Predicting Child Facial Features using Encoder-Decoder Latent Space Interpolation and Generative Adversarial Networks

Gregory Hunkins
University of Rochester
ghunkins@u.rochester.edu

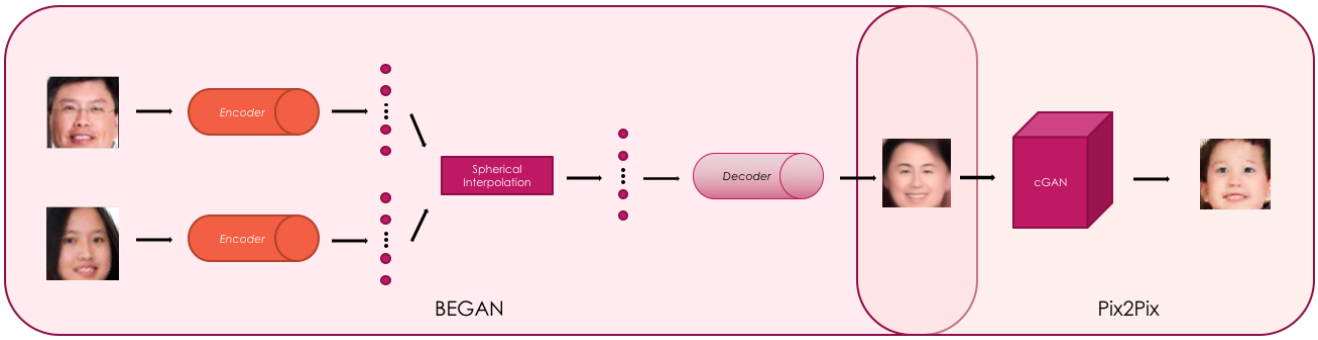

Figure 1: Illustration of the best dual-architecture combination method using BEGAN for parental facial feature encoding and interpolation and Pix2Pix for child facial feature generation. The flow of information is from left to right. Results are subjectively best-in-class on the validation data.

Abstract

Generative Adversarial Networks have shown tremendous gains in the realm of image generation, particularly for the specific task of generating human faces. An interesting test of the capabilities of these architectures is to use transfer learning to combine the learned internal representations of two adult faces from such architectures and then learn a method to predict the resulting child's face. We investigate this novel problem using a dual-architecture workflow: (1) facial feature extraction and interpolation and (2) child face generation. We show that spherical interpolation of the latent space representations in a state-of-the-art encoder-decoder Generative Adversarial Network coupled with a image-to-image Conditional Generative Adversarial Network can successfully perform this task. However, the results do not match the quality of state-of-the-art image generation networks, and thereby indicate that an end-to-end single architecture solution may be the next step in solving this issue. Our contribution is the initial investigation into this novel problem and a new dataset, Parent-Child in the Wild (PCW), for the purpose of solving such mother-father-child image generation tasks.

1. Introduction

The task of generating images involves learning an architecture that can learn a probabilistic representation of a given image class and subsequently reliably sample this target image space.

2. Related Work

2.1. Generative Adversarial Networks

Goodfellow et. al. introduced the base model for Generative Adversarial Networks (GAN) with a ground-breaking paper in [6]. The model uses a discriminator D and generator G to play an adversarial minimax game to learn value function $V(D, G)$ as described in Equation 1.

$$\min_G \max_D V(D,G) = \mathbb{E}_{p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G_x(z)))]. \quad (1)$$

