Preliminary Project Report

Abstract

GANs are generative models where we use adversarial process and simultaneously train two models, a generative model and a discriminative model. The generative model G captures the data distribution, and the discriminative model D estimates the probability that a sample came from the training data rather than G. It uses implicit density to generate new samples, unlike traditional methods such as AutoEncoders which minimize explicit density of training samples and produces unrealistic blurry images.
In this project, we aim to use GANs for face aging also known as age synthesis, where we render a face image with natural aging and rejuvenating effects on the individual face. We further discuss various versions of GANs such as Conditional GAN, where we feed the input with a conditioning variable to condition on to both generator and discriminator, Age-cGANs where we focus on identity preserving face aging.

1. Introduction

The applications of face aging include, finding lost children and cross-age face recognition. Traditional face aging are roughly of two types: prototyping and modeling. Prototyping approach of face aging are simple and fast - they estimate average faces within predefined age groups. Since they are based on general rules, personalised information is discarded, producing unrealistic images. On the other hand, modelling approaches using parametric models to simulate aging process. To train a model, we require various age sequences of the same person which is expensive.

To solve this problem, we will prefer generative models which generates new samples by estimating the probability density of given samples. Generative models are of two kinds - explicit and implicit density models. Explicit density models try to maximize the likelihood estimation of training data to generate new samples but the images produced are of low quality. Implicit density models, such as GANs, use game theoretic approach to generate realistic samples.

There are a different types of GANs, for the purpose of face aging we are going to discuss about Conditional GANs where we use target age group as conditional input.

2. Literature Review

2.1. Generative Models

Deep generative models (DGM) are neural networks with many hidden layers trained to approximate complicated, high-dimensional probability distributions using a large number of samples. When trained successfully, we can use the DGMs to estimate the likelihood of each observation and create new samples from the underlying distribution.

As of today, the three most popular generative models are Fully Visible Belief Networks (FVBN), Variational AutoEncoder and GAN.

2.2. Fully Visible Belief Networks (FVBN)

It is an explicit density model. It uses chain rule to decompose the likelihood of an image X into product of 1D distribution.

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1}) \tag{1}$$

Then, we maximize the likelihood of the training data.

2.2.1 PixelRNN

PixelRNN generates image pixels starting from a corner. It depends on previous pixels which are modelled using an RNN (LSTM). Since we are sequentially generating pixels using RNN, it is very slow.

2.2.2 PixelCNN

PixelCNN also generates image pixels starting from a corner. Its dependency on previous pixels is now modelled using a CNN over context region. Since we can use parallel convolutions, training is faster than PixelRNN. But, generation is still slow as we are proceeding sequentially.

2.3. Variational Autoencoders

In variational autoencoders (VAE), we use interactable density function and optimize the lower bound of the likelihood of the training data. We use an encoder network to

model $q_\phi(z|x)$ where, z is a latent vector and x is an input image. Further, we model $p_\theta(z|x)$ using a decoder network. Following is the objective function for the model,

$$\mathcal{L}(x^{(i)}, \theta, \phi) = \mathbf{E}\left[\log p_\theta(x^{(i)}|z)\right] - D_{KL}\left(q_\phi\left(z|x^{(i)}\right)||p_\theta(z)\right) \quad (2)$$

The variational lower bound for the likelihood,

$$\log p_\theta(x^{(i)}) \geq \mathcal{L}(x^{(i)}, \theta, \phi) \quad (3)$$

We train our model by maximizing the lower bound

$$\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_{i=1}^{N} \mathcal{L}(x^{(i)}, \theta, \phi) \quad (4)$$

Pros:

- Principled approach to generative models.
- Allows inference of $q(z|x)$ which can be a useful feature representation for other tasks.

Cons:

- It maximizes the lower bound of likelihood. But, the evaluation is not as good as PixelRNN/PixelCNN.
- The generated samples are blurry and of low quality.

