Variational Style Transfer

Jan Schopohl Dominik Fuchsgruber

Abstract

We propose a variational approach to neural style transfer, where the style of an artistic image is transferred to another. By using variational autoencoders to encode the style of an image, the latent style space is enforced to be smooth. This allows for better interpolation between different styles as well as sampling new unseen styles from the style space.

1. Introduction

The task of artistic style transfer aims to extract the style of an (artistic) image and transfer it to the content of another picture. As has been shown in [2], using an encoder-decoder architecture, where the decoder layers are conditioned on a style image using using *Adaptive Instance Normalization* (AdaIn), can achieve visually pleasant results. Instead of using the same encoder module for style and content image, we follow Kotovenko et al. [4] and make use of distinct encoders for content and style images.

Furthermore, we extend their approach by using a variational autoencoder architecture [3] to encode the style from an image. This enforces smoothness to the latent style space, which allows for better interpolation between different style embeddings. Additionally, we are able to draw samples from the latent style space, opening up a generative perspective to our approach.

2. Related Work

Artistic style transfer was introduced by Gatys et al. in [1], where the authors noticed, that the style of an image is captured by the gram matrices of different feature maps using the space of a VGG network. Optimizing a random noise image to minimize both content and style loss leads to a visually convincing stylization. The heavy computational burdens of this approach led to the development of fully feed-forward architectures for style transfer. An autoencoder-based approach that is applicable to transfer arbitrary styles unseen by the model is introduced in [2]. They use *Adaptive Instance Normalization*, which manipulates the instance-wise statistics of a content image before each decoder layer. This method was expanded upon by Kotovenko et al. [4], who proposed to incorporate two distinct encoders for content and style images.

Adaptive Instance Normalization is also used to approach different problems, one of which being synthesizing photo-realistic video sequences of speech expressions for a particular individual [6]. They use AdaIn to condition a decoder module on an embedding of the facial characteristics of a person. In our project, we heavily rely on the architecture of their approach, namely residual blocks both in the encoders and the decoder as well as bottleneck layers between the two. We extend this model by introducing U-Net [5] connections between the content encoder and the decoder.

The concept of encoding images as distributions rather than just simple points was first used in [3]. By sampling from those distributions during training, variations in the latent space can be captured, which leads to a smoother manifold. Furthermore, the distributions provided by the encoder are regularized to resemble a standard normal, which allows for sampling from the latent space as well.

3. Our Approach

Our model consists of three parts: Two distinct encoders for content and style images as well as a decoder that outputs stylizations.

- The *content encoder* E_c takes a content image c and applies a series of convolutions and poolings to it. The embeddings still retain a spatial extent, in contrast to just being flat vectors.
- The *style encoder* $E_s(s)$ is designed analogously to the content encoder, with the only difference being the use of adaptive pooling after the last convolution. Therefore, the style encoder yields a fixed-size embedding distribution by outputting the mean and variance of a normal distribution.
- The *decoder* mirrors the design of the two encoders. It is preceded by several bottleneck layers, which do not change the resolution of the input image. Both the bottleneck layers as well as the upsamling convolutions are conditioned on the style embedding using *Adaptive Instance Normalization*. Additionally, we use U-Net

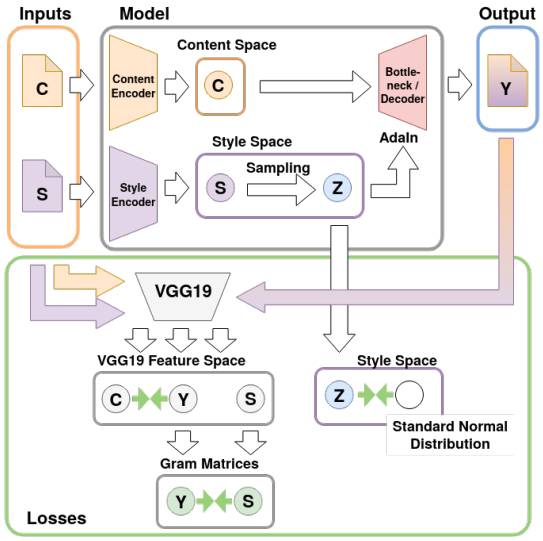

Figure 1. Visualization of the different components in our architecture as well as how the loss terms are implemented. Perceptual and style losses are computed in the VGG19 feature space.

connections that connect the output of content encoder layers to the respective components of the decoder.

