Literature Review: Style Transfer and CNN

Xu Zheng

April 17, 2018

Abstract

This is an literature review document to record the useful papers related to topic: Data Augmentation With Artistic Style

1 CNN Models

1.1 AlexNet

AlexNet from Krizhevsky et al. (2012) is the winner of ILSVRC 2012 and the first model to make CNN popular in Computer Vision field. In this work, a 8 layers CNN model are introduced. **Data augmentation techniques such as image translations, horizontal reflections, and patch extractions were used to avoid overfitting.** ReLU and dropout are also used in this paper.

1.2 VGGNet

VGGNet is a simple but deep model created by Simonyan and Zisserman (2014). This model strictly used 3x3 filters with stride and pad of 1, along with 2x2 maxpooling layers with stride 2.

- 3 3*3 conv layers back to back have an effective receptive field of 7x7.
- Used scale jittering as one data augmentation technique during training.
- But it took a very long time to train: Trained on 4 Nvidia Titan Black GPUs for two to three weeks.

1.3 ResNet

ResNet by He et al. (2016) is a 152 layer network architecture that won ILSVRC 2015.

- The idea behind a residual block is that you have your input x, after conv layer, relu layer and normalization layer series, you will get fearure maps: F(x). That result is then added to the original input x: H(x) = F(x) + x, and then continue the training.
- **Naive increase of layers in plain nets result in higher training and test error.**
- The group tried a 1202-layer network, but got a lower test accuracy, presumably due to overfitting.

1.4 ZFNet

- Reconstruction in ZFNet: Zeiler and Fergus (2014) introduced another reconstruction method with Unpooling, Rectification and Filtering are applied for visualization of an activation in some layers, But it can only reconstrut one activation one time (To examine a given convnet activation, all other activations in the layer are set to zero and pass the feature maps as input to the attached deconvnet layer).

2 Descriptive Neural Style Transfer

2.1 Initial Neural Style Transfer Work

The algorithm of Gatys et al. (2016) is the first method that use Gram matrices to to represent the style and use some layer to represent the content, and then reconstruct the stylized image by minimizing the loss by gradient descent with backpropagation. Many Neural Style Transfer ideas have been introduced in the work:

- The basic idea of image style transfer is to jointly minimise the distance of the feature representations of a white noise image from the image content representation in one layer and the painting style representation defined on a number of layers of the Convolutional Neural Network .
- The author found that replacing the maximum pooling operation by average pooling yields slightly more appealing results.
- Adjust the trade-off between content and style to create different images.
- The different initialisations do not seem to have a strong effect on the outcome of the synthesis procedure
- In this work, the author consider style transfer to be successful if the generated image looks like the style image but shows the objects and scenery of the content image.
- VGG network is applied in this work.

2.2 Image Content Reconstruction Mechanism

- In the work by Gatys et al. (2016), the author performs gradient descent on a blank image matches the feature responses of the original image to visualize the information remained in different layers. In the original work menthiond in the work by Mahendran and Vedaldi (2015), a more detailed and mathematical explanation is made: the author models a representation as a function $\theta(x)$ of the image x and then computing an approximated inverse $\theta(x)^{-1}$, reconstructing x from the code $\theta(x)$.
- Along the processing hierarchy of the network, the input image is transformed into representations that are increasingly **sensitive to the actual content** of the image, but become relatively invariant to its precise appearance. Thus, higher layers in the network capture the high-level content and higher layers of the network are used to do the content representation (Gatys et al. (2016)). In my future experiment, I can try different layers to reconstruct the images with different "level" of content.

2.3 Image Texture Synthesis Mechanism

2.3.1 Statistical Model for Texture Images

In this work, to generate a texture from a given source image:

- Extract features of different sizes homogeneously from this image.
- Compute a spatial summary statistic on the feature responses to obtain a stationary description of the source image.
- Find a new image with the same stationary description by performing gradient descent on a random image that has been initialised with white noise (Portilla and Simoncelli (2000)).

