CS230

GANimation: Facial animation from images

Qi Gao **Liang Xiang**

Abstract

In this project we implemented a GAN-based architecture which transforms facial images into a sequence of images capturing continuous facial movements. This architecture is unique in that it leverages Action Unit (AU) annotations to describe facial expressions. Using images annotated with their activated AUs as inputs, our model is intended to control the magnitude of activation of each AU. We achieved an overall accuracy of 70% on the CelebA dataset and identified specific cases where improvement can potentially be made in the future.

1 Introduction

We noted that Generative Adversarial Network (GAN) has achieved impressive results for facial image transformation of late. Among these studies, StarGAN provides a novel approach to transform the image for multiple domains using a single model. Though being flexible and generalized, StarGAN can only generate a discrete number of images, making it difficult to capture continuous facial movements.

To alleviate this limitation, in this project we aim to implement an architecture that is similar to StarGAN but is based on Action Unit (AU) annotations. Our rationale is that facial expressions are combined and coordinated action of facial muscles that cannot be classified by a discrete number of categories. The Facial Action Coding System (FACS) provides a systematic approach to describe facial expressions in terms of AUs, which are anatomically related to the contractions of facial muscles. Taking images annotated with their activated AUs as inputs, our model can control the magnitude of activation of each AU and even combine several of them.

2 Related work

Unpaired image-to-image translation. There are several existing architectures that have already achieved good performance in this area. For example, UNIT[4] pursues a variational autoencoders (VAEs) approach and combines it with a GAN framework. CycleGAN[9] and DiscoGAN[3] leverages a cycle consistency loss to preserve key attributes between the input and the target image. StarGAN[1], unlike previous models, can learn the relations among multiple domains using only a single model by conditioning the GAN's generation process on images of a specific domain. However, this state-of-the-art model can only generate a discrete number of expressions and is constrained by the content of the available dataset.

Synthesize facial expressions. Early work typically use mass-and-spring models to physically approximate skin and muscle movement[2] or apply 2D and 3D morphings[8]. Results generated by those models are unable to capture subtlety in the image or illumination changes. More recent works using convolutional networks are able to address these issues but can only be applied to discrete expression categories, e.g. happy, sad, angry, etc..In comparison, the model we implement in this project conditions the GAN model on a continuous set of muscle movements. Therefore, it becomes feasible to generate a large range of facial expressions and even show a video-like facial expression transition in a smooth manner.

,

3 Dataset and Features

The dataset we use for this project is unique in nature as all the images have to be accurately annotated with Action Units (AUs). We evaluated several datasets regarding their image quality and distribution.

The first one we looked at is Affectiva-MIT Facial Expression Dataset (AM-FED)[6] which captures naturalistic and spontaneous facial responses to three Super Bowl ads. Specifically, this dataset consists of 545 videos that have been manually FACS coded.However, after converting the 545 AU-labelled videos into 20938 photos at one-second interval, we realized that a vast majority of these images containing identical facial expressions and a significant proportion of them were taken in bad illumination conditions, making it difficult to use as inputs of our model.

The second dataset we investigated is a preprocessed CelebA dataset created by Yuedong Chen. CelebA[5] is a facial image dataset containing more than 200K celebrity images. We noted that it is more ideal for our project than AM-FED in that it offers a greater number of images. In addition, CelebA has a balanced distribution of gender, age, ethnicity, pose, background and illumination condition. The preprocessed database cropped faces in the image and all the images are annotated with Action Units. Details about the dataset can be found at `https://github.com/donydchen/ganimation_replicate.git`. We store the dictionary containing photo ids and corresponding Action Units in a pickle file (aus.pkl) which will serve as the input of our model.

