Neural Style Transfer using Variational Auto-Encoders

Dominik Fuchsgruber, Jan Schopohl

November 14, 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 [5], 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 [4] 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 [5]. 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 a Gaussian distribution, that captures variation in the given style. The decoder network of [5] is conditioned on a style representation sampled from the distribution obtained by the modified style encoding network. Also, 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 [5]. Instead, the network will be trained w.r.t. to the following loss functions.

$$\mathcal{L}_{FP-content} = \mathop{\mathbb{E}}_{z\sim E_s(y)} \|E_c(x) - E_c(G(x,z))\|_2^2 \tag{1}$$

$$\begin{aligned}\mathcal{L}_{FPT-style} = \mathop{\mathbb{E}}_{\substack{z_1\sim E_s(y_1)\\ z_2\sim E_s(y_2)}} \max(0, r+ \\ \|E_s(y_1) - E_s(G(x,z_1))\|_2^2 - \\ \|E_s(y_1) - E_s(G(x,z_2))\|_2^2)\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{FPD} = \mathop{\mathbb{E}}_{z\sim E_s(y)} \max(0, \\ \|E_s(G(x_1,z)) - E_s(G(x_2,z))\|_2^2 - \\ \|E_s(G(x_1,z)) - z\|_2^2)\end{aligned} \tag{3}$$

$$\mathcal{L}_{KL} = \mathbb{KL}(E_s(y)\|p(z)) \tag{4}$$

where E_c and E_s describe the content and style decoder networks, and $G(x,z) = D(E(x),z)$ represents the output of the decoder network D given a content image and style representation. While the fixpoint content loss (1) ensures that the content representation of x is preserved, the fixpoint triplet style loss (2) ensures that representations of different styles are far and those of similar styles are close in the style space. 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. Lastly, (4) enforces the style distribution space to resemble a standard normal distribution, since we set $p(z)$ as such. We additionally propose, to use a perceptual loss considering the first few layers of a pre-trained model (such as VGG) to replace the pixel loss of [5].

III. Dataset and Model

We propose to use the same dataset that has been used by [5], namely the places365 dataset [6] as a source of content images x and the Wikiart dataset [3] to obtain several artistic style images y. Considering our limited resources, we think of downsampling the images to a 64x64 resolution to leverage the computational burden however.

While we would like to also adapt the model architecture directly from [5] 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, 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 loss that ensures that the content of the input image is preserved, we propose to use the first few layers of a pre-trained model like VGG-16 and calculate the $L2$ distances of the feature activations of the input image x and its stylized counterpart $G(x, z)$.

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 auto-encoder 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 on the losses given by [5]. 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 auto-encoders as well [2]. As their model is a fully feed forward pipeline, it does not suffer from heavy computational requirements. Kotovenko et. al. proposed a framework which explicitly tries to find disentangled representations for the content and style of an image by introducing a fixpoint disentanglement loss. Their work [5] is the basis for our proposed approach. Additionally, the concept of variational auto-encoders, which we also rely on, was first introduced in [4].

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

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

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

[6] 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.