Face Rotation via VAE-CapsuleGAN

Zepeng Luo
Department of computer science and technology of the Institute of Information Engineering North China Polytechnic University Tangshan Hebei, China North China University of Science and Technology, China
alicture95@gmail.com

Abstract. High-quality image reconstruction and generating the new image by learning the distribution from the image are very popular in computer vision field. Nevertheless, some promising results have been reported in the literature recently. In this work, we propose a VAE-CapsuleGAN network which can generate high-quality face image with modified rotation angle. We show promising results in our experiment on the CAS-PEAL-R1 face dataset.

1. Introduction

When looking at a photo of an old friend which is taken from the side. And you would like to know what she looks like in the front view. Or you may be wondering to know what you look like from another angle. Maybe the next generation image editing software can do these magical jobs for you.

Thoughts along these lines motivated us to start looking into this problem. To accomplish this task, we need a model which can generate the high-quality image based on an encoded source image. Not like a traditional GAN, our model using variational auto-encoder as the generator and using the CapsNet to build our discriminator. An overview of our network as shown in figure 1.

Here is what we do, we choose face images which were taken in the same vertical angle, and we cropped these face images by the given eye coordinates in the dataset. We train our model by given source image and desired rotation angle vector to generate a new image as close to the target image as possible which was rotated from the source image. Rotation angle range from 45 degrees to -45 degrees, with a total of 7 positions.

To demonstrate the effectiveness of our method on the above task, we evaluate our network compared with a 2 stage Autoencoder network [1] and 2 stage VAE network [2] on CAS-PEAL-R1 dataset [3].

In a nutshell, our model attempts to satisfy two objectives: (i) -generate new face image based on source image by input the desired angle vector (ii)-make the generate image as clear as possible.
2. Related work
Recently, there are many papers proposed several networks did a great job in generating image task. There are also some old works inspired me, such as [1], in their work, they proposed a 2 stage convolutional auto-encoder network. [2] Proposed variational auto-encoder, and VAE already shows promise in generating complicate data.

3. Preliminaries

3.1. Variational Auto-encoder
[2] Introduced variational autoencoder as a very efficient approach to unsupervised learning of complicated distributions. A VAE consists of two networks that encode a data sample \( x \) to a latent representation \( z \) and decode the latent representation back to data space. The VAE loss is minus the sum of the expected log likelihood (the reconstruction error) and a prior regularization term:

\[
\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + D_{\text{KL}}(q(z|x)||p(z))
\]  

(1)

3.2. Generative Adversarial Network
[4] Introduced GAN as a new efficient generative model which learning transformations from data points sampled from prior distribution to the data distribution. A Generative Adversarial Network contained two parts: a generator network and a discriminator network. To train a GAN we arrange an adversarial game between the generator network and the discriminator network. The generator network tries to fool the discriminator network while the discriminator network tries to classify generated images from the real. To achieve this goal the author proposed the objective as shown in Equation 2.

\[
\min_G \max_D \mathcal{V}(D,G) = \log(D(x)) + \log(1-D(G(z)))
\]  

(2)

3.3. Capsule Net
Capsule introduced by [5] was designed to better model hierarchical relationships inside of internal knowledge representation of a neural network. Compared with convolutional neural networks, the most significant difference is that capsule network can encapsulate all important information about the state of the feature they are detecting in vector form. Thus, the capsule network can learn spatial relationships of different components. In CNNs, the max pooling is often used between convolution layers to achieve viewpoint invariance in the activities of neurons. But using max pooling also loses valuable information
and also does not encode relative spatial relationships between features. Loss function proposed in Capsule Network original paper as shown in Equation 3:

\[ L_c = T_c \max (0, m^+ - \|v\|^2) + \lambda (1 - T_c) \max (0, \|v\| - m)^2 \]  

(3)

In the loss function formula, the correct label determines the value of \( T_c \): it is 1 if the correct label and 0 otherwise.

4. VAE-CapsuleGAN

Our proposed model using VAE as the generator. We have two inputs to our network, source image and the face rotation attribute vectors which has been one-hot encoded. Similar to [1], we fuse the feature maps from the encoding of the input image and the attribute vector, then concatenate them before further encoding process. The remain part of the network inspired by [6]'s research on the VAEGAN. Instead of using batch normalization, we chose to apply HeNormal in convolutional layers.

| Table 1. Encoder and decoder |
|-------------------------------|
| Encoder                        | Decoder                        |
| 3x3, 1 Conv, HeNormal (input_img) | 4·4·256, fully-connected      |
| 3x3, 64 Conv, HeNormal (1)     | 3x3, 256 DeConv, HeNormal      |
| 512, fully-connected (input_pose) | 3x3, 128 DeConv, HeNormal      |
| 1792, fully-connected (2)      | 3x3, 64 DeConv, HeNormal       |
| Concatenate (1, 2)             | 3x3, 1 DeConv, HeNormal        |
| 3x3, 128 Conv, HeNormal        |                               |
| 3x3, 128 Conv, HeNormal        |                               |
| 2048, fully-connected          |                               |