2.4. Generative Adversarial Network

Generative Adversarial models uses a game-theoretic approach, where we train two networks - generator network and discriminator network.
The generator network tries to fool the discriminator by generating real looking images. The discriminator network tries to distinguish between real and fake images.
The generator network models a distribution of real looking images from a random noise. And, the discriminator network outputs the likelihood in $(0, 1)$ of real image. We train both the networks jointly using a minimax objective function.

$$\min_{\theta_g}\max_{\theta_d}[\mathbb{E}_{x\sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z\sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))] \quad (5)$$

where, $D_{\theta_d}(x)$ is the discriminator output for the real data x and $D_{\theta_d}(G_{\theta_g}(z))$ is the discriminator output for the generated fake data $G(z)$.

for number of training iterations **do**
  for k steps **do**
    • Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
    • Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
    • Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_{\theta_d}(x^{(i)}) + \log(1 - D_{\theta_d}(G_{\theta_g}(z^{(i)})))\right]$$

  end for
  • Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  • Update the generator by ascending its stochastic gradient (improved objective):

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D_{\theta_d}(G_{\theta_g}(z^{(i)})))$$

end for

Figure 1. Training pipeline for model

2.5. Conditional GAN

Generative Adversarial Networks can be extended to a conditional probabilistic model if, a conditional variable y is fed into both generator and discriminator as an input layer. In the generator, the prior input noise $p_z(z)$ and y are combined in joint hidden representation. In the discriminator, x and y are presented as input to a discriminative function.

$$\min_{\theta_g}\max_{\theta_d}[\mathbb{E}_{x\sim p_{data}} \log D_{\theta_d}(x|y) + \mathbb{E}_{z\sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z|y)))] \quad (6)$$

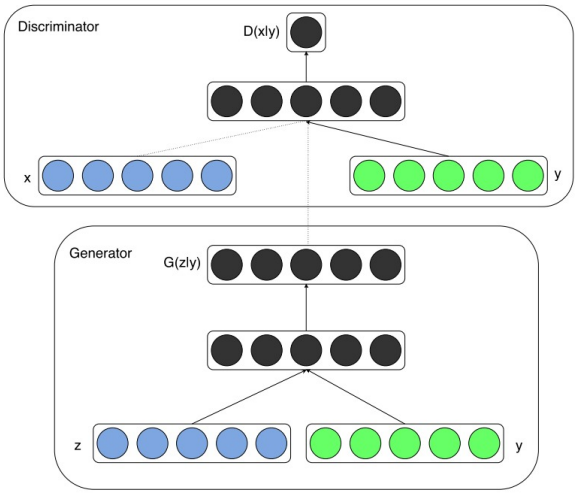

Figure 2. Conditional adversarial network

2.6. Age-cGAN

In Age-cGAN, we use a novel latent vector optimization approach which allows Age-cGAN to reconstruct an input face image preserving the original person's identity. The

Age-cGAN has multiple stages of training. It has 4 networks which get trained in 3 stages.

1. Conditional GAN training: Training of generator and discriminator networks
2. Initial latent vector approximation: Training of encoder network
3. Latent vector optimization: Training of facial recognition network

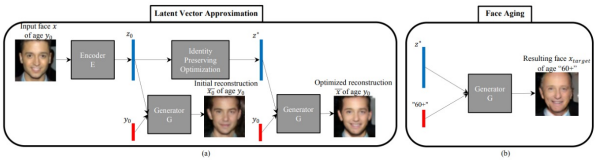

Figure 3. Face aging method using Age-cGAN

2.6.1 Latent Vector Optimization

The ultimate goal of the face aging task is to change the age of the person without changing the identity. Use of initial latent approximations results in loss of identity of the original person in about 50% of the cases. Therefore, initial latent approximations must be improved.
Pixel-wise latent vector optimization can be used for the above purpose. But, it has 2 clear downsides. It increases the blurriness of reconstruction and it focuses on unnecessary details of input face images, which have nothing to do with the person's identity. Instead, identity preserving latent vector approach is used where the difference between the identities in the original and the reconstructed images is expressed as the Euclidean distance between the corresponding embeddings. The objective is to minimize this distance.

$$z^*_{IP} = \arg\min_z \|FR(x) - FR(\bar{x})\|_{L_2} \tag{7}$$

This face aging method can be used for synthetic augmentation of face dataset and for improving the robustness of face recognition solution in cross-age scenarios.

References

[1] Aäron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu (2016) "Pixel Recurrent Neural Networks"

[2] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, Koray Kavukcuoglu (2016) "Conditional Image Generation with PixelCNN Decoders"

[3] Diederik P. Kingma, Max Welling (2014) "Auto-Encoding Variational Bayes"

[4] Ian Goodfellow et al. (2014) "Generative Adversarial Nets", NIPS 2014

[5] Mehdi Mirza, Simon Osindero (2014) "Conditional Generative Adversarial Nets"

[6] Grigory Antipov, Moez Baccouche, Jean-Luc Dugelay (2017) "Face Aging with Conditional Generative Adversarial Networks"