Child Face Generation with Deep Conditional Generative Adversarial Networks

Robert Gordan Mingu Kim Alexander Muñoz

Abstract

Child face generation is a computer vision problem in which the goal is to synthesize realistic images of a child given images of its parents. We present a model for this problem based on Deep Convolutional Generational Adversarial Networks (DCGANs). Key challenges in this domain include limited datasets with high dimensional input spaces and the multi-modal nature of the target distribution. We demonstrate convincingly that GANs have a unique ability to capture the latter feature, while use of state-of-the-art training techniques and architecture optimizations allow us to mitigate the impact of the former. As a baseline, we use a simple supervised model that minimizes RMSE with respect to target images. Qualitatively, the clarity and diversity of images reflect advantages of the GAN model when compared to the base model. Quantitatively, we find that after training the discriminator correctly classifies GAN-generated images as fake with a rate of 71.1%, compared to 99.5% classification accuracy on the images generated by the baseline model, suggesting an objective basis for our observations.

1. Introduction

Since their introduction by Goodfellow et al in 2014, Generative Adversarial Networks have gained rapid adoption as a way of modeling and sampling from latent distributions over certain output spaces. In particular, they have proved particularly effective in the image domain, allowing the synthesis of extremely realistic images.

GAN models construct a two-player game opposing a Generator and a Discriminator. The latter is trained to discriminate between output from the Generator and real-world samples, while the former is trained to fool the Discriminator. GANs have faced several practical issues, including a notoriously difficult training process and mode collapse, which occurs when the Generator learns to produce a single point in the output space. However, GANs enjoy several advantages when compared to older generative approaches. GANs do not require a costly inference step, instead learning only through backpropagation. Furthermore, GANs leverage the benefits of deep neural networks, easily incorporating many factors and complex feature interactions.

A natural extension to the model is the Conditional Generative Adversarial Network, in which a set of features are fed to both the Generator and the Discriminator. This allows the model to sample from conditional distributions. We demonstrate a successful application of the model to the child face generation domain. For this problem, we are interested in sampling from the conditional distribution over images of a childs face given images of each of his or her parents. As far as we know, no previous academic work has focused on this problem. Instead, several papers have explored kinship verification, in which the goal is to classify pairs or triples of images as representing a parent-child (or other) relationship. Among other benefits, advances in child face generation may help biological parents to find adopted and/or long vanished children.

Technically, this problem is interesting because of the properties of the distribution we seek to model. A clear cleavage exists in the distribution between male and female children, which means that the generator should be able to produce both types of children. Furthermore, child face generation is an interesting application of computer vision methods such as convolutional layers because of the complex three-way relationship between the visual objects. Rather than simply reducing an image to labels, we must capture information at some level of abstraction that best predicts the face of that parents child.

We first discuss our data and a more formal definition of our problem. We then review related work, both in model structure and in the kinship subdomain of computer vision. In section 4, we describe our model, followed by the training process in section 5 and methods in section 6. Lastly, we present our results and the conclusions we can draw from them.

2. Background

The exists some distribution $C(\theta)$ such that a random variable $A \sim C(\theta)$ consists of an image of a child with the probability that a certain child has the facial phenotype corresponding to that image. Using the universality of the uniform, we can construct a mapping from random noise Z to an arbitrary distribution. Unfortunately, theta includes many variables not easily observable, including presumably some environmental interactions. However, if we have some useful information, we can construct $f_1(Z, f_2(\theta))$ which we can then fit using a training method of our choice and a dataset of child images (in addition to theta features).

Intuitively, a convenient and useful choice for theta is the images of the parents of a given child. We employ the TSKinFace dataset, which consists of 1015 triples of 64x64 pixel images. These represent the father, the mother, and the child. The dataset is roughly balanced between female and male children. However, the limitations of the dataset pose some other challenges. First, the small size of the dataset means that overfitting is a constant concern. This concern is aggravated by the relatively large amount of data used as an input to this model. Secondly, the dataset is biased towards East Asian families, which means that the resulting model may lack the ability to generalize beyond this group of people.

