Face Aging With GAN

Siyao Sun, Alex Wang
{siyaosun, cwang16}@stanford.edu

Abstract

In this project we developed a conditional deep neural network for face age progression generation. We surveyed the many datasets to select both diverse and domain specific training data for our intended class. We tuned hyper-parameters and loss function to achieve stability and photorealism. We evaluated our results using a suite of both qualitative and quantitative metrics. Our final results were considered convincing in age progression according to our surveys.

1. Introduction

In our project, we will explore and implement face progression over different ages. People are often interested to know what they look like a few years or decades down the road. We would like to see how generative modeling can help predict what a face will look like in 10, 20 years given a base face image at a younger age (pre-teen or early teens). One significant use case is fighting human trafficking and reuniting families. This is a traumatic experience for families and leaves many without closure. Because the human face goes through marked physical transformation from childhood through adulthood, it's hard to recognize a lost child 10/15 years after abduction. Having a way to model future face given a childhood picture is extremely instrumental to continued search and rescue effort. For our use case, we will only consider age progression rather than age regression. Using Generative Adversarial Networks, we seek to generate older versions of oneself while preserving the identity of the individual. We are able to build conditional deep convolutional neural networks that achieve convincing results.

2. Related Work

2.1 Face Morphing Prior to GAN

Prior to GAN, two approaches are often used for implementing face age progression: prototype based or modelling based. Modelling based approach identifies key facial features (eyes, nose, jaw) and tracks temporal changes (wrinkles, muscle, color) in those features. This approach requires age labeled training data per individual over a long time period, which is difficult to find at scale. It's also computationally expensive. Prototype based method creates an average face based on many images within an age group and uses this to style transfer faces from one age group to another. This causes age-progressed images to lose personal traits and could average out features like wrinkles. A third way uses recurrent neural network to transform face smoothly across ages by modelling intermediate transition states. It's able to generate finer grain images, but still requires age labeled data for the same person over the years.

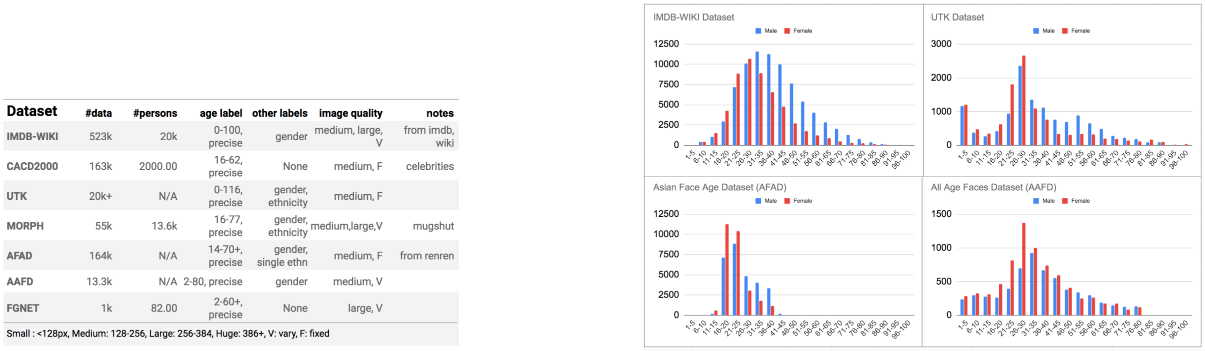
2.2 Face Morphing and GAN

Generative Adversarial Networks[12] involve a discriminator and generator competing with each other in a minmax game. The generator starts with a latent vector Z and generates an image which discriminator gives feedback on. DCGAN[7] showed that GAN can be successfully applied to generate indoor scenes and human faces. StyleGan[8] made significant extension to basic GAN network to progressively generate high resolution (1024 x 1024) images from lower resolution ones. GAN has been applied in face aging image generation as well, it has the advantage of allowing aging image generation without paired data of the same person. cGAN[9] introduced "Identity-Preserving" latent vector optimization approach to ensure resemblance of generate face w.r.t original face. Age Progression/Regression by Conditional Adversarial Autoencoder[10] introduced using autoencoder in the generator network and imposed 2 discriminators on the encoder and generator to enforce better latent vector distribution and more realistic images. Pyramid GAN[11] introduced pyramidal adversarial discriminator at multiple scales, which simulates the aging effects in a finer manner. It also introduced a set of methods for evaluating image fidelity and aging accuracy.

