Transformation of Emotions in Images Using Poisson Blended Generative Adversarial Networks (Student Abstract)

Aristidis Dernelakis,[1,2] **Jungin Kim,**[1,3] **Kevin Velasquez,**[1,3] **Lee Stearns**[1]
[1]Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Rd, Laurel, Maryland 20723
[2]University of Maryland, Baltimore County, 1000 Hilltop Cir, Baltimore, Maryland 21250
[3]Johns Hopkins University, 3400 N. Charles Street, Baltimore, Maryland 21218
ad20@umbc.edu, jkim576@jhu.edu, kvelasq1@jhu.edu, Lee.Stearns@jhuapl.edu

Abstract

We propose a novel method to transform the emotional content in an image to a specified target emotion. Existing techniques such as a single generative adversarial network (GAN) struggle to perform well on unconstrained images, especially when data is limited. Our method seeks to address this limitation by blending the outputs from two networks to better transform fine details (e.g., faces) while still operating on the broader styles of the full image. We demonstrate our method's potential through a proof-of-concept implementation.

Introduction

Generative algorithms that can transform the styles and features of an image between domains are useful for creating artwork, simulation and synthetic data generation, and numerous other applications. However, while generative adversarial networks (GANs) work well for specific, constrained applications, they are less reliable in unconstrained settings. For example, a GAN that can transform the attributes of a cropped face may perform significantly worse when presented with an image of multiple people and varied background scenery. This limitation is amplified when the size of the dataset is small, as is true for many tasks. We aim to create a process for transforming the content of unconstrained images.

To demonstrate a scenario where this approach is needed, we chose to transform the emotional content of an image to some target emotion. Emotion is difficult to define in an unconstrained image, and can be captured through many small details such as the facial expressions and poses of the people in the scene, as well as the broader colors and other stylistic elements. We observed that a traditional algorithm like CycleGAN (Zhu et al. 2017) would focus primarily on the latter, ignoring the finer details of the scene. We propose a method to address this limitation which combines the outputs of two or more GANs blended together. More specifically, we use StarGAN-v2 (Choi et al. 2020) to transform the facial expressions of the people in an image, then blend them back into the original image and transform the result using a fast style-transfer network (Engstrom 2016). Although we did not investigate them in this preliminary work, methods such as Cascade EF-GAN (Wu et al. 2020) could be used in place of StarGAN to more precisely and seamlessly alter finer expression cues on the faces in an image. We believe that our multi-stage approach will produce more meaningful transformations than any single conventional GAN.

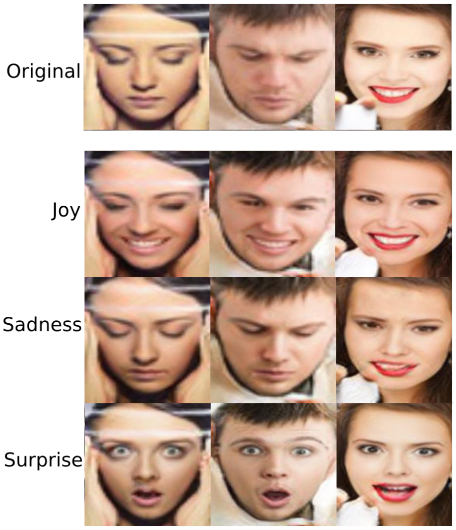

Figure 1: Example transformations of facial emotion expressions using StarGAN-v2. Trained on cropped faces extracted from WEBEmo, relabeled using FER. The top row shows the original image, and the bottom three show the transformation to Joy, Sadness, and Surprise, respectively.

Curating the WEBEmo Dataset for Training

We selected the WEBEmo dataset (Panda et al. 2018) to train our GANs. WEBEmo contains 267,441 stock photos separated into seven emotion categories (joy, sadness, surprise, anger, fear, love, and confusion), each containing a variety of image content and styles as well as different representations for that emotion. The dataset provides a good baseline for training our algorithms; however, we observed that many of the individual facial expressions in the images that contained people were not recognizable as the assigned label for the full image. This confusion resulted in poor performance in our initial experiments.

