Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230693

Multi-scaled attentive style transfer based on dilated convolution

Xuefeng Xu[1], Tuo Chen[2,4], Huanyi Chen[3]

[1]The High school affiliated to Renmin University of China, Beijing, 100080, China
[2]School of Software,Xinjiang University, Urumqi, 830046, China
[3]Shanghai New Channel JinQiu education, Shanghai, 200003, China

[4]18310223819@163.com
All the authors contributed equally to this work and should be considered as co-first author.

Abstract: Image style transfer aims to apply artists' painting styles to various images. Many approaches seek different purposes, but a general tendency is to increase efficiency and enable arbitrary style inputs. The state-of-art method is the adaptive convolutional network which expands the process of feature mixture into a layer-wise adjustment that superior previous work by presenting results that is more aware of detailed structures. However, the encoding process that guides the entire stylized revision is unaware of multi-scaled information. We designed an improved version of the style feature encoding procedure in our work. With the introduction of dilated convolution with a different rate, the output stylized image is better at spatial texture migration and color distribution determination. We also come up with a hybrid objective that better measures the spatial dissimilarity between content and style features.

Keywords: Style Transfer, Dilated Convolution, Multi-Scale Feature Extraction.

1. Introduction

In machine learning, a Convolutional Neural Network (CNN) is a deep feed-forward artificial neural network. The basic structure of a CNN consists of two layers: a feature extraction layer and a feature mapping layer. Convolutional neural networks are uniquely suited to image processing, image recognition and image segmentation due to their special structure of local weight sharing [1]. The weight sharing reduces the complexity of the network, and images with multi-dimensional input vectors can be fed directly into the network, avoiding complex data reconstruction.

The prevalent method for manipulating style is through adaptive instance normalization (AdaIN) [2], a approach that convert the mean and variance of image characteristic. However, AdaIN is a universal procedure and the local geometric structure in the style image is often ignored during migration. So DisneyResearch proposed Adaptive Convolution (AdaConv) [2], a common extension of AdaIN that allows simultaneous transfer of statistical and structural styles. In addition to style migration, the methods in this paper can be easily spread to style-based image generation, as well as to other tasks that already employ AdaIN. In this study, DisneyResearch introduces an expansion to AdaIN called Adaptive Logarithms (AdaConv) [2], which allows it to accommodate both statistical and structural styles simultaneously. In the situation of style transfer, rather than transferring a pair of simple global statistics[3], our approach is not to transfer a pair of simple global statistics of each style feature, but to

DOI: 10.54254/2755-2721/6/20230693

estimate the full convolution kernel and bias value of the style image and then convolve it with the content image features.

This is a generic extension to AdaIN [4] that allows the contemporaneous transfer of anatomical and structural styles. In addition to style migration, DisneyResearch's approach can easily be extended to style-based image generation, as well as other tasks that already use AdaIN [4].

The pre-trained VGG19 convolutional model has faster convergence and better generalisation properties, and the convolutional kernel of the whole VGG19 network is more sensitive to the stimuli of contour lines and colour stimuli of an image.However, we think VGG19 feature extractor seems outdated, we assume using a stronger feature extractor will perform better on this task.In addition, we note that both style loss and content loss are calculated via Euclidean distances, and we think that cosine distances would be better, and that designing a new target seems better than simply adding style loss and content loss together. Furthermore, we believe that the AdaConv [2] model does not incorporate Attention layers, and we will try to incorporate Attention layers so that it can convey richer information.

2. Related Work

CNN-based neural style transfer was originally proposed by Gatys et al. [5]. Although their scheme permitted arbitrary style transfer between images, it was based on a slack optimisation process. Johnson et al. solved this problem by introducing perceptual loss loss, making the optimisation much faster and achieving real-time results. At the same time, Ulyanov et al. suggested methods to improve the quality and variety of the generated examples. However. The limitation of the feed-forward methods described above is that each network is bound to a fixed style. To address this problem, Dumoulin et al. introduced a single network that is capable of encoding 32 styles and their interpolations a feedforward architecture was proposed by Li et al [6]. This architecture can synthesise 300 textures and transmit 16 styles. However, both of these methods cannot accommodate arbitrary styles that were not observed during training.

Various methods to control the output of the style transformation have been proposed. Among them, Gatys et al. suggest two universal manage way which involve the whole output, instead of a specific space region. One approach is to decompose the picture into hue, SAT [7] and brightness and to stylise only the brightness. The luminance is stylised to preserve the palette of the content picture.Gatys second way is to generate a secondary style picture to conserve the colour, or to transfer the style only from a specific scale. These types of user dominate are orthogonal to our way and can be consolidated [7].

The other type of dominate is a spatial control that allows the user to ensure that certain areas of the output should be stylised using only the manually choose style image area of the feature style image. In Gatys' approach [5], his propose some form of spatial dominate based on the user defining matching regions of the image by creating a dense mask for the style and content images.