3 Dataset

3.1 Data source & characteristics

Our model requires a lot of age labelled images as our dataset for training. Some widely cited datasets IMDB-WIKI, UTK and CACD2000. IMDB-WIKI and CACD2000 datasets are both datasets of celebrities. UTK dataset has both celebrities and ordinary people. MORPH II is a mugshot dataset but isn't free ($99). All of these datasets are diverse in terms of ethnicity. In addition, AFAD (scrapped from renren.com, an Asian social network) and AAFD also has predominantly Asian faces. They can be used to train/test our algorithm on a single demographic. All these datasets have been cropped and face aligned, some still needed more preprocessing.



Dataset	#data	#persons	age label	other labels	image quality	notes
IMDB-WIKI	523k	20k	0-100, precise	gender	medium, large, V	from imdb, wiki
CACD2000	163k	2000.00	16-62, precise	None	medium, F	celebrities
UTK	20k+	N/A	0-116, precise	gender, ethnicity	medium, F	
MORPH	55k	13.6k	16-77, precise	gender, ethnicity	medium,large,V	mugshut
AFAD	164k	N/A	14-70+, precise	gender, single ethn	medium, F	from renren
AAFD	13.3k	N/A	2-80, precise	gender	medium, V	
FGNET	1k	82.00	2-60+, precise	None	large, V	

Small : <128px, Medium: 128-256, Large: 256-384, Huge: 386+, V: vary, F: fixed

Figure 1. Distribution and demographics of various datasets that we used in this project.

3.2 Dataset preprocessing

Although the aforementioned datasets abundance of data from twenties to forties years old, we found data in our key interest age groups (pre-teen or early teen) quite lacking. This is probably due to privacy concerns. IMDB and CACD have very few images below 14 years old; AFAD, although rich in post 15-yr images lack images under. A big part of our data preprocessing has been to combine these datasets. We further cropped IMDB and AAFD datasets to match the face proportion of UTK and AFAD respectively. In addition, because the data size in younger age is limited, it's more sensitive to noise. We manually removed abnormal looking data such as adult mislabeled as children, wrong face orientation, masked face etc.

3.3 Dataset composition

We grouped images into age buckets and selected 16-20 and 26-30 as the target age group. We chose 6-10 years old as the source age group. For the source age group, we selected 1600 from combined IMDB and UTK combined set, and 800 from AFAD/AAFD combined dataset. We used 6000 images for 16-20 age groups and 10000 images for 26-30 age groups. All the age groups are balanced in terms of gender. We did 80%, 20% training/validation split, but did not use train/validate/test split because in lecture it's said that for generative models' people can skip test set, and we think our generative model fit the bill.

4. Methods

4.1 Model Architecture

In this project, we use a deep convolutional neural network to generate face aging patterns. Our architecture is shown below. Our model has a generator/discriminator component as well as identity preserving component. We used FaceNet pretrained on VGGFace2 dataset to extract facial features as encoding and compared encodings faces. Unlike pixel-wise comparison, which will force pixel to pixel similarity, encoding comparison allows some variation to exist in output image (natural due to age progression) while preserving more prominent features. We did not incorporate age encoding as input to generator or discriminator, but instead rely on training on images in specific age groups for the particular generator. This "Group GAN"[4] approach has been used in other literature as well. It's simpler architecturally but fits our problem domain well because we are only interested in a few age groups. Our input is a young face encoding concatenated with some noise vector Z to stabilize training. This latent vector is applied several layers of deconvolution to generate a face, which is critiqued by the discriminator. At test time, we use our generator model to make inference based on test data and evaluate our results.