3. Related Work

Goodfellow et al. (2014) first introduced the GAN model (Goodfellow et al., 2014). GANs have demonstrated success in many image-based domains, including image generation (Denton et al., 2015), representation learning (Radford et al., 2015), text-to-image synthesis (Reed et al., 2016), upsampling images (Ledig et al., 2016). The key innovation is the adversarial loss in which the loss of the generator is based on the ability of a discriminator to correctly classify the image as coming from the generator. The weights of the generative model can be trained directly by backpropagating through the (fixed) weights of the discriminator, since the output of the generator and input of the discriminator align.

Conditional Generative Adversarial Networks quickly followed GANs. The first application of this model conditioned only on a single number from 0-9, in order to generate MNIST images. However, one can condition on any type of data that can be easily fed into a neural network (Mirza & Osindero, 2014).

The most natural parallel to the problem of child face generation is image-to-image translation, a domain which generalizes the problem of taking an input image and producing an output image. While such a problem can be approached with a standard CNN and deconvolutions, the CGAN produces more realistic results (Isola et al., 2016). Further developments in this domain include the introduction of models that include an inverse mapping to translate back (Zhu et al., 2017). Our only adjustment to the image-to-image translation model is the use of two images instead of one as the input to our generator.

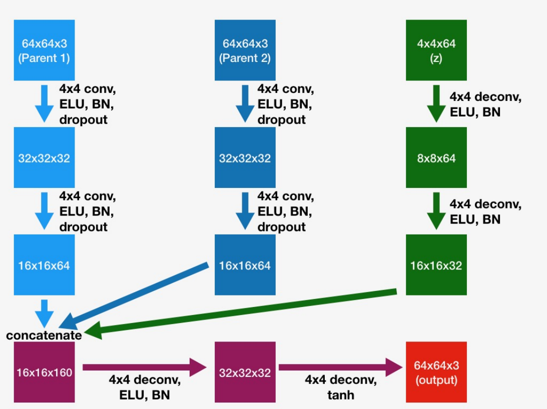

Figure 1. Model architecture for the generator network of childGAN. Separate convolutional parts of the network process the two parent images, while a fully connected layer processes the noise. These inputs are combined at the end of the network to produce the output.

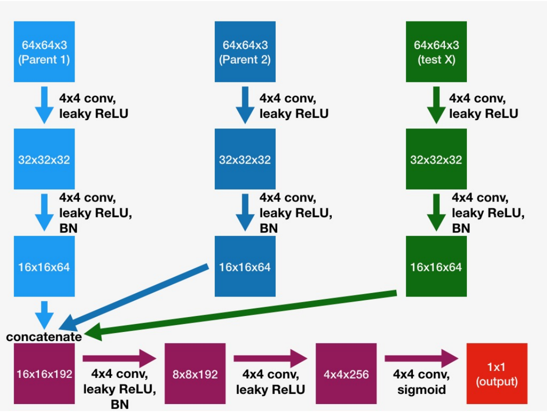

Figure 2. Model architecture for the discriminator network of childGAN. As in the generator, separate convolutional components process the parent images.

In the domain of kinship verification, we find a variety of approaches. The paper we draw our dataset from employs an RBSM (relative symmetric bilinear model) and feature extraction (Qin et al., 2015). However, others have found success with this problem using convolutional neural networks (Zhang12 et al., 2015).