Radford et. al. [13] expanded on this idea by combining Convolutional Neural Networks with the GAN architecture and introduced Deep Convolutional Generative Adversarial Networks (DCGAN), effectively revolutionizing the task

of image generation. Recent successful adaptations of this architecture style for the specific task of human face generation include Energy Based Generative Adversarial Networks (EB-GAN) by Zhao et. al. [16], Boundary Equilibrium Generative Adversarial Networks (BEGAN) by Berthelot et. al. [2], and Progressively Growing Generative Adversarial Networks (PG-GAN) by Karras et. al. [8]. The EB-GAN architecture modeled the discriminator D using a novel energy function paradigm and introduced an encoder-decoder structure for use as the discriminator. The BEGAN architecture [2] expanded on this work by combining a loss inspired by the traditional Wasserstein loss [1] combined with an equilibrium term designed to balance the generator and discriminator. Additionally, Berthelot et. al. [2] showcase the power of latent space interpolation and investigate its capabilities to find mid-point representations between two images. Finally, the PG-GAN architecture [8] showcases the ability to generate state-of-the-art, high resolution images by introducing a novel growth mechanism that progressively grows the generator and discriminator throughout training. These papers are the foundation of the experiments described in this paper.

2.2. Latent Space Mathematics

Radford et. al. in their DCGAN work [13] demonstrated the ability of their architecture to perform mathematics in the latent space. Famously, the simple example of adding a pair of sunglasses to a woman using image algebra highlighted the power of latent space manipulations. Berthelot et. al. in their BEGAN work [2] use interpolation to investigate the latent space between image encodings and thereby show that their architecture successfully generalized the image contents. An illustration of this can be seen in Figure 5. The mid-point interpolation between two images serves as the inspiration to use latent space interpolation to find combined parental facial features for this task.

3. Datasets

3.1. CelebA

For the task of training existing architectures for facial generation, the Large-scale CelebFaces Attributes Dataset (CelebA) was used. It contains 202,599 facial images of 10,177 unique identities in a variety of forward-facing poses.

3.2. Parent-Child in the Wild

Upon investigation into the community at large, an appropriate mother-father-child facial image dataset for learning a mapping from parental features to child features was not found. As such, the Parent-Child in the Wild (PCW) dataset is created and released to the public. It contains 767 verified and cleaned family facial photo collections. Each images is 128x128.

Figure 2. A representative sample of the pruned collected images for PCW. The top image is the original image. Below this, from left to right, are the identified family members. Gender-age annotations from left to right are: F (38-43), M (8-12), F (4-6), M (15-20).

3.2.1 Scraping

The *google-images-download* Python package [15] was the primary interface for the scraping of images. Using a variety of keywords, a diverse set of families with young children was manually collected. Using the *similar images* feature, images that displayed similar characteristics to these oracle images were subsequently automatically collected. As such, a total of 780 family pictures were successfully scraped. Only images with Acceptable Use Policies were attained.

3.2.2 Image Standardization

For image standardization, three main transformations were performed on each scraped image: (1) facial detection and cropping, (2) facial alignment, and (3) re-sizing. For these tasks, the *Face Recognition*, *Pillow* [11], and *OpenCV* [4] Python packages were used, respectively. A FaceNet-inspired Convolutional Neural Network (CNN) model is used for facial detection [14]. The facial alignment method identified facial keypoints then used affine warping to force a consistent facial profile. Finally, the images were resized using the nearest-neighbors method [5].

3.2.3 Dataset Verification

The final step before a family candidate image set was accepted to PCW was verification of a biological family structure. As such, the following criteria X for a family F in the dataset was enforced:

$$X = dad \wedge mom \wedge child$$

where

$$\begin{aligned} dad &= \exists!\ p \in F\ s.t.\ p \in Person(Male, age > 18) \\ mom &= \exists!\ p \in F\ s.t.\ p \in Person(Female, age > 18) \\ child &= \exists\ \ p \in F\ s.t.\ p \in Person(age < 18) \end{aligned} \tag{2}$$

For gender and age verification, two pre-trained CNNs were used from an online repository [12]. They are unofficial implementations of Levi and Hassners' architectures from Age and Gender Classification Using Convolutional Neural Networks [9].

In future releases of this dataset, using a kinship verification method [10] in addition to family structure verification would allow for more robust verification and allow for larger automated collections with quality assurance.