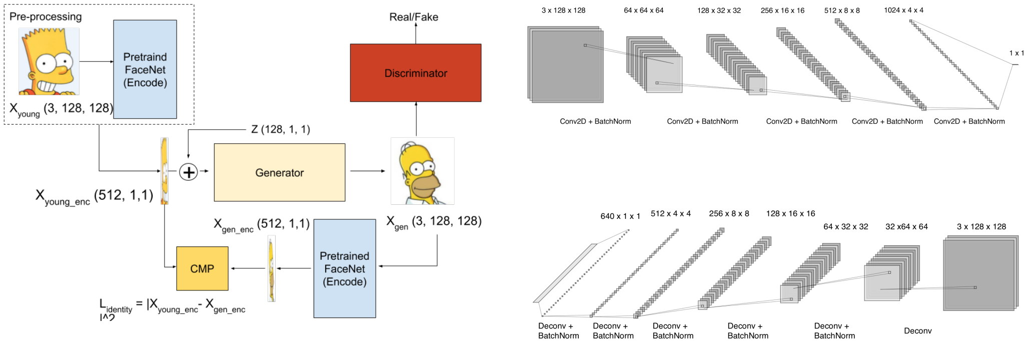


Figure 2. On the left: overview of architecture of age-GAN model. On the right: dimensions of generator and discriminator networks.

4.2 Loss Functions

Our baseline loss function is the one introduced in class. This measures Kullback–Leibler (KL) divergence between data distributions. We would like the distribution of our generator network to approach that of real images to achieve photorealism, thus it's necessary to minimize the divergences. In addition, we incorporated other loss functions into our network. WGAN paper introduced Wasserstein distance (i.e. Earth Mover's distance) which seeks to make two distributions similar while moving the least amount of distance. Wasserstein loss is better than KL divergence because even when two distributions do not share overlaps, it still provides meaningful representation of distance between the distributions, allowing the network to learn. The key to Wasserstein distance is Lipschitz constraint, which ensures that function is smooth. Weight clipping is used as an approximation to achieve that constraint. In WGAN-GP, gradient penalty penalty is used to enforce the Lipschitz constraint as well. Both WGAN and WGAN-GP[5] are shown to be able to facilitate GAN training and reduce mode collapse. The loss functions are listed below.

AgeGan Loss

Discriminator Loss:

$$L^{(D)} = y_{real} \bullet log(D(x_{old})) + (1 - y_{gen}) \bullet log(1 - D(G(x_{young})))$$

Generator Loss (second term represents $L_{identity}$):

$$L^{(G)} = log(1 - D(G(x_{young}))) + \lambda * \left|\left|Enc(G(x_{young})) - Enc(x_{young})\right|\right|_2$$

AgeGan Loss + WGAN Loss

Discriminator Loss:

$$L^{(D)} = f(x_{old}) - f(G(x_{young}))$$

Generator Loss:

$$L^{(G)} = f(G(x_{young})) + \lambda * \left|\left|Enc(G(x_{young})) - Enc(x_{young})\right|\right|_2$$

AgeGan Loss + WGAN-GP Loss

Discriminator Loss:

$$L^{(D)} = f(x_{old}) - f(G(x_{young})) + \tau * (\left|\left|\nabla f(t * x_{young} + (1 - t) * x_{old})\right|\right|_2 - 1)^2$$

Generator Loss:

$$L^{(G)} = f(G(x_{young})) + \lambda * \left|\left|Enc(G(x_{young})) - Enc(x_{young})\right|\right|_2$$

5. Experiment

5.1 Artifacts Tuning and Resize Convolution

In the early iterations of our training, we observed various "artifact" patterns or "checkerboard" effects on our generator result. These artifacts are a natural result of applying layers of deconvolution which resulted in the stacking of pixels on specific columns on the image.