3.1. Losses

Our network is trained with respect to three different objectives, two of which are computed in the features space of a pre-trained VGG19 network. Let $VGG^{\{i\}}$ denote the feature maps of the i-th layer and $G(x)$ the gram matrix.

- The content loss $\mathcal{L}_{content}$ penalizes deviations of the stylizations regarding the content image and is implemented as a perceptual loss. That is, we consider the feature maps of the $relu_{4,2}$ layer of the VGG19 network.
- The style loss $\mathcal{L}_{style}$ penalizes deviations from the style image by comparing the gram matrices of different VGG19 activation maps. We include all layers up to $relu_{4,2}$ into the style loss term.
- The variational loss $\mathcal{L}_{KL}$ regularizes the latent space to resemble a standard normal distribution. It uses the Kullback-Leibler divergence to compare the two distributions.

$$\mathcal{L}_{content} = \sum_i \|VGG^{\{i\}}(c) - VGG^{\{i\}}(y)\|_2^2$$

$$\mathcal{L}_{style} = \sum_i \|G(VGG^{\{i\}}(c)) - G(VGG^{\{i\}}(y))\|_2^2$$

$$\mathcal{L}_{KL} = KL(\mathcal{N}(E_s(s)_{:D_s}, E_s(s)_{D_s:}) \| \mathcal{N}(0, I))$$

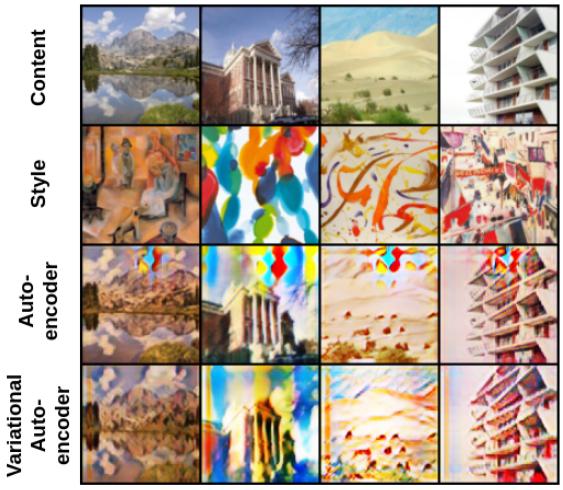

Figure 2. Stylizations of Non-Variational Autoencoders and Variational Autoencoders

The content is denoted with c, the style image with s and the stylization with y. We assign different weights λ to each summand of the content and style loss terms, as well as the variational loss.

3.2. Datasets

We use images from the *Places365* dataset [7], containing 1.8 million scenic photographs as content images. A subset of 100 images from the *WikiArt* dataset is selected that exhibits a broad range of abstract and artistic styles. We explicitly exclude realistic paintings as they do not contribute to the learning of styles. For both datasets, images are downsampled and cropped to a resolution of 96x96 pixels.

4. Results

All the results we report are with respect to content images from a test dataset, while the style images are taken from the training data. This is due to the fact, that we observed our model to be unable to generalize to unseen styles beyond just simple color transformations. We conjecture that this is due to the model complexity being to low. Because we lack the necessary GPU resources however, we were not able to verify this hypothesis.

4.1. Variational and Non-Variational Autoencoders

First, we compare stylizations obtained by a model that uses a non-variational style encoder and compare it with the outputs of our proposed method. As can be seen in Figure 2, we find that the introduction of sampling can prevent artifacts in the output image, as it also imposes regularization on the model itself. Furthermore, we find the overall quality of the variational approach to be visually more pleasing.