3.3. Large Age-Gap Face Verification

Due to the small size of the collected PCW dataset, it was necessary to diversify and augment it for the task of child face generation. As such, the Large Age-Gap Face Verification dataset (LAG) from the Imaging and Vision Laboratory was incorporated [3]. The adult images were auto-encoded to maintain interpolated parental facial features style, and the children images . The facial bounding box identification and cropping, facial alignment, and re-sizing cleaning methods applied to PIW were also applied to LAG.

4. Method

The method primarily focused on a dual-architecture system: (1) a feature extraction and interpolation architecture, (2) a child facial image generation architecture. Figure 1 showcases the best architecture and a best-in-class result on the validation data.

4.1. Data Cleaning

Verification of input image size (128x128 for all architectures) was included in each of the architectures dataloaders. Additionally, the BEGAN architecture required a tighter facial profile and as such cropped input images to the appropriate profile and then re-sized the results to the appropriate size. No other data cleaning occurred during this stage.

4.2. Feature Extraction and Interpolation

Two architectures, BEGAN and PG-GAN, were investigated for the purpose of extracted parental facial features and finding an appropriate mid-point representation. The method for finding spherically interpolated vector for two arbitrary vectors p_0 and p_1 is shown in Equation 3. Ratio t is equivalent to $\frac{1}{2}$ to find the mid-point representation.

$$Slerp(p_0, p_1; t) = \frac{\sin{(1 - t\Omega}}{\sin{\Omega}} + \frac{\sin{t\Omega}}{\sin{\Omega}} p_1 \tag{3}$$

The interpolated vector representations using Equation 3 were the input vector and the encoder output for the BEGAN and PG-GAN architectures, respectively. These interpolated vectors were then expanded using the generator and decoder for the BEGAN and PG-GAN architectures, respectively, to give the input image for child facial feature generation architecture.

No deviations from the base architectures for both works were incorporated into this work.

4.3. Child Facial Feature Generation

Using the output of the feature extraction and interpolation architecture, an image-to-image translation network was used to learn a mapping from the interpolated parental facial features the child facial features. A single architecture style, the Pix2Pix architecture from Isola et. al.'s Image-to-image Translation with Conditional Adversarial Networks [7] was used.
No deviations from the base architecture were incorporated into this work.

4.4. Training Implementation

All networks were trained using two Telsa K20 GPUs on the University of Rochester's BlueHive supercomputer. Training the base BEGAN network took approximately five days. Training the Pix2Pix architecture took approximately 8 hours. All other architectures were pre-trained.

5. Experiments

Figure 1 showcases the best architecture and a best-in-class result on the validation data as verified from the experiments.

5.1. Feature Extraction & Interpolation

5.1.1 Progressively Growing Generative Adversarial Network

Due to the impressive quality of the state-of-the-art PG-GAN architecture [8], interpolation of the input vectors to a PG-GAN pre-trained architecture was investigated.

Three representative results from an initial input vector interpolation experiment can be visualized in Figure 3. As

Figure 3. Illustration of two representative interpolations of the input vector using the PG-GAN architecture. The middle image of each image trio is the interpolation.

Figure 4. Illustration of representative parental facial features spherical mid-point interpolations using the BEGAN architecture. From left to right in each concatenated image trio, the images are: father, interpolation, mother.

can be seen, the architecture learned no meaningful correlation between the input vector and output facial features. As the authors did not design the architecture for this purpose, this is to be expected. As such, due to these results, this architecture was not investigated for full

5.1.2 Boundary Equilibrium Generative Adversarial Network

The proven capability of the BEGAN architecture by et. al. to provide meaningful interpolation between facial images can be seen in Figure 5. Figure 4 showcases the model's capabilities at capturing representative facial features from both parents. Due to these experimental results, this architecture was chosen for the stage of parental feature extraction and interpolation.

5.2. Child Face Generation

As noted in the Method section, the Pix2Pix architecture was initially investigated for the generation of child facial features from the output of the feature extraction and interpolation architecture. Two experiments were undertaken to investigate which expansion network strategy worked best for this task: upsampling or deconvolution.