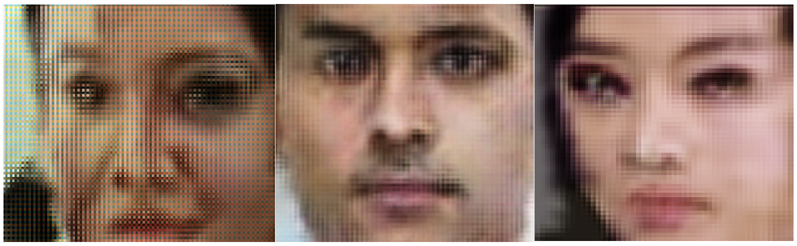


Figure 3. Illustration of hyper parameter tuning in order to alleviate the checkerboard effect. From left to right: without any artifact tuning, tuning with stride and kernel size, using resize convolution architecture instead of deconvolution.

Although these artifacts do not affect the training loss, they do affect the visual inspection from human eyes and will generate unnatural results. Our first approach is to re-tune the kernel and stride sizes of our deconvolution layers. "Checkerboard" effect tends to happen when you have kernel size that's not divisible by the stride size, hence we tried various divisible pairs such as (5, 1), (4, 2), (6, 2) and etc. The result is better image quality with less artifacts.

Other than tuning the kernel and stride size, we also explored with resize convolution layers for the generator instead of deconvolution. The resize convolution layer we tried includes two components: 1. Nearest neighbor interpolation for up sampling and 2. Same size convolution to reduce the output channel while keeping the image dimensions. Resize convolution is naturally immune to the checkerboard effect as the layer does not project pixels in a congested area. We did observe the artifacts completely disappear in our experiment. However, we ended up not going with the resize convolution as it performed worse in terms of other generated image quality metrics.

5.2 Mode Collapsing

Mode collapsing is also another typical effect that has been observed through GAN training, in which the generator only produces a limited variety of samples. To counter this effect, we tried adding various dimensions of random pixels as an input to our generator. Adding too large of random noise size in generator training will make generator less stable and the generator training loss tends to increase drastically, while adding too few random noise sizes to the input will have less counter effect to mode collapsing. We found the (512, 128) split between FaceNet encoding and random noise size to be ideal for our GAN training process. In addition, WGAN and WGAN-GP also helped reduce mode collapse.

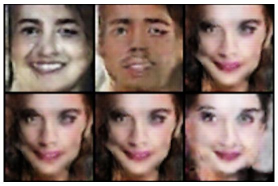


Figure 4. A typical mode collapsing training batch. Many faces ended up being very similar from visual inspection.

5.3 Learning Rate Decay and Encoding Loss Weight

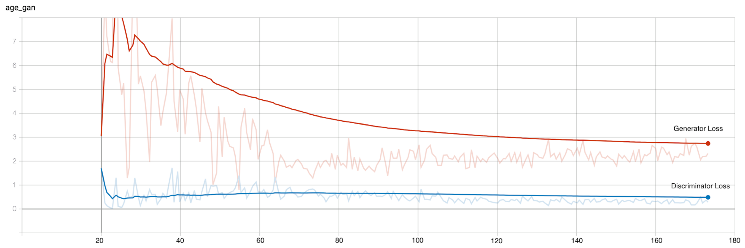


Figure 5. Using step decay learning rate in age GAN training. Both generator and discriminator loss eventually converge.

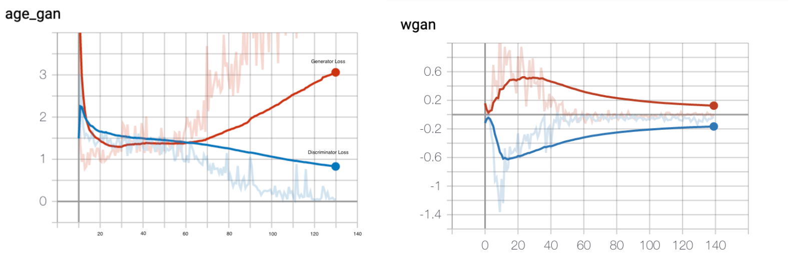


Figure 6. On the left, no learning rate decay is applied, generator eventually diverges during training. On the right, using exponential learning rate for WGAN raining. Both generator and discriminator loss eventually converge.

