Neural Style Transfer using Variational Auto-Encoders

Dominik Fuchsgruber, Jan Schopohl

November 17, 2019

I. Introduction

Neural Style transfer describes the task of extracting the style information of a style image y and applying it to a content image x, in order to obtain a stylized version of x. Finding two disentangled representations of the content and style of an image is a crucial ingredient for models that are capable of describing various styles while providing visually appealing results. Following closely the approach of [6], a content representation can be obtained using an auto-encoder architecture, where the decoder D is additionally conditioned on a style embedding, which is produced as the output of a second style encoder network. We propose to make use of variational auto-encoders [5] instead in order to enforce the model to learn a smooth latent style space, which aims at improving interpolation between different artistic styles.

II. Proposal

The proposed model is a direct extension of the model elaborated in [6]. Instead of using a style encoder E_s network to output a single style representation of an image, our approach outputs the mean and variance of a Gaussian distribution, that captures variation in the given style. The decoder network of [6] is conditioned on a style representation sampled from the distribution obtained by the modified style encoding network. Because of the lack of computational resources and to limit the scope of the project, we will not employ an adversarial setting to train the generator model, in contrast to [6]. Furthermore, we propose replacing the pixel-

and content-loss of [6] with a perceptual loss [3]. This loss can be implemented by considering the first few layers of a pre-trained model such as VGG-19 and calculating the $L2$ distances of the feature activations of the input image x and its stylized counterpart $G(x, z)$, where $G(x, z) = D(E(x), z)$ represents the output of the decoder network D given a content image and style representation. This loss ensures that the content representation of x is preserved.

To enforce a similar style in the output as the style input image y, in contrast to [6], we propose using a perceptual loss for the style as well. To calculate the loss, the Gram matrices of features activation maps of a pretrained VGG-19 network between the input image x and the stylized output $G(x, z)$ can be compared [3].

Following [6], to minimize the influence of content information in the style representation, we adopt the fixpoint-disentanglement loss:

$$\mathcal{L}_{FPD} = \mathop{\mathbb{E}}_{z \sim E_s(y)} \max(0,$$
$$\|E_s(G(x_1, z)) - E_s(G(x_2, z))\|_2^2 - \qquad (1)$$
$$\|E_s(G(x_1, z)) - z\|_2^2)$$

where E_c and E_s describe the content and style decoder networks. The fixpoint disentanglement loss ensures that the discrepancy between two stylizations is smaller than those of a stylization and the style image in the style space. Thus, it tries to prevent the content of the style image influencing the stylization.

As is standard for variational auto-encoders, the loss function given in equation (2) enforces the style distribution space to resemble a standard normal distribution, since we set $p(z)$ as such.

$$\mathcal{L}_{KL} = \mathbb{KL}(E_s(y)\|p(z)) \qquad (2)$$

III. Dataset and Model

We propose to use the same dataset that has been used by [6], namely the places365 dataset [9], as a source of content images x and the Wikiart dataset [4] to obtain several artistic style images y. In contrast to [6], we will however not group images by the same artist together as sharing a style, as this is not required for the loss functions included in our setting and may also be an invalid assumption on its own, as artistic styles may vary over time for even a single artist. Considering our limited resources, we plan on downsampling the images to a 64x64 resolution to lower the computational burden.

While we would like to also adapt the model architecture directly from [6] and most likely scale it down to fit our setting, the source code of their model has not been published to the current day and the supplementary material does not describe the architecture used in-depth. Thus, we plan on implementing the structure of a (shallow) VGG or ResNet architecture for the content and style encoders E_c and E_s respectively. The decoder architecture follows a similar structure, that should mirror the encoder networks, and the conditioning on the style is achieved by replacing the affine parameters of the instance normalization with the values obtained from the z it was conditioned on.

In order to calculate the perceptual content and style losses, we propose to use the first few layers of a pre-trained model like VGG-19.

IV. Project Plan and Milestones

We plan on successively building our architecture starting from scratch. That is, the first step is to train a fully functional standard autoencoder architecture that is able to faithfully reconstruct any image given its content representation $E_c(x)$. Keeping these model parts fixed, we will implement the style encoder $E_s(y)$ as a standard encoder, training it solely on the losses proposed above. Lastly, we propose to extend the model by replacing $E_s(y)$ with a variational version and thus complete the proposed setup.

Evaluating the results our model provides quantitatively proved to be quite a challenging task in neural style transfer. Thus, we will rely on qualitative metrics only. That is, we plan on conducting a small survey including for example friends or chair members to come up with a deception rate value, since we most likely can not rely on art experts to participate. We also plan on showing results on the enhanced style interpolation capabilities of our model.

Thus, we come up with the following milestones:

- Implement and train a standard auto-encoder
- Implement a style encoder and conditioning on the decoder network
- Implement the variational style encoder
- Perform a qualitative survey on the results

V. Related Work

The concept of neural style transfer was first introduced by Gatys et al. in [1], where however optimization was performed on the image directly resulting in the lack of real-time capabilities. Follow-up works introduced Adaptive Instance Normalization as a concept, where the affine parameters of the instance normalization layers are viewed as being able to capture the style of an image and thus are transferred from the style to the content image utilizing autoencoders as well [2]. As their model is a fully feed forward pipeline, it does not suffer from heavy computational requirements. Achieving style transfer by applying a transformation to a hidden representation of an image was recently also used by [7], where linear transformation is learned. Furthermore, Li et al. [7] also train their architecture using perceptual losses for content and style, which we also plan on utilizing. The VGG-19 that we intend to evaluate for retrieval of those losses was introduced in [8]. Kotovenko et al. proposed a framework which

2

explicitly tries to find disentangled representations for the content and style of an image by introducing a fixpoint disentanglement loss. Their work [6] is the basis for our proposed approach. Additionally, the concept of variational auto-encoders, which we also rely on, was first introduced in [5].

References

[1] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. "A Neural Algorithm of Artistic Style". In: *CoRR* abs/1508.06576 (2015). arXiv: 1508.06576. URL: http://arxiv.org/abs/1508.06576.

[2] Xun Huang and Serge J. Belongie. "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization". In: *CoRR* abs/1703.06868 (2017). arXiv: 1703.06868. URL: http://arxiv.org/abs/1703.06868.

[3] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution". In: *CoRR* abs/1603.08155 (2016). arXiv: 1603.08155. URL: http://arxiv.org/abs/1603.08155.

[4] Sergey Karayev et al. "Recognizing Image Style". In: *CoRR* abs/1311.3715 (2013). arXiv: 1311.3715. URL: http://arxiv.org/abs/1311.3715.

[5] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. arXiv: 1312.6114 [stat.ML].

[6] Dmytro Kotovenko et al. "Content and Style Disentanglement for Artistic Style Transfer". In: *The IEEE International Conference on Computer Vision (ICCV)*. 2019.

[7] Xueting Li et al. "Learning Linear Transformations for Fast Arbitrary Style Transfer". In: *CoRR* abs/1808.04537 (2018). arXiv: 1808.04537. URL: http://arxiv.org/abs/1808.04537.

[8] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv 1409.1556* (Sept. 2014).

[9] Bolei Zhou et al. "Learning Deep Features for Scene Recognition using Places Database". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 487–495. URL: http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf.