

Manipulação de Atributos Faciais

Rafael de S. Toledo

rafaeltol@gmail.com

Universidade Federal de Santa Catarina

Maio 2021

Motivação

- Há um número limitado de poses e atributos que recordamos um rosto em imagens.
- Explorar a síntese de um rosto com n possíveis atributos permite edição e ampliação de base de dados.
- *Deep Learning* já mostrou seu grande poder de síntese de imagens artificiais [1, 4, 6]. Uma de suas principais vertentes para esse fim é o **Autocondificador Variacional (VAE)** [2].
- Demonstrar a habilidade pra isolar e manipular características de uma cena de forma **não supervisionada**, seja facial ou não.

Autocodificadores Variacionais

Autocodificadores Variacionais têm a arquitetura similar de autocodificadores comuns, exceto que o espaço latente modela uma Normal Multivariada Z .

De Z são amostrados valores dessa distribuição que será alimentado no decodificador.

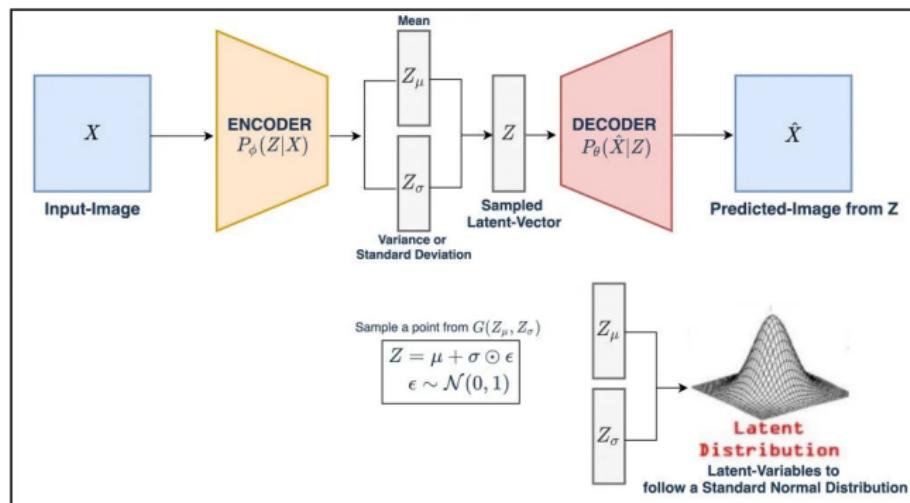


Figure 1: Diagrama tirado de [5].

CelebA

- Conjunto de dados *open-source* com 202k imagens de celebridades. [5].
- Rótulos binários da presença de 40 atributos faciais como sorriso, óculos, chapéu, barba, cabelo loiro e outros.
- Mais de 10 mil identidades distintas.

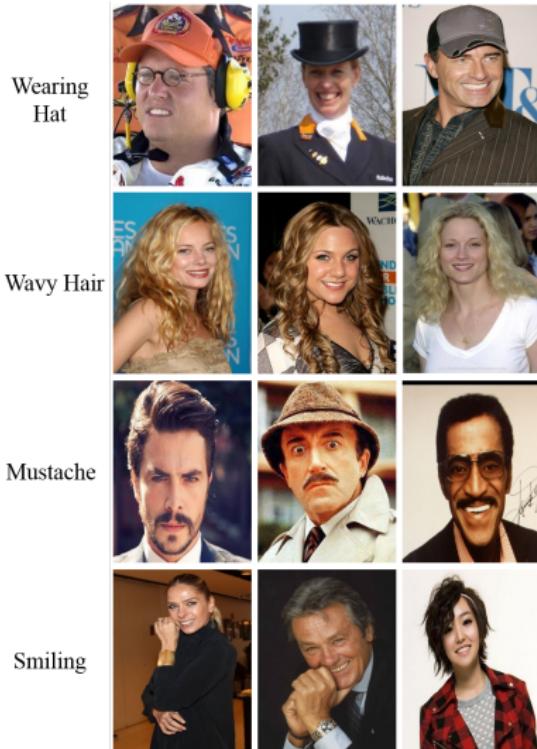
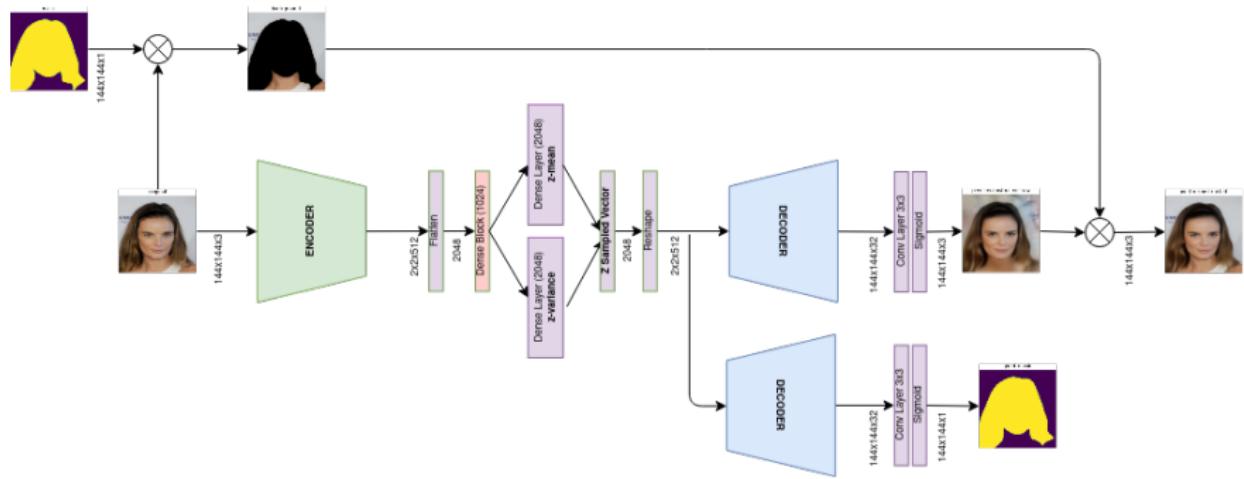


Figure 2: Imagem tirada de [3].

Arquitetura

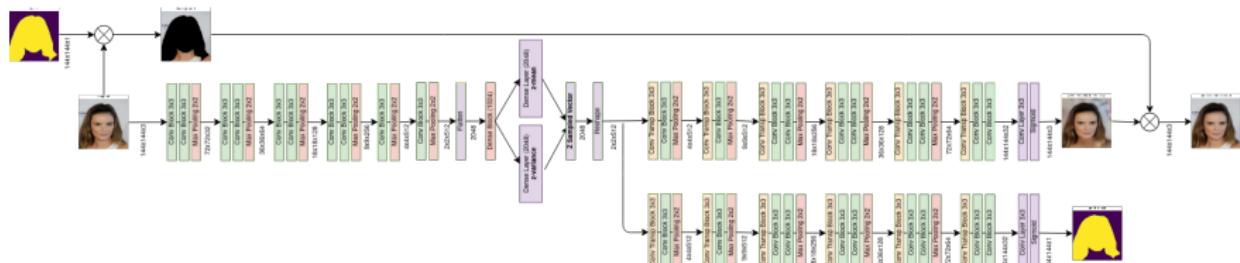
A solução proposta é um VAE com uma estrutura simétrica entre o codificador e o decodificador.

Há dois decodificadores: um gera a máscara do rosto mais cabelo e chapéu (caso haja), o outro reconstrói a imagem inteira.



Arquitetura Completa

Há 145 camadas na rede neural, sendo essas convolucionais, *batch normalization*, ativação, e densas. Somando 16.6M de parâmetros.



As funções perdas são divididas em três partes:

- ➊ Reconstrução da imagem: SSIM (*Structural Similarity Index Measure*) e MAE (*Mean Absolute Error*).
- ➋ Reconstrução da máscara: *Binary Cross Entropy* e *DICE loss*.
- ➌ Regularização do espaço latente Z: divergência de Kullback-leibler. Essa escalada no fator de 1/1000.

Treinamento

O treinamento teve 113 steps de 4000 iterações de *batch size* 32. Durando mais de 50 horas na Tesla P100 do Google Colab Pro.

Usou-se o otimizador Adam e taxa de aprendizado de 1e-4, cortando gradiente no máximo de 1e-3, e learning decay de 1/3 a cada 10 épocas até atingir o mínimo de 1e-5.

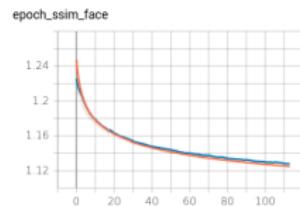


Figure 3: custo de 2 – SSIM da reconstrução da Imagem.

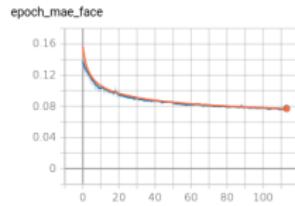


Figure 4: custo de MAE da reconstrução da Imagem.