3. Motivation

The AdaConv model [2] processed the data stream of paired content and style images separately. The backbone of the model is the downsampling and upsampling of content images, with no style normalization to content features at the latent space. Besides content reconstruction, the model views high-level style features and respectfully adjusts the up-sampling process by predicting adaptive convolutional blocks. The AdaConv block includes spatial and depth-wise convolution with channel-wise bias, which receives weight from the kernel predictor. The design of the global style encoder caters to kernel prediction by further transforming style features to style embeddings. It consists of three successive average pool operations with kernel size (2,2) for down-sampling and a dense layer to predict style embeddings. The original feature is scaled eight times to filter helpful information. Another purpose is to decrease the in-feature size for the dense layer, significantly saving the entire model size.

However, directly reducing the size of feature maps has potential risks of missing valuable information. Although a convolutional kernel is embedded to aid feature reduction, it is insufficient to obtain multi-scaled features with a fixed kernel size. Therefore, we assume that the style encoder could be improved for more robust feature down-sampling.

DOI: 10.54254/2755-2721/6/20230693

After comparing AdaConv's method [2] with other compatible style transfer approaches, we come up with three possible improving directions:

3.1. Learning Objective

Style loss and content loss are all computed via Euclidean distance; we assume REMD loss for style-loss computing will be the better option. Besides, designing a new objective seems better than adding style loss and content loss [8].

3.2. Attention mechanism

Attention blocks tell the model which part to pay attention to, which is widely used in other computer vision models, such as image segmentation and super-resolution. We assume that having a model self-aware of the vital part will enhance its performance. Therefore, we sought a way to produce a weighted mask that tells the model which part of the style feature is essential.

3.3. Dilated convolution

Regular convolution has a successive receptive field; dilated convolution pined holes on a kernel which sparse the receptive field. It has been used in various object-detection models for obtaining multi-scaled features. Our work uses the convolution kernel with different dilation rates for predicting an attention mask.

4. Experiment

We use two different loss functions to test the first assumption. REMD loss and moment-matching loss. They are used respectfully during the training process.

We redesigned the global style encoder to a hybrid network with different procedures to predict attention masks to test the second assumption. The original method is that three consecutive down samples are directly connected to the dense layer. However, we split the single down-sampling channel into a two-phase prediction. We downsample the original features in two parallel processes, each consisting of a similar kernel, to predict a mask with the dilated convolution of different rates simultaneously. Output from dilated kernels is passed to the Sigmoid function to produce a pixel-wise weight. The attention mask multiplies down-sampled features to make the final style embeddings. Various scaled information is extracted in this process by different dilated receptive fields.

We, therefore, choose the AdaConv model with single stride and single scale as the benchmark model, and the loss function is the moment-matching loss l_m. We also trained our model with a refined style encoder under equal optimizer parameters. Feature extractors are both VGG19 net with a scaling rate of eight, producing feature maps with shape (512,32,32) at the latent space. Kernel predictors also remain unchanged, yielding adaptive convolutional blocks inserted before the upsampling layer.

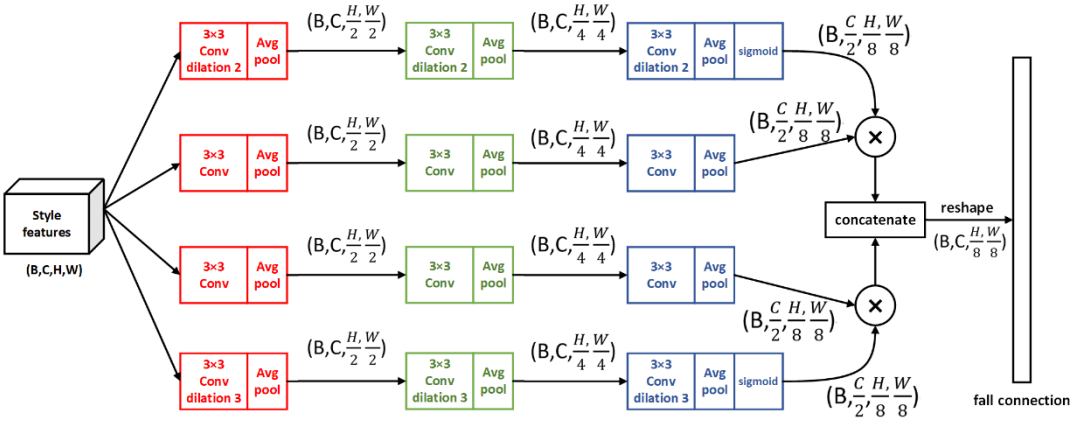

Figure 1. The structure of our proposed module.

Inspired by the objective function in STROTSS [8], we add the REMD loss between style and content

DOI: 10.54254/2755-2721/6/20230693

features for measuring l_s. The exact process is also computed between stylized and style images producing self-dissimilarity measurement l_c. Therefore, we rearranged the style loss as follows:

$$l_{style} = \frac{\alpha * l_m + l_c + \frac{1}{\alpha} * l_s}{1 + \alpha + \frac{1}{a}}$$

For content loss, Euclidean distance performed better in this model compared to cosine loss after conducting experiments. Therefore, we utilize MSE loss to measure content reconstruction [9].