4. Model

As discussed previously, we are interested in modeling and sampling from the distribution over children given their parent images. Because we process each input separately initially, we can express our generator as:

$$f_1(Z, f_2(p_1), f_3(p_2))$$

Here each p_1 and p_2 represent each parent. We call this function G. G is therefore a mapping from random noise onto the space of our data x, which is the images of children. The discriminator then maps from the space of the child images to a single scalar representing a boolean value indicating whether or not the image was produced by the generator.

$$D(x, f_4(p_1), f_5(p_2))$$

We then train D to minimize the probability of misclassifying an image as coming from the generator or the real data. In contrast, G is trained to minimize:

$$1 - D(G(Z))$$

Thus, the generator network is trained to produce images that closely conform to the distribution, as evaluated by the generator. The original GAN paper characterizes the relationship between the two neural networks as that of a two player minimiax game with the following value function:

$$\min_G \max_D V(D, G) =$$

$$E_{x \sim p_{data}(x)}[\log D(x, f_4(p_1), f_5(p_2))]$$
$$+E_{Z \sim p_z(z)}[\log(1 - D(G(Z, f_2(p_1), f_3(p_2))), p_1, p_2)]$$

where in this case log-loss is used.

The parameters of this model are all weights in the neural networks. This presents a challenge because the number of weights is large relative to the the amount of data that we have.

However, the advantages of using a neural network-based model include their ability to represent nearly arbitrary functions and using highly efficient backpropagation to train. This allows us to make very few assumptions about parameters in our model. Nonetheless, we bake a few assumptions about the structure of the data and the problem into our model. For example, we assume we can process the parents' images separately because the features that affect the looks of their child can be encoded at a higher level of abstraction before they interact. We also use convolutional layers in our network components that take images, which reflect an assumption of positional invariance for the features in the images. Intuitively these are reasonable assumptions, and they help us limit the number of parameters we have to train.

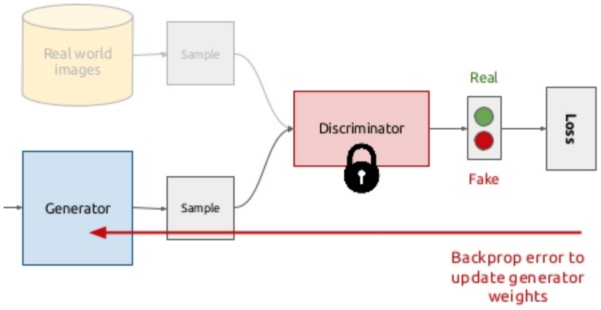

Figure 3. The basic training process for Generative Adversarial Networks. (ima)

Our baseline supervised CNN model is nearly identical to the generative network of our GAN. The key difference is that the loss of this model is calculated with respect to the child image from the data.

5. Training

Because our model is composed of neural networks, we use backpropagation to fit reasonable values for the parameters. When compared to other generative methods, GANs have the advantage of not requiring an inference step. Computationally, this makes it possible to estimate parameters for much larger models. GANs learn approximate parameters, and have a notoriously unstable training process, which makes appropriate training procedures very important.

As mentioned previously, we are limited by a relatively small dataset of 1015 examples of parent-parent-child trio image sets. Because of this, we used transfer learning as a way of pseudo-augmenting our data. For example, much of what the generator must learn at a higher level is simply making human-like faces; regardless of the parent images, the output needs to look like a child. This fundamental feature of human faces and higher level abstraction is not limited to our problem. We assume that this encoding occurs from the first two convolutional layers. Therefore, we can instead pretrain a different DCGAN that takes in a random z matrix input and outputs a child image. The advantage of splitting this problem up in this manner is that we can use any image dataset of children. We therefore used the Large Age-Gap (LAG) database, an image database with pictures taken of people at a variety of ages (Bianco, 2017). We only take the child images, resulting in 9846 photos.

In summary, the following is the basic outline of our overall training process for the DCGAN:

1. Train a DCGAN that receives random array z as an input and generates child images.
2. Initialize the weights of the first two convolutional

layers which initially takes a random array as input, which are shown in Figure 1, of a different DCGAN (intended to generate children from parent images) to be the first two layers of the pre-trained DCGAN from the previous step.

3. Keeping the weights in these layers fixed, train the DCGAN conditioned on parent images to generate potential children images.

In general, we alternate the training of the discriminator and the generator. The objective functions are as described in Section 4. We use binary cross entropy loss as the loss function.

In order to evaluate the quality of the output of our GAN we also implemented a supervised model in which the loss function for the generator is RMSE with respect to the actual child image. In this case we backpropagate directly from the final layer of the generator instead of doing so through the weights of the discriminator first.

6. Methods

In this section we specify the exact details of our implementation of the model and the dataset.

We have two DCGANs total- one of them (which we will call PreDCGAN) is used to generate child images from random input, for transfer learning purposes. The other DCGAN (which we will refer to as childGAN) uses these pretrained weights for two of its layers, and trains on parent images to generate possible children images.

To train PreDCGAN, we used the Large Age-Gap Dataset, which comprises of 9846 child photos. To train childGAN, we used the TSKinFace dataset of 1015 triples of father-mother-child images. Each image was mean-centered around 0 and normalized before training. The architectures of the generator and discriminator for childGAN are shown in figures 1 and 2. The generator takes in two 64x64x3 images of parents, along with a 4x4x64 random array drawn from Unif(0, 1), and outputs a 64x64x3 image. The discriminator takes in two parent images as well as either an image of a real child or a generated image of a child.

During our first attempts at the problem, we tried to train the childGAN without transfer learning. Although we received results, we suffered from mode collapse, where the Generator learned to produce a single point in the output space for input images. Expanding our dataset by the use of transfer learning with PreDCGAN, along with fixed weights for the two convolutional layers shown in green in Figure 1, allowed for greater diversity of output child images.

Much of the architecture was inspired by previous work (Radford et al., 2015), but we made several key changes for our problem. For example, our model is a conditional GAN that takes in 2 parent images, so we had to adjust our model accordingly. Also, although previous work used ReLUs for activation, we found that ELUs allowed the learning to be much quicker. Also, we have found that too much batch normalization causes the discriminator to 'overpower' the generator in the early training process, so some of the batch normalization layers were omitted. We used Adam optimizers for all generators and discriminators, with learning rate of 0.0002 and betas = [0.5, 0.999].

For our final model, we trained for 12 hours, which equated to 4000 epochs.

7. Results

One weakness of the generative model is that it lacks clear quantitative evaluation metrics. We can see in Figure 4 that the generator is successfully learning to confuse the discriminator, such as the spikes in the probability fooling the discriminator around t=590 and t=430. 4 However, the quality of the generated images is subjective and best evaluated by humans. We observed the largest quality increases in the plateaus of the generator performance. We've included sample images from both our GAN and our baseline supervised model. The GAN produces clearer images, and is also capable of outputting a much greater variety of images. Meanwhile, as expected, the fully supervised model does not produce variety and produces comparatively blurry images.

We also explored the performance impact of various tweaks to our model architecture. We found that pre-training the discriminator had an adverse impact on our model, because when the discriminator was much stronger than the generator, the gradients during generator training were not meaningful because it could never successfully fool the discriminator. Lastly, as discussed in Sections 5 and 6, the pseudo-augmentation of data improved subjective performance.

Finally, we compared the generative adversarial model to our baseline model by pitting both against the discriminator network from the GAN. We found that the generator fooled the discriminator in 28.9% of instances, while the supervised CNN model only did so in 0.005% of cases. This suggests the the GAN is a more robust model for sampling child images.