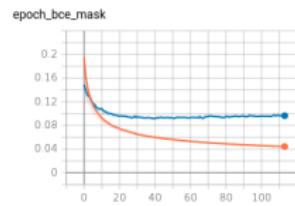


Figure 5: custo de BCE + Dice da reconstrução da Máscara.

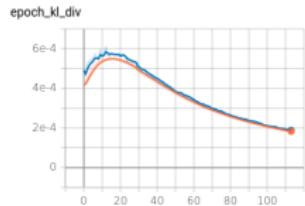


Figure 6: custo da divergência de Kullback-leibler.

Maior parte das funções custos apresentaram nenhuma variância, exceto a reconstrução da máscara.

Predição

Resultado da reconstrução nos conjuntos de treinamento e de validação.



Figure 7: Reconstrução no conjunto de treinamento.
Primeira coluna é imagem original, segunda coluna a reconstruída, depois a máscara original e a versão reconstruída.



Figure 8: Reconstrução no conjunto de validação.

- A performance é similar em ambos conjuntos.
- Partes com muito detalhes como fios de cabelo ou brincos acabam ficando borradadas.
- Preserva bem pose, cor, expressão.

Aritmética no espaço latente

Como observado em [3], o uso de aritmética no espaço Z para isolar um atributo quando tirando a média do resultado de uma larga amostragem funciona bem.

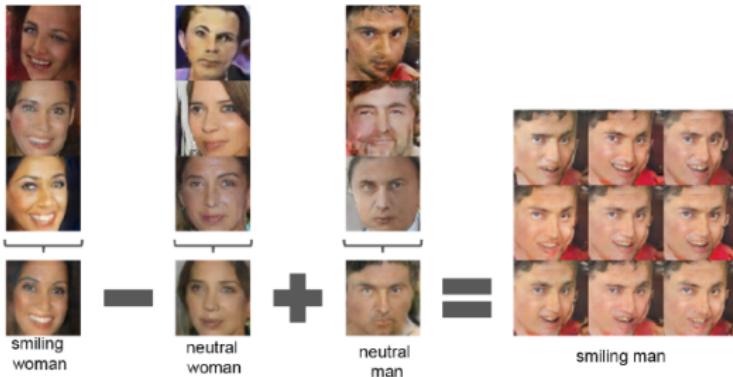


Figure 9: Exemplo da aritmética, imagem retirada de [3].

No trabalho, para cada atributo usei 10k de amostras de exemplos positivos e de exemplos negativos.

Exemplo 1



Figure 10: Espectro de atributos, percorrendo da escala negativa até a positiva do atributo, a primeira coluna é a imagem original.

Adicionando Barbas



Figure 11: Várias cores de barbas, primeira coluna é a foto original.