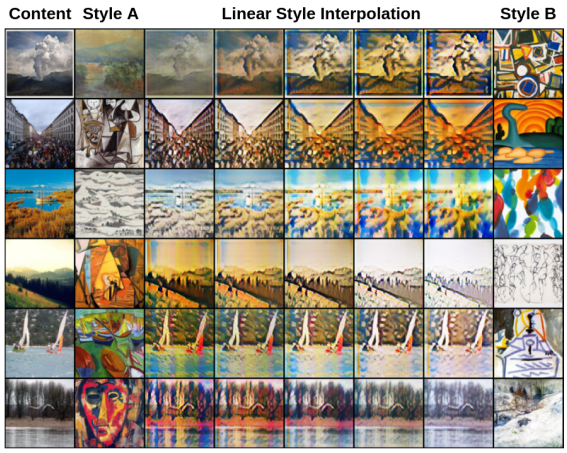

Figure 3. Interpolating between two different styles

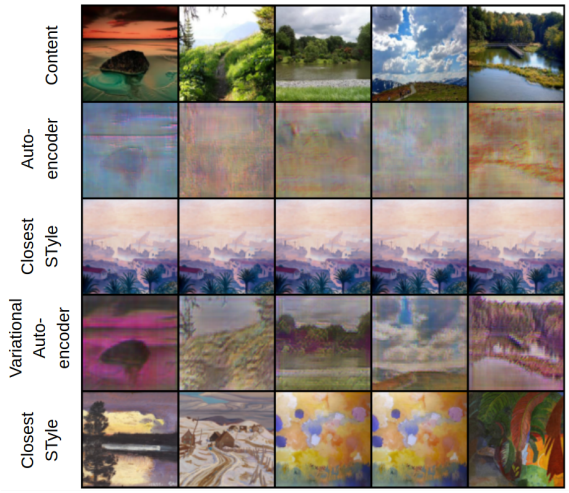

Figure 4. Sampling random styles and comparing with the style image most similar to the sample in style space

4.2. Style Interpolation

Figure 3 shows, that the smoothness of the latent space allows for a meaningful interpolation between two styles. That is in particular noteworthy, as our model has highly overfitted to 100 style images and nonetheless is able to provide appealing stylizations that mix two different style images. That means, that even a model with no generalization capabilities at least yields a smooth latent style space containing the styles it was trained on.

4.3. Style Sampling

Novel styles can be sampled by drawing random vectors from a standard normal distribution and conditioning the decoder on those. Figure 4 visualizes stylizations with randomly sampled style embeddings. It can be observed, that the variational approach visually outperforms the plain autoencoder architecture. In order to verify that the styles

Figure 5. Deception rate among non-experts for non-variational autoencoders and our approach.

are indeed novel and unrelated to the style embeddings the model learned during training, we also show the style image that is most similar to the sampled style with respect to their latent representation.

4.4. Non-Expert Deception Rate

We also conducted a survey among non-experts that were shown real artworks, stylizations provided by a non-variational autoencoder and outputs of our approach. Among five exemplars of each category the participants had to decide if an image they were presented was real or artificially generated. The images were shuffled before, and the participants were given a hard time-limit of five seconds to come up with a decision.

Figure 5 shows the results of our survey. While in general both approaches seem to perform similarly well, we observe that participants in general seemed to prefer non-abstract styles. Even though excluded from the figure, even in the case of real artworks low scores were recorded for abstract paintings. Nonetheless, two stylizations yielded by our approach outperform all five stylizations of the non-variational autoencoder. Not considering abstract styles - taking into account none of the participants had an arts-related background - this shows the superiority of our approach.

5. Conclusion and Future Work

We propose an approach to artistic style transfer that embeds artworks into a smooth latent space making use of variational autoencoders. We show that this method allows for a smooth interpolation between different stylizations and also makes it possible to sample novel styles unseen by the model. Furthermore, introducing a variational prior on the style manifold imposes a regularization on the model that reduces artifacts in the output.

Due to the lack of resources we were unable to train a model with generalization capabilities beyond simple color transformations. However, comparing our approach with related work, we conjecture that a more complex model might in fact to overcome this limitation.

References

[1] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[2] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[4] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[6] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *CoRR*, abs/1905.08233, 2019.

[7] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014.