Style Transfer for Videos with Audio

Gaurav Kabra[1] and Mahipal Jadeja[2]

Malaviya National Institute of Technology (MNIT), Jaipur, India
gauravkabra12@gmail.com[1]
mahipaljadeja.cse@mnit.ac.in[2]

Abstract. In the art of painting, from early era of beginning of human civilization, human beings have been creating artistic images with content from the real world but style from their imagination. Consider one such art called The Starry Night by painter van Gogh. Here the mountains, moon and houses are content taken from real world but the style of painting is totally from the painter's imagination and is unique. Style Transfer is the problem of taking content of an image and style of other image to create third image having content of the former but style of the latter. Clearly, such work can not be obtained by simple overlapping the two images. Until recently, due to not so optimized GPUs and slower hardware, image processing was a time consuming computation problem. But now we can use technological optimizations to use Convolutional Neural Networks (CNNs) to do the Style Transfer. In this paper, we will discuss how style transfer can be done in videos having audio in them. We shall also compare one of the existing methods with our implementation. Our proposed work has potential applications in the domains of social media communication, entertainment industry and mobile applications.

Keywords: Image Processing, Style transfer/filters, Video processing, Audio processing, Computer Vision (CV), Deep Convolutional Network (CNN), Deep painterly harmonization

1 Introduction

Consider an art image A and your photograph P. We want to construct image X such that X matches the style of A and content of P at the same time by optimizing the pixel values of X. This work is emerging topic in image processing due to the GPU and TPU optimizations possible only recently. Existing approaches focus on only video or only audio. Our work serves to stylize simultaneously both video and audio.

The work for video can be used to generate synthetic artworks/ artistic videos such as to give effects from famous painting and can be used in movies such as James Cameron's Avatar so that actors can shoot without makeup and later using computers, we can make them appear as if originally they had makeup. The work for audio can be used for creating a new audio from two input audio. For instance, background music can be modified by mixing style of other audio.

Many mobile applications may be developed using our approach. Most similar (yet not the same) application is "Voice Changer with Effects" where your voice may be changed to that of a robot, for example. It may extend its functionality in future updates by incorporating our audio work that will be stylizing background audio, not your voice.

The rest of the paper is organized as follows. In Section 2, related work is discussed along with explanation of key concepts of various types of losses during style transfer. In Section 3, methodology for the proposed approach is discussed including information about dataset and processing technique. Section 4 discusses about key results and implications. Our proposed method is compared with an existing method in the same section. The final section summarises our work and provides future directions for research.

2 Literature Review

To understand underlying style of the image is a very crucial aspect of computer vision but not it has gained a very little attention in terms of research [4] [7] [3]. Gatys el al [1] put forward first working methodology for style transfer in images. In his method, he used a type of CNN [10] called VGG-19 architecture that is pre-trained on ImageNet dataset of Stanford Vision Lab, Stanford University. This was used to extract content and style of image. He showed that it is possible to separate content and style of an image.He used average pooling instead of max-pooling as former was giving more appealing output. Moreover, his model did not have any fully connected (FC) layers. Other related approaches are discussed in [5] and audio specific ideas are discussed in [9].

2.1 Method for Images Style Transfer

We want to construct image X such that X matches the style of A and content of P simultaneously. We start by initializing X to some random values (called noise).

Content Loss: Let us select a hidden layer (l) in VGG-19 to calculate the content loss. Let p: original image and x generated image. Let P_l and F_l denote feature representations of the respective images corresponding to layer l. Then the content loss will be defined as: $\mathcal{L}_{\text{content}}(\text{p,x},l) = \frac{1}{2}\sum_{i,j}\left(F_{ij}^{l} - P_{ij}^{l}\right)^2$

Style Loss: Before understanding what style loss is we first need to understand what is style of an image. Intuitively it is texture information and local color scheme but not global arrangement. Mathematically, it is the correlation between filters at a given layer l. For example, in Figure 1 different filters (or channels or feature maps) at layer l have been shown.

Calculation of correlation between different filters/ channels involves the dot product between the vectorized feature maps i and j at layer l.The matrix thus obtained is called Gram Matrix (G). Two channels are correlated if and only if the value of dot-product across the activation of two filters is large.

$G_{ij}^{l} = \sum_{k} F_{ik}^{l} F_{jk}^{l}$

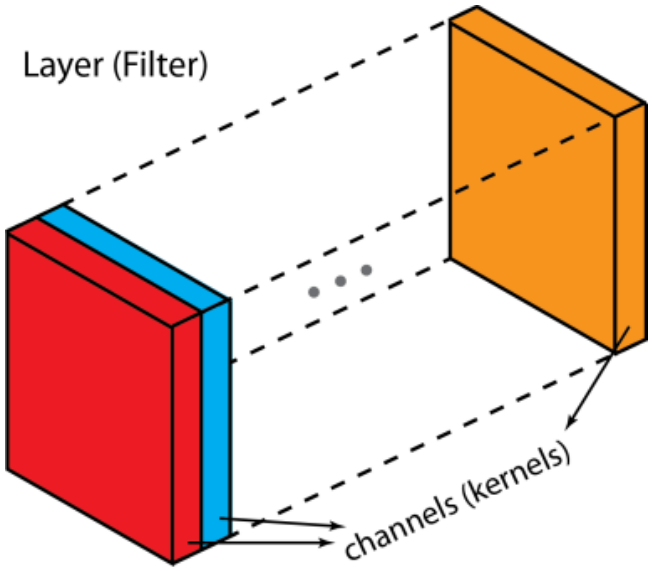

Fig. 1: Different filters at layer l

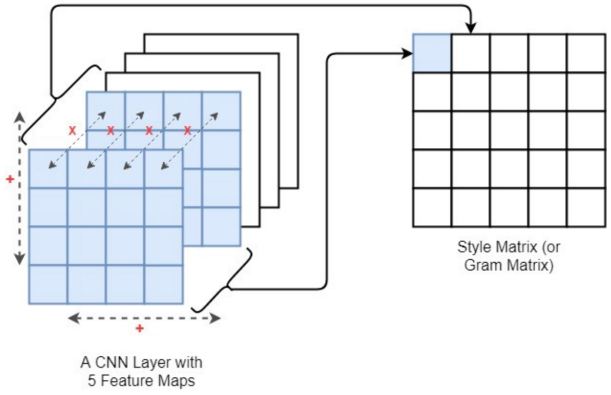

Fig. 2: Gram Matrix Calculation

Figure 2 shows a scheme for Gram Matrix calculation.

Style loss is the square of difference between the Gram Matrix of the style image with the Gram Matrix of generated Image. So let a: style image and x: generated image. Let A_l and G_l denote respective style representations in layer l. The contribution of layer l to total style loss is:
$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - A_{ij}^l\right)^2$
And hence the total style loss is:
$\mathcal{L}_{\text{style}}$ (a,x) $= \sum_{l=0}^{L} w_l E_l$ where w_l are corresponding to weighting factors of the contribution of each layer to the total loss. The process of optimizing pixel values includes two tasks at the same time - minimize both the content loss($L_{content}$) and style loss(L_{style}) by using backpropagation.

So the loss function to be minimized turns out to be:
$L_{total}(P, A, X) = \alpha \times L_{content} + \beta \times L_{style}$

Here α and β are hyperparameters that need to be set. The ratio α/β determines emphasis is on content or on style:

1. A large α/β ratio means more emphasis is on content of P in X.
2. A small α/β ratio means more emphasis is on style of A in X.

Other audio related work are [9] [2] [8] [6]

3 Methodology

Our proposed methodology is shown in Figure 3.

3.1 Dataset

Since this work includes Style Transfer on both videos and audios, the input is a small duration video (about 1 min) downloaded from YouTube called storm.mp4. For applying style to frames of video, we used The Starry Night by von Gogh. For audio in the downloaded video we used an instrumental music crescent.mp3

3.2 Processing Technique

The most prominent deep neural networks able to process images are Convolutional Neural Networks (CNNs). Each layer of CNN learns certain features of input image. The output of each layer is called a feature map.

In the Figure 4, at Layer 1 using 32 filters the network may capture simple patterns, for example, lines or edges which is of immense importance to the network. As we move deeper to Layer 2 with 64 filters, the network starts to record more complex features by using the edges detected in previous layer(s) to form corners (intersection points of edges) and then parts of objects. This process of recording different simple and complex features is called feature extraction.

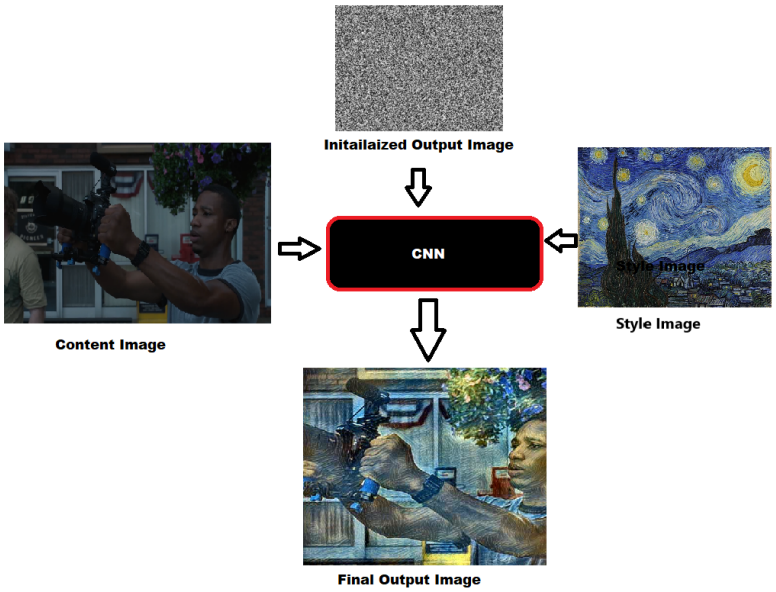

Fig. 3: Proposed Methodology

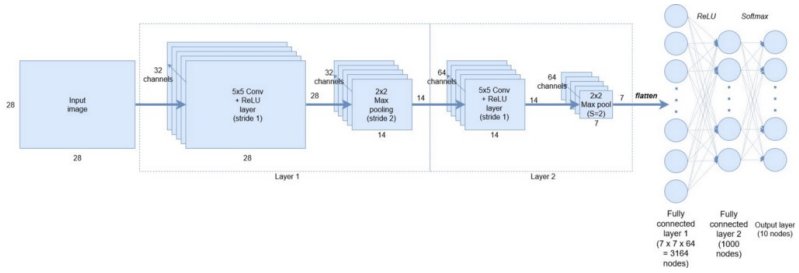

Fig. 4: General Scheme of a CNN

Method for Audio Style Transfer An approach similar to Image Style Transfer (as described in Section 2) is used but modifications are made for audio signals (See our implementation). AlexNet architecture is trained having smaller receptive size of 3×3 instead of original filter size to maintain resolution.

Training of AlexNet is done on spectrograms of instrumental sounds with 3×3 convolution filters and 2×2 max pooling. It has 6 layers in total and Adam optimizer.

A video is collection of frames which are swapped sequentially at a sufficiently fast rate. So we first obtained the frames of video before applying style. Next we applied our code to each of the frame and saved the results. Last thing for Video Style Transfer was to re-attach the frames to form a video.

For applying Audio Style Transfer, we extracted the audio from the original video and then applied to it the style audio(crescent.mp3) and saved the result. The last thing to do was to combine generated video and audio. For that we used a software ffmpeg that was downloaded locally on our computer from https://www.ffmpeg.org/download.html.

We used VGG-19 instead of VGG-16 that is 19 layer deep CNN trained on Stanford Vision Lab's ImageNet Dataset with about 14 million images. VGG-19 can detect higher level features of an image. A schematic representation of VGG-16 and VGG-19 has been shown in Figure 5(a) and Figure 5(b) respectively.

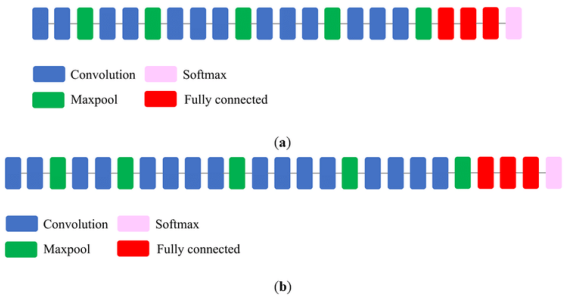

Fig. 5: Schematic representations of VGG-16 and VGG-19

4 Results

For the given input, how our method works is shown using following example. Here input and style images are shown in Figure 6 and Figure 7 respectively. The output of our proposed method is shown in Figure 8.

Fig. 6: Input

Fig. 7: Style Image

Fig. 8: Result

Implementation Details: The major novelty of our work is styling in terms of video along with audio styling. The existing approaches only focus on video [1] styling without audio or just audio [9].

- Click here for the existing code.
- Click here for the code of proposed approach.

The potential applications of out work are in the following domains :

1. Social Communication: Due to its eye-catching output, Style Transfer can be used in networking sites such as Facebook and Twitter.The comments by users can be further used to improve the algorithm.
2. Entertainment World Domains such as gaming and animations may find Style Transfer useful.
3. Mobile Applications Apps such as Prisma have been developed and gained a wide popularity in short term of release.

Comparison With Some Previous Methods: Determining quality of images is almost a task subjective in nature. Most widely accepted method is how user rates/ prefers the aesthetics of the result. Other approach can be how faster the algorithm converges. Our work outperforms in visual plausibility when compared with some of the existing works. For example, one of the implementation beaten by our code is found here:

Fig. 9: Output of an existing method

5 Conclusion and Future Directions

In the present work we implemented an algorithm that can apply style transfer to not only visual part of a video but also to its audio part. Although the output of our work is quite impressive visually but it requires high computation

power(for example, on local computer it took 9 hours 30 minutes for just 85 iterations which is not in real-time for obvious reasons). Hence we turned to Google Colaboratory which significantly sped up the execution with Google's GPUs as Runtime Type.

Also potential future work for this project may include:

- The duration of video and audio must be the same so that both conclude at the same time in final resultant video.
- The trade-off between time efficiency and quality of output can be adjusted.
- Ensuring that in no case should the style transfer leave any geometric effect (e.g. the window pane in some image should remain as window pane and should not get distorted on applying Style Transfer).
- The quality of audio may be improved further.
- In our final output the frames appear to be cropped. This can be taken care of.

Bibliography

[1] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[2] Eric Grinstein, Ngoc QK Duong, Alexey Ozerov, and Patrick Pérez. Audio style transfer. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590. IEEE, 2018.

[3] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[4] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[5] Haochen Li. A literature review of neural style transfer.

[6] Dhruv Ramani, Samarjit Karmakar, Anirban Panda, Asad Ahmed, and Pratham Tangri. Autoencoder based architecture for fast & real time audio style transfer. *arXiv preprint arXiv:1812.07159*, 2018.

[7] Tiancheng Sun, Yulong Wang, Jian Yang, and Xiaolin Hu. Convolution neural networks with two pathways for image style recognition. *IEEE Transactions on Image Processing*, 26(9):4102–4113, 2017.

[8] Maciek Tomczak, Carl Southall, and Jason Hockman. Audio style transfer with rhythmic constraints. In *Proceedings of the International Conference on Digital Audio Effects*, pages 45–50, 2018.

[9] Prateek Verma and Julius O Smith. Neural style transfer for audio spectograms. *arXiv preprint arXiv:1801.01589*, 2018.

[10] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.