Adicionando Outros Atributos

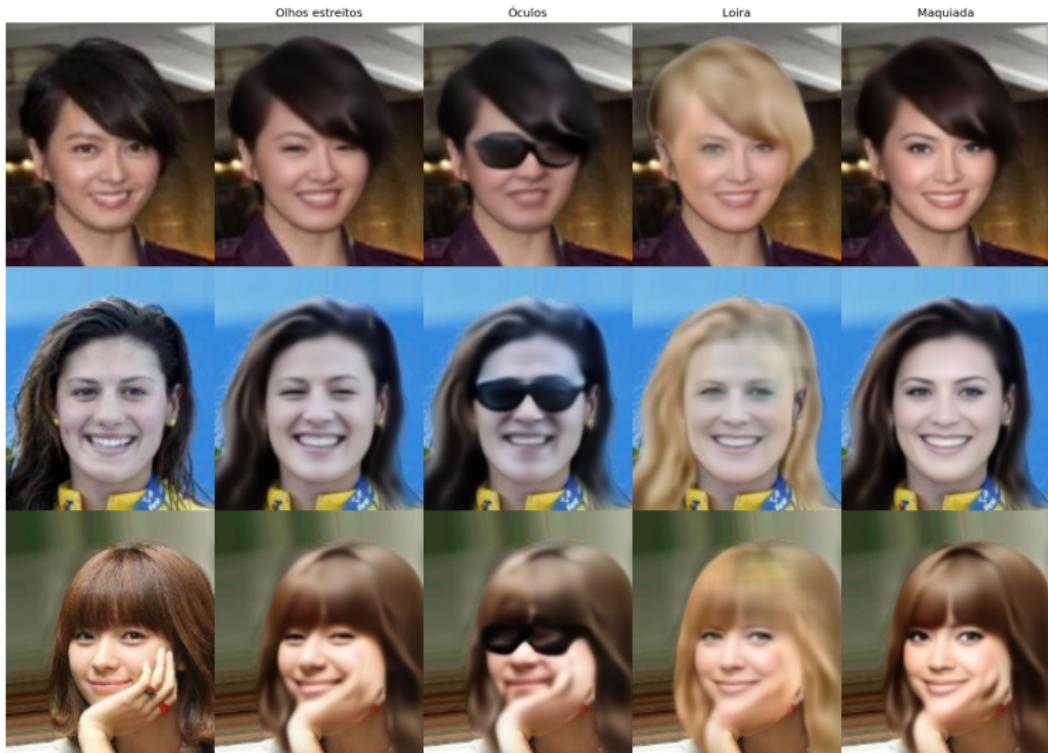


Figure 12: Primeira coluna é a foto original, seguido de Olhos estreitos, Óculos, Cabelos loiros e Maquiagem.

Adicionando Cabelo

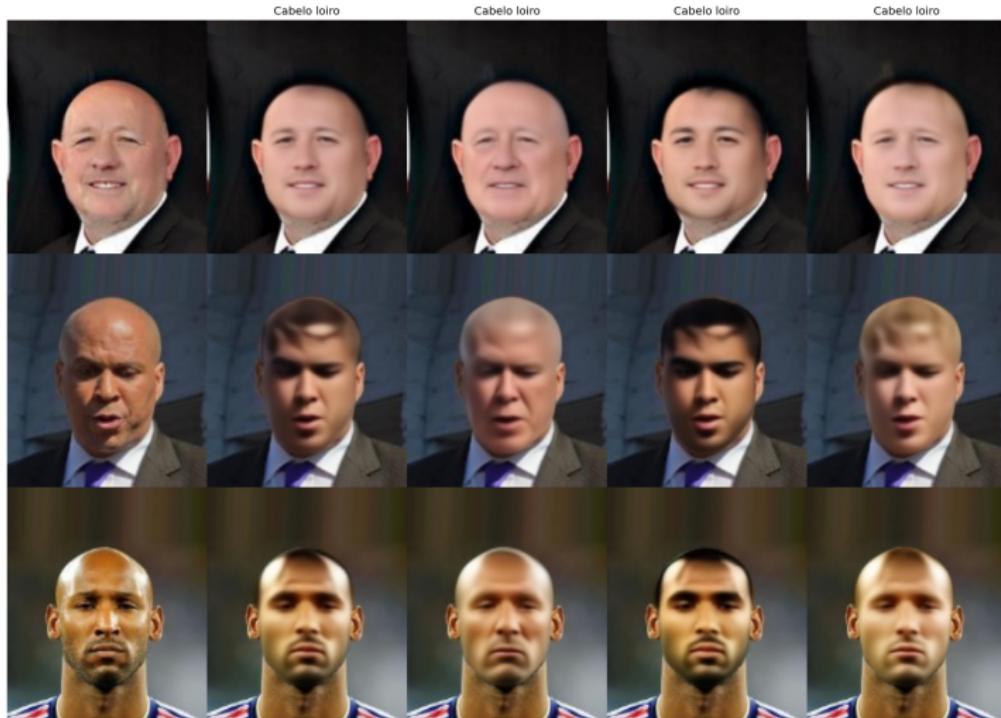


Figure 13: Várias cores de cabelo, primeira coluna é a foto original.

Conclusões

- Para a reconstrução da imagem, constatou-se pouquíssima presença de *variância*, não tendo diferença nos resultados entre os conjunto de treinamento ou de teste.
- Mostrou-se possível a adição de atributos específico dentro de uma figura sem corromper, com ressalvas, o resto da imagem original.
- A inserção ou alteração de atributos faciais tendem a se alinhar a pose a expressão do rosto original da foto.
- Detalhes sutis tendem a ficar borrados na reconstrução da imagem. Sendo um ponto de melhora para trabalhos futuros.
- Existe uma correlação de atributos, por exemplo o cabelo loiro estava correlacionado a óculos.

Referências

- [1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August, 15(2018):11*, 2018.
- [4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.
- [5] Aditya Sharma. Variational autoencoder in tensorflow, May 2021. URL <https://learnopencv.com/variational-autoencoder-in-tensorflow/>. Last accessed 15 May 2021.
- [6] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.