**Year -** 2019

**Idea -**

We propose a word image generating method called Synth-Text Transfer Network to solve the lack of in training data problem through synthesizing more samples with arbitrary textures and different content text in a style transfer pipeline.

Synth-Text Transfer Network utilizes a style transfer approach to synthesis images with arbitrary text content with preserving the texture of the referenced style image in the target dataset.

The large amount of synthesized images can help to alleviate the overfitting problem and improve the accuracy in latter scene text image recognition tasks.

Our proposed method has two contributions showed as follows:

Firstly, we put forward a novel pipeline to generate massive text images based on a few collected scene text images.

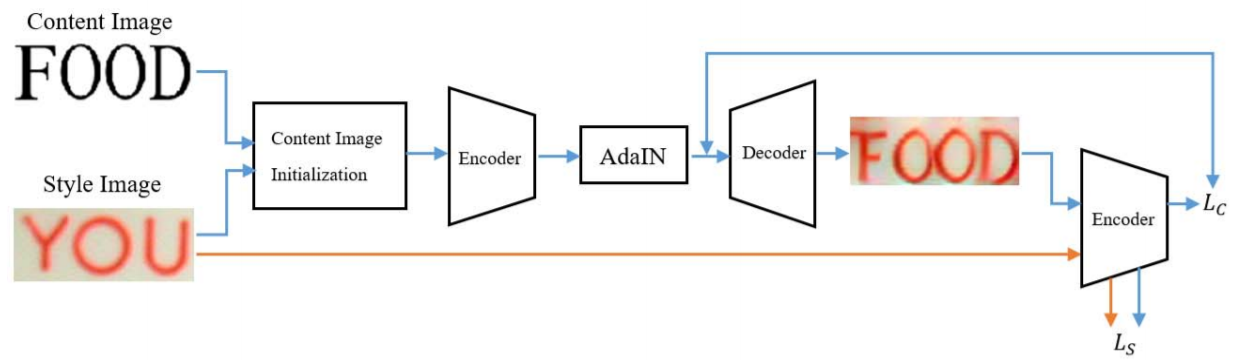
Secondly, the Synth-Text Transfer Network can produce appealing and adequate synthesized images, which may further make the accuracy of current text recognition models better and avoiding the overfitting problem.

**Method -**

We select an image from the training dataset as the style image and a binary image as the content image which is generated by the corresponding dataset. The Content Image Initialization module that we newly proposed initializes the content image according to the mean grey value of the style image. And then the Style Transfer Network generates a synthetic image that preserves the content word and have similar texture of the referenced style image.

The input of our method is one image from the target dataset as style image and a binary image as the content image. Then, the style image and the content image are fed into the Content Image Initialization module to decide the color of the background area and the text area (the background area is white and the text region is black or the background area is black and the text region is white). After that, The STN transfers the content image to one particular style, which is a feedforward transmission network. When it has been trained, it will able to convert the input image into arbitrary style by a single forward pass with a considerable speed.

We use a simple encoder- decoder architecture where the encoder is a pre-trained VGG-19 with first few layers fixed. After encoding the content and style image to feature space, we feed feature maps of style and content image to the AdaIN [8] layer, which aligns the mean and variance of the content feature map with the mean and variance of the style feature map, thus generate the target feature maps. We can see the full structure in Figure 1.

Figure 1. The architecture of the proposed method. Using an Content Image Initialization and an Encoder-AdaIN-Decoder architecture. The first few layers of VGG-19 network is fixed to encode both content and style images. The decoder is learned to invert the AdaIN output to the image spaces. The AdaIN layer is used to conduct style transfer in the feature space. Then we compute the content loss Lc and the style loss Ls by using the same VGG encoder.

**Results -**

The consequence of our experiments showed that AdaIN is the most proper module for arbitrary style transfer. Synth Text-Transfer's interpretable and flexible pipeline allows users to tightly control the fonts, content text, distortion or random noise. Moreover, it is convenient for us to create training datasets effectively in arbitrary language for text recognition task. If there is not enough training image for this language, it will bring significant improvement.

As shown in Figure 2, the AdaIn [8] module concentrate on the similarity of global statistic variable. Also it can render the texture and color of the style image onto the content image and keep the overall text structure unchanged.

Figure 2. Examples generated by our method. Alphabets and numbers in content images are all standard Times New Roman font. On the left is the referenced style images and its corresponding generated images, whose content word are all “food”. And the content word on the right is randomly selected.

However, our method has limitations when the style image has complicated background texture or uncommon font. In Figure 3, the synthesized image has irregular texture and the background is unlike the referenced style image. When style image has low resolution like the last row in Figure 3, then the transferred image cannot keep original font and hard to identification. These images are considered as noise data and will influence the accuracy of text recognition.

Figure 3. More samples generated by Synth-Text Transfer. The referenced style images from benchmark datasets with uncommon font, complex background texture or low resolution are in the left column. The right column is transferred results, which cannot learn the style very well and even reverse the color of background and text region.