4 Methods

The architecture we implement is proposed by Pumarola, etc in their 2018 paper[7]. As shown in Figure 1, there are two sets of generator and discriminator in the model. The first set generates and discriminates a image while the second set is designed to replicate the input image. Each generator of this architecture regresses out two masks, a color mask C and attention mask A. The final image can be obtained as: $I_{y_f} = (1 - A) \cdot C + A \cdot I_{y_0}$,where $A = G_A(I_{y_0}|y_f) \in \{0, ..., 1\}^{H \times W}$ and $C = G_C(I_{y_0}|y_f) \in R^{H \times W \times 3}$. The mask A indicates to which extend each pixel of the C contributes to the output image I_{y_f}. This approach allows the generator to focus exclusively on the pixels defining the facial movements, leading to sharper and more realistic synthetic images.

The discriminator maps the input image I to a matrix $Y_I \in R^{H/2^6 \times W/2^6}$, where $Y_I[i, j]$ represents the probability of the overlapping patch ij to be real. To ensure the image is conditioned on the specified AU, on top of it we add an auxiliary regression head that estimates the AUs activations $\hat{y} = (\hat{y}^1, ..., \hat{y}^N)^T$ in the image.

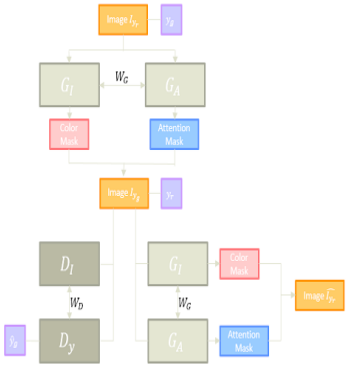

Figure 1: Architecture

The loss function consists of four terms: image adversarial loss, attention loss, identity loss and conditional expression loss. Details about these four terms are discussed below.

Image adversarial loss. The image adversarial loss is applied to learn the parameters of the generator G by pushing the distribution of the generated images to the distribution of the training images. A gradient penalty is added for the critic network computed as the norm of the gradients with respect to the critic input.

Attention loss. The attention loss is to drive the attention masks to be smooth. In addition, it prevents the mask from saturating. When training the annotation for the attention mask as well as the color mask are learnt from the resulting gradients of the critic module and the rest of the losses. However, the attention mask tends to saturate to 1 which makes the generator having no effect. This problem is addressed by regularizing the mask with an L2-weight penalty.

Conditional expression loss. The conditional expression loss conditions the expression of the generated images to be close to the desired one. While reducing the adversarial loss, the generator G must also satisfy the target facial expression encoded by y_f. Therefore, this loss is defined with two components: an AU regression loss with fake images to optimize G and an AU regression loss with real images to learn the regression head on top of the discriminator D.

Identity loss. The identity loss is defined to preserve the person texture identity. With the rest of the losses the generator G is enforced to generate photo-realistic face transformations but there is no guarantee that the face in both the input and the output images are for the same person. To address this, we force the generator to maintain the identity by penalizing the difference between the original image I_{y_0} and its reconstruction.

The code is available here: `https://github.com/vickyq0513/GANimation`

5 Experiments/Results/Discussion

- Training

We implemented the same structure and hyperparameters suggested by original authors in paper[7], which was already fine tuned for EmotioNet dataset. While tuning at a small epoch, it shows that it has already reached the optimal status by provided set of values. Learning rate for generator and discriminator Adam is set to be 0.0001, the weights in loss function for real/fake discriminator loss, condition discriminator loss, cycle loss, mask loss, gradient penalty loss and mask smoothness loss are: 1, 4e3, 10, 0.1, 10, 1e-5.

- Results and Discussion

Figure 2 shows two sample outputs from our test model. In these cases, the target expression is a randomly selected another image among our dataset. The first, second and last column in Figure 2 are the input image, the fake target image and replicated input image, respectively. The third and the fourth columns show attention masks and color masks generated by the first generator while the fifth and the sixth columns contain masks created by the second generator.

Figure 2: Sample Output

To evaluate the overall performance of the model, we randomly picked 100 test samples out of 10,374 total test images, and got 70% natural and smooth transitions to target expression (as shown in Figure 3) while the rest are not as real or natural as expected. Since our original cited paper did not provide any quantitative metrics to compare with, we are trying to collect more information from error analysis as below.

