over MNIST after 1 hour on a GTX1080Ti. Each plot is obtained by sampling one layer uniformly in the [−3, 3] range, and other layers randomly. Left: Represents stroke style and width. Middle: Represents digit inclination. Right: Represents digit identity.

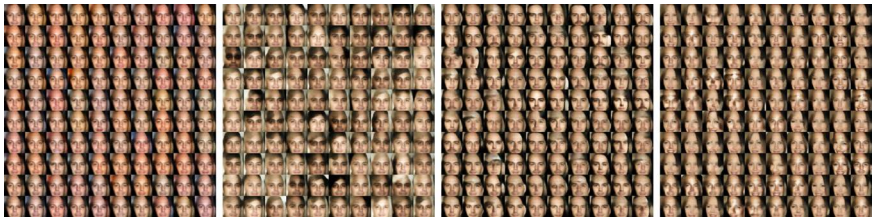

Figure 3. Training results over CelebA after roughly 8 hours on a GTX1080Ti. Left: Represents lighting and white balance. Middle Left: Represents hair color, face color, and minor variations of facial feature. Middle Right: Represents face identity. Right: Represents pose and expression of the face.

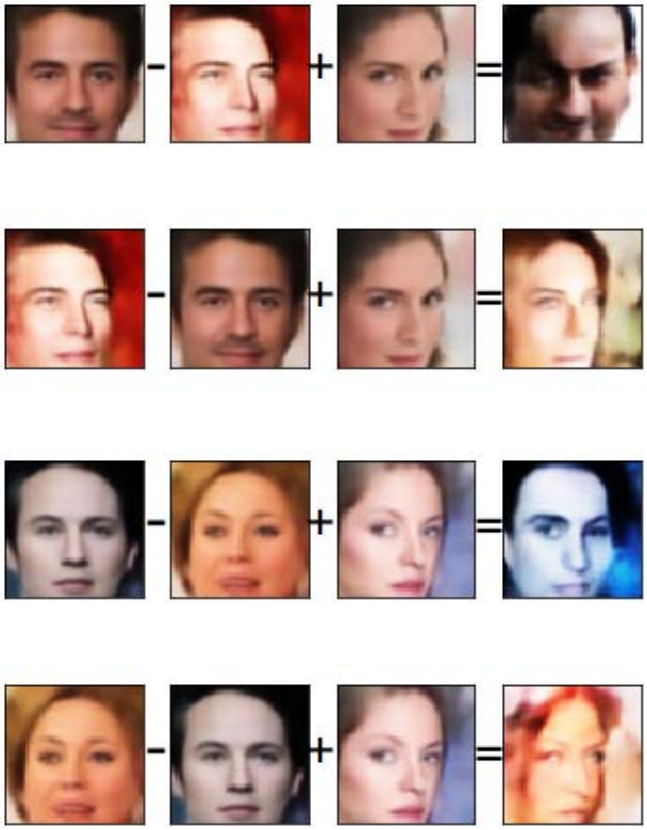

Figure 4. Faces of the fourth column are acquired by subtracting the second column from first column, then by adding the third column to the first column.

certain properties of faces. In addition, we can blend multiple faces together by adding or subtracting latent codes from or to each other.

We observed convincing results from these experiments (as shown in Figure 4). The final result of fourth column has shown various arithmetical properties. For example, faces of colors and brightnesses on all images are explicitly represented by the arithmetic result: the forth images share similar colors and brightnesses with the first and the third images, while these properties differ from the second images. Moreover, more complicated features are also learned and applied, the most specific one being the facial expression.

5. Conclusions/Discussion

In this paper we proposed MMD Variational Ladder Autoencoder and its applications on various tasks, especially on facial recognition and modification on the CelebA dataset. It is capable of disentangling various features of human face and also capable of modifying or blending different faces.

Possible future works might include further discussion on the accuracy and readability of its latent code, its overfitting tendency, and its application on more unlabeled datasets.

References

[1] Kingma, D.P. and Welling, M. (2013) Auto-Encoding Variational Bayes. https://arxiv.org/abs/1312.6114

[2] Rezende, J.D., Mohamed, S. and Wierstra, D. (2014) Stochastic Backpropagation and Approximate Inference in Deep Generative Models. https://arxiv.org/abs/1401.4082

[3] Zhao, S., Song, J. and Ermon, S. (2017) Learning Hierarchical Features from Generative Models. http://arxiv.org/abs/1702.08396

[4] Chen, X., Kingma, D.P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I. and Abbeel, P. (2016) Variational Lossy Autoencoder. https://arxiv.org/abs/1611.02731

[5] Zhao, S., Song, J. and Ermon, S. (2017) InfoVAE: Information Maximizing Variational Autoencoders. http://arxiv.org/abs/1706.02262

[6] Liu, Z., Luo, P., Wang, X. and Tang, X. (2015) Deep learning face attributes in the wild. https://arxiv.org/abs/1411.7766 https://doi.org/10.1109/ICCV.2015.425