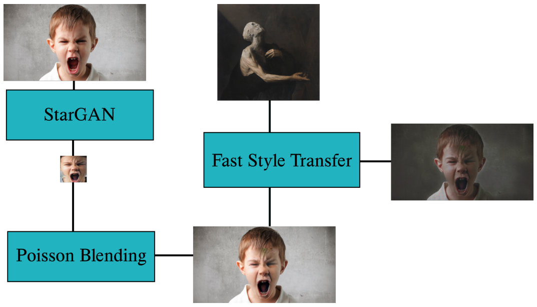

Figure 2: Illustration of the steps used to transform an image's emotional content to a target emotion. Source images from the WEBEmo dataset.

To more reliably train our facial expression GAN, we first curated the dataset using an automated process. We first applied YOLO v3 (Redmon and Farhadi 2018) to select the images containing people, and FaceNet (Schroff, Kalenichenko, and Philbin 2015) to crop out all recognizable faces in those images. Next, we applied the FER classifier (Goodfellow et al. 2013) to reclassify the cropped face images. Our goal was to select the cropped face images that could be classified with sufficiently high confidence by FER–we discarded any images where the highest confidence was below 30% as well as those classified as one of the two classes (neutral and disgust) that did not exist in the WEBEmo dataset. A confusion matrix from this process is available in the accompanying Supplemental Materials.

Proof of Concept

To demonstrate our approach, we selected two images from the WEBEmo dataset: an image of a person with an expressive facial expression and another of a painting with an expressive style to use as the target emotion. We crop the face from the first image and use our trained StarGAN-v2 model to transform the facial emotion to the emotion label from the second image. Examples of this transformation are illustrated in Figure 1. Next we place the transformed face back into its original position in the first image and apply Poisson blending (Pérez, Gangnet, and Blake 2003) to smooth the remaining artifacts. Last, we apply the fast style transfer network to transform the image to match the style of the selected painting.

An example of this process is illustrated in Figure 2. In this example, we selected an image containing a person's face with the label ”anger”, then transformed the cropped facial expression to ”fear” and blended it back onto the original image. We then selected a painting with the label ”fear” that had a similar structure and applied a style transformation to create the final image. The final image highlights fear on the subject's face as additional furrows above the eyes and subtle changes to the eyebrows themselves, causing them to slope more horizontally than the original image, while the style transformation primarily results in darker, less saturated colors.

Conclusion

Our proof-of-concept demonstration shows that our method for transforming both the broad and fine-grained representations of emotions in an unconstrained image is feasible. We plan next to explore how well our techniques will extend to the more varied images contained in the WEBEmo dataset. Additionally, our process for selecting a style image to use for this demonstration was entirely manual. However, we believe that that process could be automated by analyzing the structure of the original image and selecting a style image with the desired emotion that is as similar as possible to the structure of the original. A similar process could be used to replace or make adjustments to objects or other fine details that a single GAN could not reliably transform. The process would likely need to incorporate additional heuristics and modules to allow it to handle the highly varied types of image content present in the WEBEmo dataset. However, these techniques would allow the task to be subdivided into more constrained, easier to solve problems, which can also function with limited training data. If this process can be fully automated, then it provides a baseline for other researchers to more reliably transform the content of unconstrained images between domains.

References

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Engstrom, L. 2016. Fast Style Transfer. https://github.com/lengstrom/fast-style-transfer/. Accessed: 2021-09-15.

Goodfellow, I.; Erhan, D.; Carrier, P.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; Zhou, Y.; Ramaiah, C.; Feng, F.; Li, R.; Wang, X.; Athanasakis, D.; Shawe-Taylor, J.; Milakov, M.; Park, J.; and Bengio, Y. 2013. Challenges in Representation Learning: A Report on Three Machine Learning Contests. *Neural Networks*, 64: 8.

Panda, R.; Zhang, J.; Li, H.; Lee, J.-Y.; Lu, X.; and Roy-Chowdhury, A. K. 2018. Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias. In *European Conference on Computer Vision*.

Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, 313–318.

Redmon, J.; and Farhadi, A. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

Wu, R.; Zhang, G.; Lu, S.; and Chen, T. 2020. Cascade EF-GAN: Progressive Facial Expression Editing with Local Focuses. arXiv:2003.05905.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.