Deep Learning

Artistic Style Transfer For Videos

Mario Hany Hunter 48

Merit Victor 69

4th March, 2020

Introduction

Style transfer is a computer vision technique that made a great impact in 2015. It is an image processing and manipulation technique in the field of non-photorealistic rendering. In the past, manually re-drawing an image in a certain artistic style required a professional artist and a long time. Doing this for a video sequence single-handed was beyond imagination. Nowadays computers provide new possibilities. In traditional supervised techniques, image style transfer requires two images; the original image and the artistic representation of the original images. So the model can learn the transformation from the original to the stylized image and apply it to a new one.

Practically, it is difficult to find the artistic representation for every image and hence this approach is inefficient. In recent years, researchers have used the Neural Style Transfer (NST) approach to solve the problem of style transfer. Neural networks are used to extract statistical features of images related to content and style so that we can quantify how well the style transfer is working without the explicit image pairs. With this improved approach, only a single style reference image is needed for the neural network to apply it to original content images.

Motivation

Content Style Transfer has been a hot topic for a while. Many proposed models have been used to learn the style from an image and use it to create content resembling another image with the stylized features learned from the former, effectively accomplishing a style transfer for images using Deep Learning.

As a logical step, Video Style Transfer was the next milestone to tackle. Many approaches that were built upon the Image Transfer were proposed. Each is trading off, efficiency and performance with accuracy. As an example, Instead of using a per-pixel loss function, *Johnson et al.,* have proposed a Perceptual Loss function which makes use of outputs of higher layers outputs instead of each pixel, effectively decreasing run-time while maintaining a decent level of accuracy.

However, Video Style Transfer imposes an additional challenge. Namely, flickering and popping. By processing the video frame by frame and applying style transfer to each, each frame converges to a different local minima which when compressed produces an unstable flow of motion, and an erratic changes in contrast and texture. Techniques have been developed to tackle such instability, but comes at a great cost.

Further, when applying these models to devices with limited computational power, they can overwhelm the device.

Style Transfer in general is an interesting problem on its own. In addition to having peculiar results, it can be useful to different factions from professionals in the video editing industry to amateurs producing videos with mobile devices.

Related Work

Early work in Image Style Transfer was introduced by *Gatys et al.[1],* the model they proposed used a per-pixel error function and applied loss to minimize error in style and content.
Later models proposed by *Johnson et al[2].,* improved the efficiency of such models by incorporating a perceptual loss function instead of a per-pixel loss. By using a pre-trained network on images network such as VGG, the loss is computed by the outputs of the jth layer instead of each pixel.
Building upon these approaches, Naive models then tackled the video transfer problem by using a divide-and-conquer approach. Divide the video into frames, style transfer each image, conquer by recompressing frames again. However, this suffered from instability. Each frame converged to a different local minima, effectively creating flickering in colors and textures.

More sophisticated models that counted for the nature of videos, the motion and the correlation between frames have been proposed. *Manuel Ruder et al.[3],* introduced a new loss term that penalized the optical flow between frames to create temporal consistency. While producing smooth results, the model suffered from complex structure and needed longer time to train and produce results, which prevented real-time models to be produced. See the video here.
A variation of the losses were introduced with the goal to produce a temporal consistency. These variations can be found in *Chen et al.[4],* and *Huang et. al[5].*
A different approach was tackled in ..., which instead of penalizing the optical flow, penalized the fact that changes between frames should be no more than random noise and not due to style transfer output, which essentially fasten the training times.

The problem of real-time models for computationally limited devices are tackled by performing techniques such as model compression and knowledge distillation as described by *Hinton et al[6].*

Datasets

For this problem, no particular Datasets are needed. Only pre-trained networks such as VGG, and any mix of a stylized image and a content video will work. However, picking different styles will be adamant to detect the model prowess in detecting and learning to transfer different styles.

Proposed Approach

Since the optical flow penalty method produced smooth results, it's natural to pursue it. However the model is slow. So, techniques of optimizing the models discussed before will be tried to optimize it for real-time applications.
A different approach is to use the noise addition method while increasing accuracy by incorporating a simpler model of the loss mentioned before, gaining the best of both worlds.
Finally, it may be worth the investigation of replacing the VGG with other networks such as Inception to check the most efficient network while maintaining accuracy.

Evaluation Metrics

In the field of Style transfer, the final evaluation will be mainly qualitative. However, quantitative measures such as the style loss, content loss, optical flow penalty, and random noise metric can be used to train the model and gain insights on how well the model is doing.

Arabic Sentiment Analysis

In recent years, the interest in analysing the opinion of the people and the new trends has increased. Sentiment Analysis (SA) as well as opinion mining is broadly investigated research areas]. It is an application of natural language processing (NLP), computational linguistics, and text mining to extract people' opinions or emotions towards an event, product, or others.

Arabic language has received less effort compared with other languages; however, hundreds of studies have been proposed for Arabic Sentiment Analysis (ASA).

In our graduation project we propose a Deep Learning (DL) approach to classify Arabic Tweets into 3- or 5- sentiments, namely, positive (and very positive), neutral, negative (and very negative).

References

[1] Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. CoRR abs/1508.06576 (2015), http://arxiv.org/abs/1508.06576

[2] Justin Johnson, Alexandre Alahi, and Li Fei-Fei: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. arXiv:1603.08155 (2016), https://arxiv.org/abs/1603.08155

[3] Manuel Ruder, Alexey Dosovitskiy, Thomas Brox: Artistic style transfer for videos. arXiv:1604.08610v2 (2016), https://arxiv.org/abs/1604.08610

[4] Dongdong Chen1, Jing Liao2, Lu Yuan2, Nenghai Yu1, Gang Hua2: Coherent Online Video Style Transfer. arXiv:1703.09211 (2017), https://arxiv.org/abs/1703.09211

[5] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, Wei Liu.CVPR pp. 783-791 (2017), http://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Real-Time_Neural_Style_CVPR_2017_paper.html

[6] Geoffrey Hinton, Oriol Vinyals, Jeff Dean: Distilling the Knowledge in a Neural Network. arXiv:1503.02531 (2015), https://arxiv.org/abs/1503.02531