2.3.2 Texture Synthesis Using Convolutional Neural Networks

Within the model Gatys et al. (2015), textures are represented by the correlations between feature maps in several layers of the network.

- The author used the feature space provided by a highperforming deep neural network and only one spatial summary statistic: the correlations between feature responses in each layer of the network.

- The style representation is a multi-scale representation that includes multiple layers of the neural network. The number and position of these layers determines the local scale on which the style is matched, leading to different visual experiences.
- The evaluation criterion for the quality of the synthesised texture is usually human inspection and textures are successfully synthesised if a human observer cannot tell the original texture from a synthesised one Gatys et al. (2015).

3 Fast Neural Style Transfer

Although the the work by Gatys et al. (2016) can produce impressive stylized images, there are still soem efciency issues "since each step of the optimization problem requires a forward and backward pass through the pretrained network" Johnson et al. (2016). We are going to apply the Style Transfer as a data agumentation strategy, so the performance is of great importance. Based on the previous work, many Fast Neural Style Transfer methods has been proposed.

3.1 Perceptual Losses and FeedForward Network

Based on the algorithm proposed by Gatys et al. (2016), Johnson et al. (2016) introduced a much faster approach.

- Their system consists of two components: an image transformation network and a loss network
  - The image transformation network is a deep residual convolutional neural network parameterized by weights W; it can transform input images to output images.
  - Each loss function computes a scalar value to measure the dierence between the output image and a target image, including Feature Reconstruction Loss and Style Reconstruction Loss.

3.2 N-Styles FeedForward Network

Although the work by Johnson et al. (2016) are much faster than descriptive methods, their limitations are also obvious: each generative network is trained for a single style, which means that we need to train multiple networks for different styles, so it is time consuming and not flexible. Based on the observation, Dumoulin et al. (2016) proposed an algorithm to train a conditional style transfer network for multiple styles. Their work "stems from the intuition that many styles probably share some degree of computation, and that this sharing is thrown away by training N networks from scratch when building an Nstyles style transfer system."

By using their method and tuning parameters of an conditional instance normalization, we "can stylize a single image into N painting styles with a single feed forward pass of the network with a batch size of N."

So we could build a collection of style embeddings upon which a generative model could be trained.

4 Dataset and Model Choosing

4.1 Image Dataset

There are lots of image dataset. Since we need to do some analysis in the future, the image quality should be good. Here are some popular dataset avaliable online.

- Caltech 256: http://www.vision.caltech.edu/Image_Datasets/Caltech256/ (1.2 GB. image size/quality is fine)
- Caltech 101: http://www.vision.caltech.edu/Image_Datasets/Caltech101/ (131MB, image quality)
- CIFAR10 / CIFAR100: image size is too small (32*32)
- ImageNet(more than 100GB. select specific classes: how to select would be a problem)
- Tiny ImageNet(image Size is too small 64*64)

- Pascal VOC2007 and VOC2012: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html(around 2 GB)
- Coil-100/Coil-20: images are too small.

Conclusion: The first dataset can be Caltech101, which is a small but effective dataset. This dataset has been tested in the original VggNet Paper. I can take it as my first dataset and do some experiment. I can also do some testing in Caltech256 and VOC2007/2012 after I get some results from Caltech101.

4.2 Classification Model

The first model I use will be vgg16. Then I can also try vgg19. The output would look like:

Model	**Dataset**	**result**
Vgg16	Caltech101	...
Vgg19	Caltech101	...
Vgg16	Caltech256	...
Vgg19	Caltech256	...

Bibliography

Dumoulin, V., Shlens, J., and Kudlur, M. (2016). A learned representation for artistic style. *CoRR, abs/1610.07629*, 2(4):5.

Gatys, L., Ecker, A. S., and Bethge, M. (2015). Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2414–2423. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Mahendran, A. and Vedaldi, A. (2015). Understanding deep image representations by inverting them.

Portilla, J. and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.