Table 1: Architectures for the encoder and the decoder that comprise the generator of our VAE-CapsuleGAN.

| Table 2. Architecture summary of our proposed VAE-CapsuleGAN |
|---------------------------------------------------------------|
| Layer (type) | Output Shape | Param # | Connected to |
|----------------|---------------|---------|--------------|
| Input_img (InputLayer) | (None, 64, 64, 1) | 0 |                 |
| Input_pose (InputLayer) | (None, 7) | 0 |                 |
| Model_enc (Model) | (None, 64) | 1415818 | Input_img [0] [0] | Input_pose [0] [0] |
| Model_dec (Model) | (None, 64, 64, 1) | 1226753 | Model_enc [1] [0] |
| Model_discriminator (Model) | (None, 1) | 22763377 | Model_dec [1] [0] |

Our proposed model adopts CapsNet [7] as our discriminator. Inspired by [8]’s Capsule GAN architecture, the objective of ours VAE-CapsuleGAN can be formulated in Equation 4.

\[ \min_G \max_D V(G, D) = -L_c(D(x), T=1) + L_c(D(G(z)), T=1) \]  

(4)

5. Experiment

We conduct a series of experiments to evaluate the performance of our network as described below, in which we compared ours network with some of the others both qualitatively and quantitatively. We implement Ghodrati, A. et al. (2015)’s network based on their code which is given in their paper. And
our 2 stage VAE network using open source keras implementation. Our proposed VAE-CapsuleGAN model is implemented using open sourced CapsNet-Keras package.

5.1. Data
We choose to use CAS-PEAL-R1 dataset to conduct our experiments. Before we start our experiments, we need to crop face images first. All faces were cropped based on eye coordinates [9], then resized to $60 \times 60$ pixels. From 7273 images, we randomly chosen 90% of them for training, and 10% for testing.

![Target Images](image1)

![VAE-CapsuleGAN](image2)

![2 stage AE](image3)

![2 stage VAE](image4)

Figure 2. 3 sets of 25 generated CAS-PEAL-R1 images from different network and target images from CAS-PEAL-R1.

5.2. Visual Quality of Generated Images
In figure2, we demonstrate 25 randomly chosen images generated by 3 different models. Figures 2c show images generated by 2 stage Auto-encoder network, and figures 2d show images generated by 2 stage variational auto-encoder image. Compare these two sets of generated images to the target images as shown in figures 2a, 2 stage Auto-encoder network did a good job in reconstruction. But 2 stage Auto-encoder network lost some face details in a few pictures due to the input image is not cropped well. As we can see in the figures 2b, images generated by VAE-CapsuleGAN generated some very high-quality face images even when the input image is not cropped nicely. And compare figures 2d and figures 2b, those faces generated by the VAE-CapsuleGAN network look cleaner and crisper than those generated using plain VAE network. The generated image of ours model somehow looked more beautiful than the others.

Table 3. Mse result of 2 stage ae, 2 stage vae and the proposed VAE-CapsuleGAN

| Model          | n=50   | n=100  | n=1000 |
|----------------|--------|--------|--------|
| 2 stage AE     | 0.01303| 0.01231| 0.01244|
| 2 stage VAE    | 0.02139| 0.01808| 0.01993|
| VAE-CapsuleGAN | 0.01630| 0.01705| 0.01811|
To compare our proposed model with others quantitatively, we choose to use mean square error to measure the reconstruction error between input images and target ones. As shown in table 3 we randomly chose n(50,100,1000) samples from the dataset as the input image, then compute MSEs for our proposed VAE-CapsuleGAN, 2-stage AE model and 2-stage VAE model. The reconstruction error of our proposed model is little bit higher than the 2 stage AE model, but compare to the 2 stage VAE model, ours is better, keep in mind, our model’s generator using the same architecture used in the VAE model.

6. Conclusion
In this work, our goal is to generate images with modified attributes while maintaining as much as possible. Our proposed VAE-CapsuleGAN generated significant high-quality face images, but on the quantitively quality, our proposed network did not achieve a good score in our experiment. For the future work, we would like to extend our model to achieve much lower reconstruction error, and try to use CapsNet instead of CNNs in the generator.

References
[1] A. Ghodrati, X. Jia, M. Pedersoli, and T. Tuytelaars, “Towards Automatic Image Editing: Learning to See another You,” Artif. Neural Networks–ICANN, 2015.
[2] A. Santerne et al., “Auto-Encoding Variational Bayes,” Astron. Astrophys., 2016.
[3] W. Gao et al., “The CAS-PEAL large-scale chinese face database and baseline evaluations,” IEEE Trans. Syst. Man, Cybern. Part ASystems Humans, 2008.
[4] I. Goodfellow et al., “Generative Adversarial Nets,” Adv. Neural Inf. Process. Syst. 27, 2014.
[5] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming auto-encoders,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011.
[6] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” Dec. 2015.
[7] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” Oct. 2017.
[8] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, “CapsuleGAN: Generative Adversarial Capsule Network,” Feb. 2018.
[9] L. El Shafey, C. McCool, R. Wallace, and Ś. Marcel, “A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., 2013.