8. Discussion

Our results demonstrate that the appealing properties of GANs hold even for complex models in which the distribution is conditioned on a large set of features such as two images. The most important property in this set is the ability

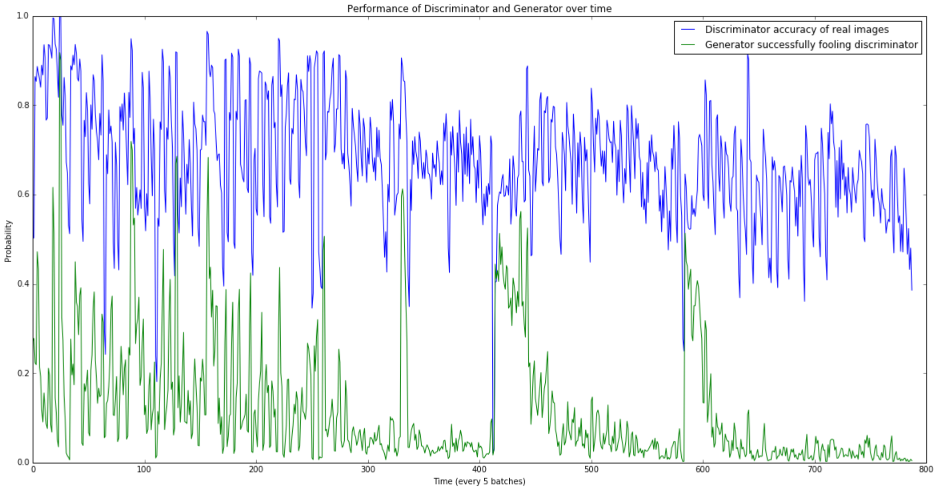

Figure 4. Results for training: probability of generator fooling the discriminator over time

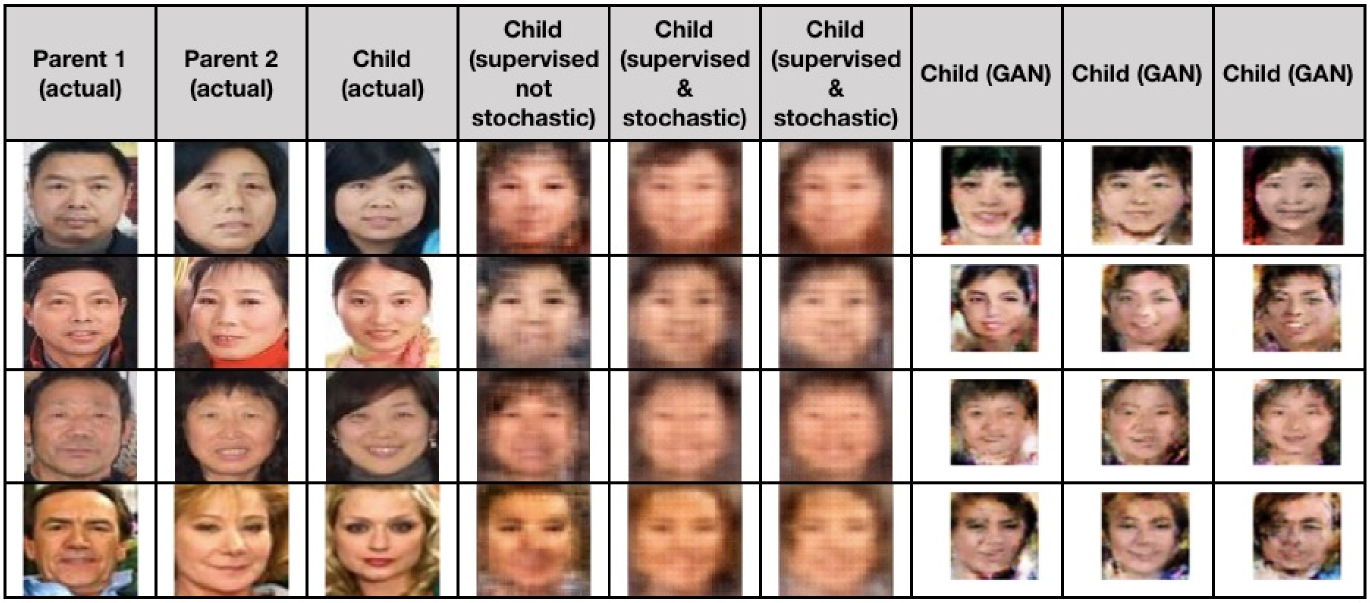
Parent 1 (actual)
Parent 2 (actual)
Child (actual)
Child (supervised not stochastic)
Child (supervised & stochastic)
Child (supervised & stochastic)
Child (GAN)
Child (GAN)
Child (GAN)