The content images are COCO datasets with real-life scenes and objects, and the style images are WikiArt datasets with art paintings from different periods. For each iteration, our data loaders create random pairs of content and style images equal to batch size [10].

5. Results

We combined ten content images with ten style images to test our model's stylization performance. We randomly choose five stylization images to compare the results.

As shown below, our model is more aware of the feature generalization in style images. With the introduction of the attention mechanism, our refined style encoder has done a better job conserving the detailed structures of content than in the baseline model. For example, the facial organs is well preserved in our results. For objective function, measuring the self-dissimilarity weakens style migration on useless spots but enhances generalized texture migration [11, 12].

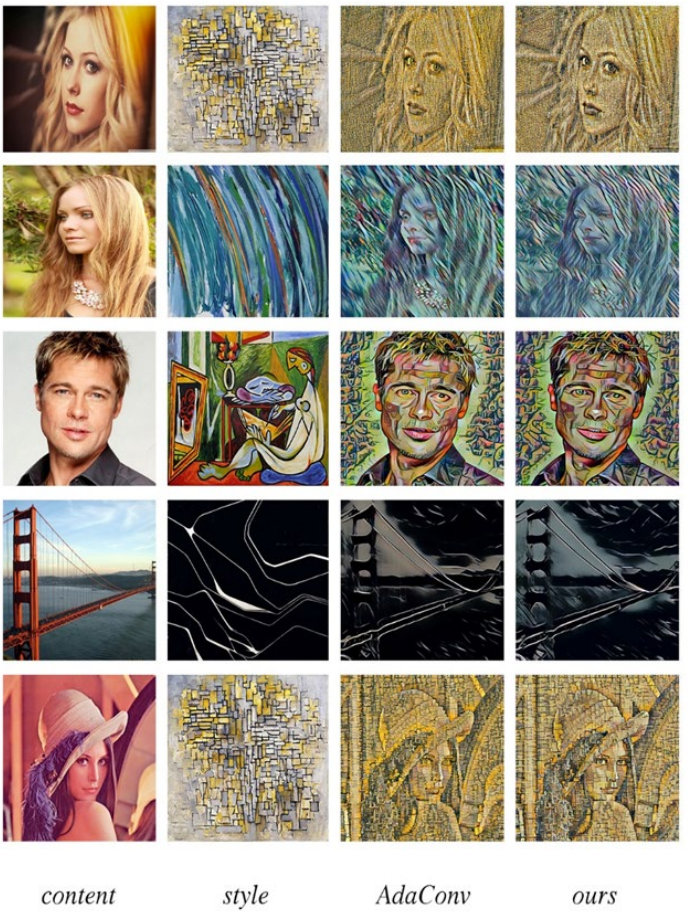

Figure 2. Comparison with AdaConv methods in arbitrary image style transfer.

DOI: 10.54254/2755-2721/6/20230693

6. Conclusion

As an improvement to the state-of-art model, our model does better in generalizing style textures and colors, with a more robust reference of style embeddings to reconstruct the content image. We also redesign the objective functions regarding self-dissimilarities to compute style losses. For future work, designing a more complex AdaConv block seems feasible for better performance. Also, switching the feature extractor with a more efficient architecture could reduce parameter size. For application concerns, adding a head-to-head CLIP model with image segmentation for generating a mask that limits area for style transfer may further improve users' experience.

Acknowledgement

All the authors contributed equally to this work and should be considered as co-first author.

References

[1] Liu S, Lin T, He D, et al. AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer[C]// 2021.

[2] Chandran P, Zoss G, Gotardo P, et al. Adaptive Convolutions for Structure-Aware Style Transfer[C]// Computer Vision and Pattern Recognition. IEEE, 2021.

[3] Tian Q C, Schmidt M. Fast Patch-based Style Transfer of Arbitrary Style[J]. 2016.

[4] Huang X, Belongie S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In CVPR, 2016.

[6] WANG Ting, LI Hang, HU Zhi. DESIGN AND IMPLEMENTATION OF IMAGE STYLE MIGRATION ALGORITHM BASED ON VGGNET [J]. Computer Applications and Software, 2019, 36(11):5.

[7] Park D Y, Lee K H. Arbitrary Style Transfer With Style-Attentional Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[8] Kolkin N, Salavon J, Shakhnarovich G. Style Transfer by Relaxed Optimal Transport and Self-Similarity[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[9] ZHANG Jinglei, HOU Yawei. Image-to-image Translation Based on Improved Cycle-consistent Generative Adversarial Network [J]. Journal of Electronics & Information Technology, 2020, 42(5):7.

[10] Zeng Xianhua, Lu Yuzhe, Tong Shiyue. Photorealism style transfer combining MRFs-based and gram-based features [J]. Journal of Nanjing University (Natural Sciences), 2021, 57(1):9.

[11] ZHANG Ying-tao, ZHANG Jie, ZHAN Rui. Photorealistic Style Transfer Guided by Global Information [J]. Computer Science, 2022, 49(7):6.

[12] LIU Hong-lin, SHUAI Ren-jun. Method of Fast Neural Style Transfer with Spatial Constraint[J]. Computer Science, 2019, 46(3):4.