5.3. Pix2Pix using Upsampling

Figure 6 showcases a random, representative output of the validation dataset using the upsampling method of the Pix2Pix architecture. As can be seen, a high degree of noise is present in the generated images.

5.4. Pix2Pix using Deconvolution

Figure 7 showcases a random, representative output of the validation dataset using the deconvolutional method of the Pix2Pix architecture. While many outputs are still quite fuzzy, this experiment reveals the best investigated architecture style for this task.

5.5. Fine-Tuning a Pre-Trained Model

The final experiment to fine-tune a pre-trained BEGAN architecture for the task of creating an end-to-end solution to the problem was undertaken. Unknown causes disallowed the architecture from returning meaningful results, however, and this experiment will be undertaken at a later date.

6. Codebase & Dataset

All code and the PCW dataset used in this paper have been made open-source and are available via the link below. Any updates will be reflected in the README of the Github repository. Questions about either the code or the dataset should be directed to the author.

Link: `https://github.com/ghunkins/Child-Face-Generation`

7. Conclusion

In this work, the novel task of predicting child facial features using parental facial features is undertaken. Contributions include a working, initial investigation of this task and the Parent-Child In The Wild dataset for further experimentation regarding this task. A dual-architecture workflow is investigated for this task using the following paradigm: (1) a parental facial feature extraction and interpolation architecture and (2) a child facial feature generation architecture. Experiments reveal that spherical interpolation of the latent space parental representations from the state-of-the-art encoder-decoder BEGAN architecture coupled with the state-of-the-art Pix2Pix image-to-image architecture for child facial feature generation can successfully perform this task using subjective analysis. Further experimentation developing an end-to-end architecture using the knowledge attained from these experiments is hypothesized to be the next best step forward.

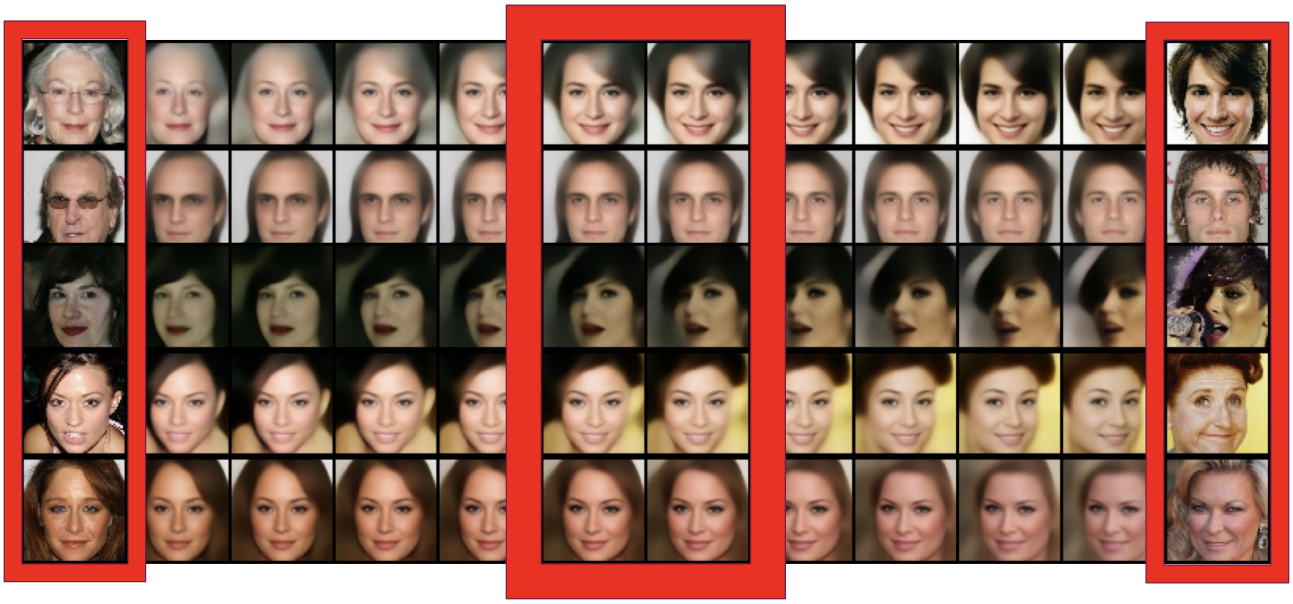

Figure 5. Illustration of the interpolation capabilities of BEGAN. Using a variety of spherical interpolation methods from with t [0, 1] in steps of $\frac{1}{10}$. The middle red box highlights the midpoint interpolations between the far left and far right images.

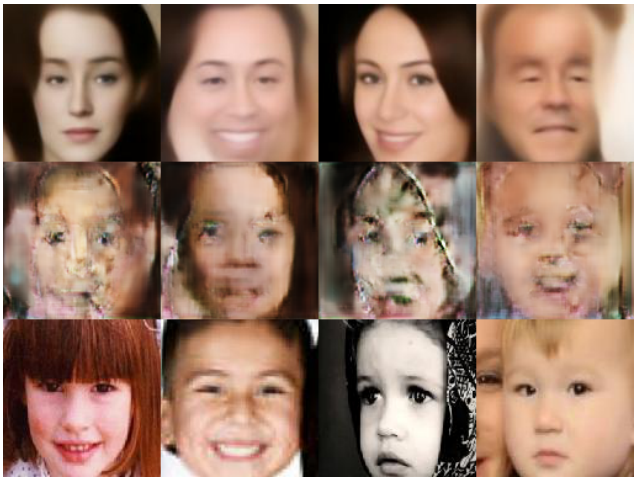

Figure 6. Illustration of random, representative output of the validation data using the Pix2Pix upsampling method. From top to bottom the images are: input interpolated parental facial feature representations, output child facial features, ground truth.

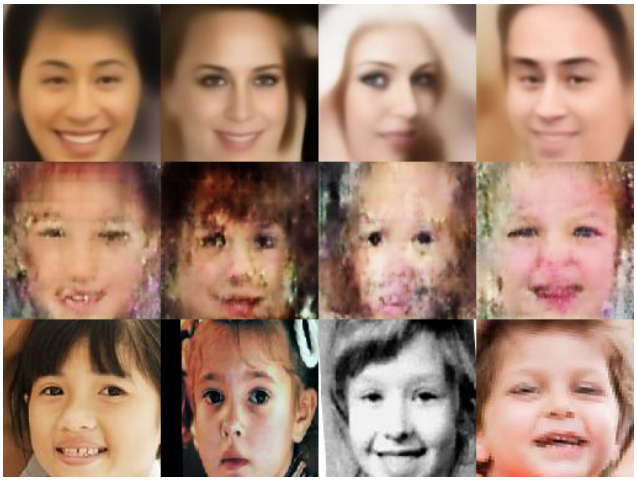

Figure 7. Illustration of random, representative output of the validation data using the Pix2Pix deconvolutional method. From top to bottom the images are: input interpolated parental facial feature representations, output child facial features, ground truth.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[3] S. Bianco. Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters*, 90:36–42, 2017.

[4] G. Bradski and A. Kaehler. Opencv. *Dr. Dobbs journal of software tools*, 3, 2000.

[5] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information*

processing systems, pages 2672–2680, 2014.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[9] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[10] J. Lu, J. Hu, and Y.-P. Tan. Discriminative deep metric learning for face and kinship verification. *IEEE Transactions on Image Processing*, 26(9):4269–4282, 2017.

[11] F. Lundh, M. Ellis, et al. Python imaging library (pil), 2012.

[12] D. Pressel. Age/gender detection in tensorflow, 2016–. [Online; accessed May 14, 2018].

[13] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[14] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[15] H. Vasa et al. Google images download, 2015–. [Online; accessed May 14, 2018].

[16] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.