Figure 5. Output images by model

to sample from highly multi-modal distributions. The most obvious example of this is that the GAN successfully produces both male and female children. However, the issue of image clarity can also be viewed as a similar problem. While two very clear images, when viewed by a human, may both be judged to be likely images of real children, the pixelwise average of these images is quite unlikely to get the same reaction. Therefore the clarity of images produced by the GAN reflects its ability to find modes, peaks in probability density, instead of averaging them. In contrast, the supervised baseline model, driven by its RMSE, ”hedges” its guesses for pixel values between multiple possibilities.

GANs are notorious for ”mode collapse” a problem that occurs when the generator learns to produce a single value without any variation. We were able to avoid this problem by using several random re-initializations. This ability to produce a highly diverse set of images is important for this domain. Potential applications include synthesizing hypothetical images of lost children to help biological children find them or information retrieval (searching for online images of children for given parents). In these cases, having several diverse candidate images is valuable.

The other interesting result from our experiments is that the GAN was even able to produce remotely reasonable results given the small size of the dataset. This may reflect the fact that generated data augments the real data when training the discriminator. In effect, this doubles the size of the data. Furthermore, our use of transfer learning gave a sort of pseudo-augmentation to the dataset. Our success with this technique could potentially be replicated in other domains with limited data from which to train generative data. By introducing data from spaces which overlap the target space, the generator can quickly learn valuable information applicable to the target space.

9. Conclusion

The results in this paper suggest Generative Adversarial Networks are well suited to problems that involve complex three way relationships between visual objects. With appropriate training and architectural adjustments, these models can be effective even for low-data problems. The techniques, such as pseudo-augmentation, can potentially be extended to other domains.

GANs successfully capture important properties of the parent-child kinship relationships, most importantly the multi-modality. We show that GANs are capable of producing candidate children with highly diverse features.

There has been, unfortunately, relatively little work on the problem of generating child faces given parents. While clearly a small niche, it has the potential for specific real-world impact. The most useful contribution in this area would be the development of larger and more diverse datasets. The limits of the dataset limit the strength and generality of the model.

Further exploration is also needed into techniques for generative models in low-data environments. For example, the incorporation of data from similar spaces proved a promising addition to our model, and large volumes of unlabeled data could also contribute success on this and similar problems.

References

Deep learning for computer vision: Generative models and adversarial training (upc 2016). URL `https://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-mode`

Bianco, Simone. Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters*, 90:36–42, 2017. doi: 10.1016/j.patrec.2017.03.006.

Denton, Emily L, Chintala, Soumith, Fergus, Rob, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pp. 1486–1494, 2015.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Ledig, Christian, Theis, Lucas, Huszár, Ferenc, Caballero, Jose, Cunningham, Andrew, Acosta, Alejandro, Aitken, Andrew, Tejani, Alykhan, Totz, Johannes, Wang, Zehan, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Qin, Xiaoqian, Tan, Xiaoyang, and Chen, Songcan. Tri-subject kinship verification: Understanding the core of a family. *IEEE Transactions on Multimedia*, 17(10):1855–1867, 2015.

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

Reed, Scott, Akata, Zeynep, Yan, Xinchen, Logeswaran, Lajanugen, Schiele, Bernt, and Lee, Honglak. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

Zhang12, Kaihao, Huang, Yongzhen, Song, Chunfeng, Wu, Hong, Wang, Liang, and Intelligence, Statistical Machine. Kinship verification with deep convolutional neural networks. 2015.

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, and Efros, Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.