Problem Definition/Introduction

Neural Style Transfer is a technique that takes two images or videos– one as the source and the other as the reference– and applies the style of the reference image to the source image. The model uses deep learning to extract the stylistic features of the reference image, and apply them to the source image without changing the content of the source image. StarganV2 is a specific type of Neural Style Transfer that performs image-to-image translation, working with animals to animals or humans to humans; it changes the physical features surrounding the face of the source image to the reference images– such as the gender, hairstyle, and skin color– while maintaining the source image's distinctive facial features– such as the source's facial shape and expression. For this project, we analyzed StarganV2's translational ability by testing it on a multitude of faces from the CelebA-HQ dataset, making sure to pull from a variety of races, hair colors, and image backgrounds for both males and females. We did so so that we could see if the model is biased in any way towards different types of people, and how realistic the images it generates are.

Method, Implementation, and Experiments

For the project, we used the StarganV2 model developed by ClovaAI that had been pre trained on the CelebA-HQ and AFHQ datasets. Using their pretrained weights, we ran the model on a randomly selected set of images from the CelebA-HQ dataset, making sure to get a diverse population of faces as mentioned above; we also decided to include our faces for the tests. With this data, we obtained an output image matrix of crosses/translations between the various source and reference images, which is what we used for our analyses below.

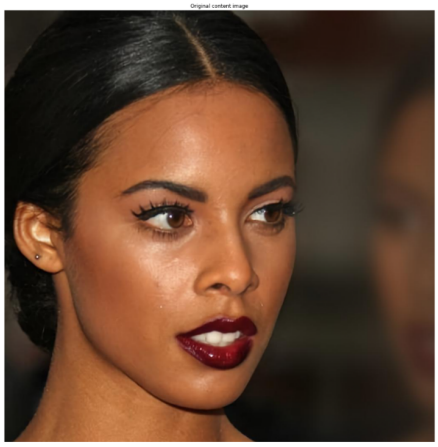
Results

Model Evaluation Metrics:

Dataset	FID (latent)	LPIPS (latent)	FID (reference)	LPIPS (reference)	Elapsed time
`celeba-hq`	13.73 ± 0.06	0.4515 ± 0.0006	23.84 ± 0.03	0.3880 ± 0.0001	49min 51s

Face Translation Output Matrix:

Fast Style Transfer Output:



Discussion

For the evaluation of their model's performance on the CelebA-HQ dataset, ClovaAI reported metrics for both the Frechét Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS), from averages 10 measurements each using different latent vectors and reference images. Both the FID and LPIPS scores are measures of the distance between generated images and real images via different mathematical comparisons, so smaller scores mean the generated images are more similar to the real images. As seen in the table above, the averages for both of the FID and LPIPS scores aren't very low, but aren't very high either, indicating that the model has done a decent job at generating realistic images; the score would never be able to reach zero for either since the generated images are an interpolation of two different images, so in this case it is reasonable that they are a little above zero.

From the generated output matrix of faces, we can also visually see that the model and it's pretrained weights do well in extracting certain reference image features and applying them to the source image to create a semi-realistic new image. In general, the model will apply the reference image's background, hair style, and skin color to the source image, while preserving the source image's distinctive facial features and geometry– such as the face's angle of orientation, expression, distances between eyes, nose, and mouth, etc. It doesn't appear very biased towards any specific age/skin color/gender, and it's handling of hair is especially impressive, seeing that it is able to generate the intricacies of hair roots and flow in a realistic manner.

However, while the model is able to generally generate realistic images, there are scenarios in which the model struggles; to focus on the areas of improvement of the model, its deficiencies are pronounced when there are physical augmentations to the face (e.g. glasses, earrings), when it needs to interpolate features hidden in the reference but present in the source (e.g. hair covering ears), and poor lighting condition in the source image. The model is very rigid when dealing with glasses– as it seems to just copy and paste the glasses onto the generated image– but is very liberal with earrings– as it seems to nonsensically decide to occasionally include earrings in generated images when they aren't present in the reference and occasionally exclude them when they are present. The issue with the earrings is especially present when hair covers the ears of a source image but not reference image or vise-versa, suggesting that the model has yet to learn how to reasonably deal with earrings. Similarly, the model also struggles with reconstructing eyebrows that are hidden by hair in either the source or reference, suggesting that it has yet to learn how to model eyebrows as well. Lastly, when the lighting from above causes a shadow under the chin (as can be seen in Rishab and my source images), the model will produce artifacts under the chins of the generated images. Nonetheless, its capabilities are impressive, and with further augmentations, can be even more robust to the varieties of input images it may see.

Fast Neural Style Transfer doesn't transfer facial features, but instead transfers the overall image style - which when tested on faces results in an output image that contains multiple repetitions of the reference face within the source face. StarGAN has been trained to specifically

transfer facial features and is thus much better at performing facial feature transfer. Fast Neural Style transfer works very well when trying to make an image look like another image - for example transforming images to look like a specific painting from an artist, or the same color scheme as another image and so on. It isn't suited for specific tasks.

Questions

- Which styles does the network work on? Which fail?
    - The network seems to work on older and primarily white individuals. The network doesn't do very well with glasses and earrings, also bad lighting seems to create a double chin.
- Does the method work equally well on all faces?
    - No, it doesn't.
- What are good quantitative metrics for evaluating this?
    - FID and LPIPS are two metrics used by the authors of the paper to evaluate performance of the network. FID compares the distribution of generated images with the distribution of real images that were used to train the generator, while LPIPS evaluates the distance between image patches - higher means more different and lower means more similar.

Conclusion

To conclude, we have analyzed the performance of StarGAN and Fast Neural Style Transfer on the task of transferring facial features from the reference image to the source image. StarGAN performs much better than Fast Neural Style Transfer simply because the latter is trained to work on transferring general image style and not face specific features. StarGAN, while much better at facial style transfer has its limitations that we discuss in the discussion above.

Credits and Bibliography

https://github.com/clovaai/stargan-v2
https://arxiv.org/pdf/1912.01865.pdf
https://arxiv.org/pdf/1706.08500.pdf
https://arxiv.org/pdf/1801.03924.pdf
https://nealjean.com/ml/frechet-inception-distance/