In Figure 4, we look at cases where the expression fails to achieve desired image, or the continuous transition is not natural as expected. We note that there are 3 main types. The first type of error involves teeth creation. Like the first row in Figure 4, a series from a face revealing no teeth in the source expression gets unnatural when we are trying to create teeth and achieve the desired facial expression on the far right. Unmatched teeth shapes, sizes and even colors could lead to unreasonable transition image. This has contributed to 50% of error samples.

Figure 3: Successful Cases

A second case has to do with covered or obscured facial features.Some pictures include people placing hands or microphones in front of their key features, which would apparently affect the application of trained model to update the action unit associated with the features. The last case involves failure to process part of images caused by their specific angles, or labeling initially was not accurate enough.

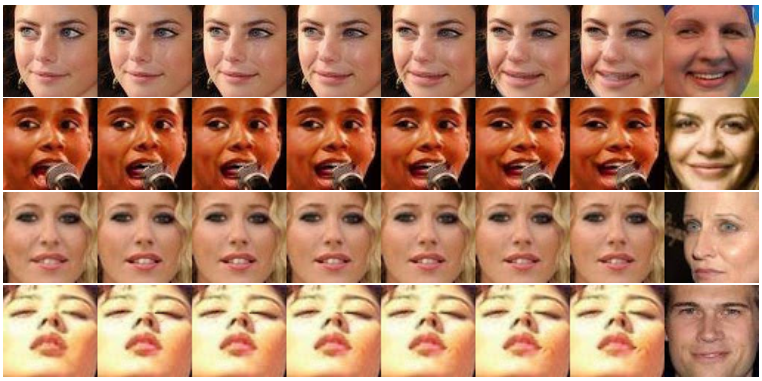

Figure 4: Unsuccessful Cases

6 Conclusion/Future Work

To recap, in this project we implemented a conditional-GAN model to synthesize facial expressions in a continuous manner. We achieved an accuracy of 70% and were able to smoothly transit the input image to a target expression in most cases.

For edge cases that have to do with teeth creation and special poses or angles, improvement can potentially be made by including more images with such expressions in the training set, or add one more hyperparameter as the weight of these scenarios in the loss function which could be tuned in the process. Another way is to create an extra model to find matched teeth from other open-mouth faces with similar facial features and skin types. More scrutiny on labelling training dataset and more diversity of poses and angles should also help improve the smoothness of the generated series.

7 Contributions

We worked together in discussing and deciding the topic of the project and searching for related literature/data. After that, Gao was primarily focused on training the model, fine-tuning parameters and evaluating performance while Xiang preprocessed datasets, created prototype models on Colab and vetted the test code.

References

[1] Yunjey Choi et al. *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*. 2017. arXiv: 1711.09020 [cs.CV].

[2] M. A. Fischler and R. A. Elschlager. “The Representation and Matching of Pictorial Structures”. In: *IEEE Transactions on Computers* C-22.1 (Jan. 1973), pp. 67–92. ISSN: 2326-3814. DOI: 10.1109/T-C.1973.223602.

[3] Taeksoo Kim et al. *Learning to Discover Cross-Domain Relations with Generative Adversarial Networks*. 2017. arXiv: 1703.05192 [cs.CV].

[4] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. *Unsupervised Image-to-Image Translation Networks*. 2017. arXiv: 1703.00848 [cs.CV].

[5] Ziwei Liu et al. “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

[6] Daniel McDuff et al. “Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected "In-the-Wild"”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2013.

[7] Albert Pumarola et al. *GANimation: Anatomically-aware Facial Animation from a Single Image*. 2018. arXiv: 1807.09251 [cs.CV].

[8] Hui Yu, O. Garrod, and P. Schyns. “Perception-driven facial expression synthesis”. English. In: *Computer Graphics* 36.3 (May 2012), pp. 152–162. ISSN: 0097-8493. DOI: 10.1016/j.cag.2011.12.002.

[9] Jun-Yan Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. 2017. arXiv: 1703.10593 [cs.CV].