GAN training is an unstable process. From training age GAN and age GAN + WGAN we found that the generator can easily diverge after some amounts of epochs and pictures will "explode" eventually and never converge. We were able to solve this problem using learning rate decay algorithms and the GAN training process becomes more stable which can benefit higher epochs in training. Another issue is not knowing when to stop training, as loss function can gradually rise again beyond many epochs, we did early-stopping by observing the trends in Tensorboard.

6. Evaluation

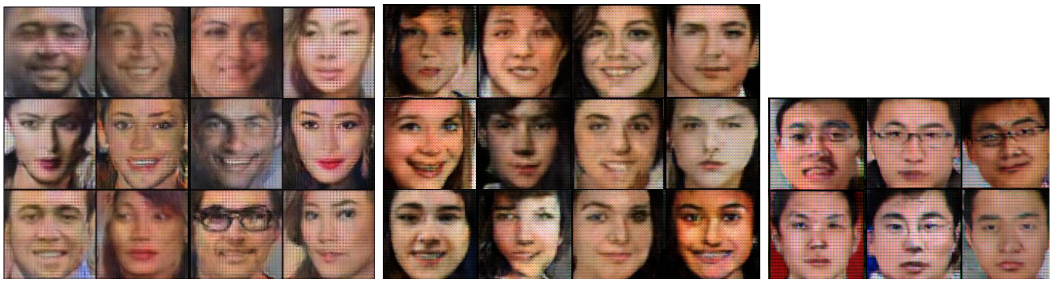


Figure 8. Left two: training results for two different age groups, 26-30, 16-20 respectively. Right: Training results from using a specific demographic dataset.

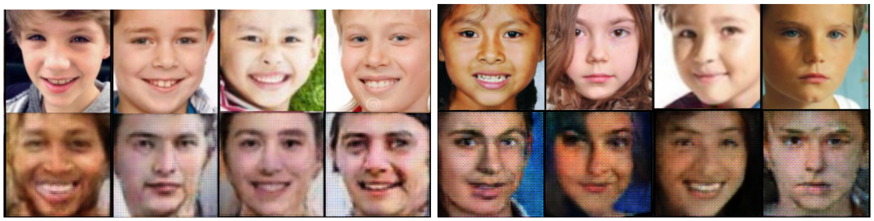
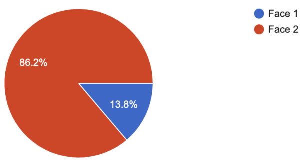


Figure 9. Inference results for two different age groups, 26-30, 16-20 respectively. Using age group 1-5 pictures as input.

We selected a suite of evaluation, both qualitative and quantitative to measure our results. Quantitative evaluation can be automated and done at scale and can be used to evaluate a large number of generated images quickly. Qualitative evaluation aligns better with real human perception and is the true north.

For quantitative evaluation, we used FID score and age estimation[14]. Inception scores were deemed inappropriate as it evaluates class (bird, ship, TV) diversity of many classes, and is not suited for human face. FID score tells how real the generated images (of faces) are compared to real images (of faces). FID uses a deep network to extract facial features and compares ground truth image against target image is through some matrix operation. This core informs image quality, age estimator verifies quality of age effect generation, and gender estimator verifies basic identity information is preserved. There's also a semi quantitative measure which is doing survey and getting mean opinion score. We created a survey of 8 questions that asks respondents two types of answer: 1. Does our model generate more realistic human face image than baseline model? 2. Does our model generate convincing progression given target age.

We conducted a survey with 424 images and here is the mean result:



Total Participants	Total Images Surveyed	Mean Opinion Score (scale 1 -5)
53	424	3.73

For FID score and age gender evaluation, we used online implementations with pretrained models. Perhaps because of the training set, age and gender estimation was not particularly accurate, and the model had almost random gender judgement and age had overestimation by 5 years for 26-30 age on ground truth data. For FID score, we compared 300 generated images from 6-10 source target group against 1000 real images. We had an FID score of 340 for AgeGan model and 360 for WGAN model. Surprisingly AgeGan had better image quality. Because of the difference in training data, our FID score is very different from results in published literature. For sanity check, we calculated FID score on between real images and got ~60, and FID score between cat/dogs against human face and got 1000+. Our generated FID score is directionally correct.

7. Conclusion

We developed and evaluated multiple GAN models for generating face progression. We started out with Conditional Deep Convolutional GAN architecture and modified several loss/divergence functions to improve model stability and enhance image generation result. We modified model architecture to account for issues in GAN such as checkerboard artifacts, and fine-tuned hyperparameters such as learning rate, weight, hidden layer size, D vs G training time ratio. We researched and identified many datasets to power our training and used preprocessing to decrease noise. We adopted a suite of evaluation metrics to gauge the quality of our network and provide guidance for improvement. Ultimately, we were able to generate good quality age progressed images. Hopefully this technique will mature into a powerful tool to help find lost people and combat human trafficking

8. Future work

Improve evaluation. Better model to get high resolution images. Maybe gather higher quality labeled data through web scraping and video image extraction.

9. Contributions & Acknowledgements

Both members of the team contributed equally in the project. We want to thank all the TAs for their support in guiding us through this project. Especially Mohamed, Hao, Vineet for all the office hours Q&As.

10. Code Repository

https://github.com/siyaofd/GANSTA

11. References

[1] A. Lanitis, C. J. Taylor, and C. J. Taylor, T. F. Cootes. 2002. Toward Automatic Simulation of Aging Effects on Face Images. IEEE Trans. Pattern Anal. Mach. Intell. 24, 4 (April 2002), 442-455. Jinli Suo, Song-Chun Zhu, Shiguang Shan, and Xilin Chen. 2010. A Compositional and Dynamic Model for Face Aging. IEEE Trans. Pattern Anal. Mach. Intell. 32, 3 (March 2010), 385-401.

[2] Burt, D. M. and David I. Perrett. “Perception of age in adult Caucasian male faces: computer graphic manipulation of shape and colour information.” *Proceedings. Biological sciences* 259 1355 (1995): 137-43 .

[3] Bernard Tiddeman, Michael Burt, and David Perrett. 2001. Prototyping and Transforming Facial Textures for Perception Research. IEEE Comput. Graph. Appl. 21, 5 (September 2001), 42-50.

[4] Palsson, Sveinn & Agustsson, Eirikur & Timofte, Radu & Van Gool, Luc. (2018). Generative Adversarial Style Transfer Networks for Face Aging. 2165-21658. 10.1109/CVPRW.2018.00282.

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein GANs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus (Eds.). Curran Associates Inc., USA, 5769-5779.

[6] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): n. pag. Crossref. Web.

[7] Alec Radford and Luke Metz and Soumith Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. 2015.

[8] Tero Karras and Samuli Laine and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks”. 2018.

[9] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. 2014.

[10] Zhifei Zhang and Yang Song and Hairong Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. 2017.

[11] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep generative image models using a Laplacian pyramid of adversarial networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15), C. Cortes, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 1. MIT Press, Cambridge, MA, USA, 1486-1494.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Vol. 2. MIT Press, Cambridge, MA, USA, 2672-2680.

[13] Ali Borji. Pros and Cons of GAN Evaluation Measures. 2018.

[14] R. Rothe, R. Timofte and L. V. Gool, "DEX: Deep EXpectation of Apparent Age from a Single Image," 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, 2015, pp. 252-257.
doi: 10.1109/ICCVW.2015.41