CAPTIONING DISTORTED IMAGES USING GENERATIVE ADVERSARIAL NETWORK AND DEEP NEURAL NETWORK

Anand Vardhan, Arjun M, Chirag Kashyap S , Harish S, Rama Devi P
PES University, Bengaluru, India

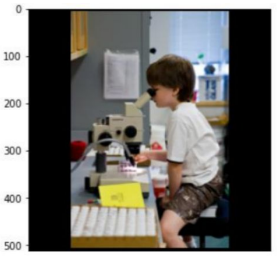
{anandvardhan371, arulakiman2000, chiragkashyap15, sarasharish2000, ramadevip}@gmail.com

Abstract

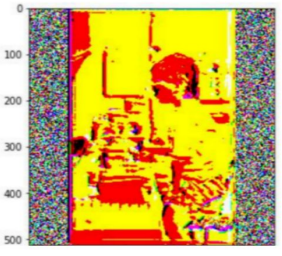
There has been immense progress in the field of Image Captioning over the past decade. Models have evolved from sputtering disjoint tokens to predicting meaningful excerpts. Despite all the progress, one enormous problem remains unaddressed in the domain of Image Captioning: Captioning of hazy, distorted images. Most of the recent advancements in the field of the Image Captioning deals with captioning of still high resolution images which contains a lot of details and clarity. The real world scenario is very different compared to what these research deals with. One cannot expect images to be clear, taken from a still frame of reference. Real world examples are often distorted or hazy. In spite of great research in the field of Image Captioning, there hasn't been an attempt to intertwine the world of Image Captioning and Image Denoising. This paper presents, PixCaption, an unorthodox coupling of a Generative Adversarial Network with a VGG-GRU model for captioning distorted and hazy images. To our knowledge, this is the first framework capable of accurate Image Captioning on distorted images. To achieve this we have coupled a denoising network with a state-of-the-art Image Captioning mechanism. Similar to the Super Resolution Generative Adversarial Network (SRGAN), we use perceptual loss function which consists of Content Loss and Adversarial Loss to enhance image details. This processed image is piped into a VGG model coupled with a GRU to caption image.
This method outperforms any other state-of-the-art method that seeks to caption image trained on the same dataset.

1. Introduction

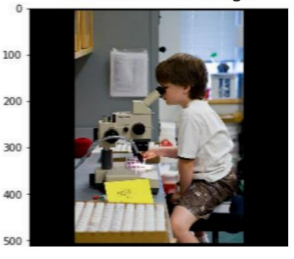
There exist two highly challenging tasks in Front of us - the task of denoising a blurry and hazy image and the task of captioning the image.
A highly important real world problem, image captioning , which till very recently most people couldn't imagine solving has suddenly become a simple task with the advent of deep neural networks.



Noisy Original Image



Denoised BRY Image



Denoised RGB Image

Generative Adversarial networks proposed in Goodfellow et al.[1]
is an adversarial network architecture like min-max game, between
generator and discriminator with adversarial loss function based on
JSD(Jensen Shannon Divergence).

For denoising we took a similar approach taken in Image SuperResolution using a Super Resolution Generative Adversarial Network (SRGAN) proposed by Ledig et al.[2] with a duplex loss function - one to measure Content Loss and the other to measure Adversarial Loss.

By using the methodology adopted by Ledig et al., denoising can result in sharp images. This helps the captioning model improve at detecting features like grass, trees, and the others with a fine separation.

The detailed images are then fed into Simonyan et al. VGG network[3]. VGG breaks down the given image into tokens of items perceived within the image. Once the tokens are received, we need to order them correctly to build up the symphony.

Gated Recurrent Unit was introduced in 2014 by Kyunghyun Cho et al.[4]
with each unit containing an update and forget gate, solving the problem
of vanishing gradient and also increased efficiency over Vanilla RNN.

A Gated Recurrent Neural Network reads these tokens and constructs the caption for the given image. This is how the PixCaption is able to generate captions for a distorted image.

2. Architecture

While attempting to solve a large and tedious problem one usually splits it into sub-problems or sub tasks. In the same way to overcome the problem of captioning distorted images we have chosen to split our model into two main parts, one to handle the image denoising and another to handle the task of captioning.

2.1 Adversarial Network

2.1.1 Design of the Adversarial network

In today's world we have seen that deep neural networks outperform most other networks though it is quite tough to train them. The generator network consists of nine convolutional layers and the discriminator has a set of four convolutional layers

In our model we not only aim to increase our accuracy but also improve our training speeds. This is why we have implemented Batch Normalization[5],which is a well known method to reduce the internal covariate shift. Also Residual blocks[6] and Skip connections[7] used in the generator network help improve the gradient flow throughout the deep neural network .In turn we see that these deep networks provide better accuracy.

2.1.2 Method and Loss Function

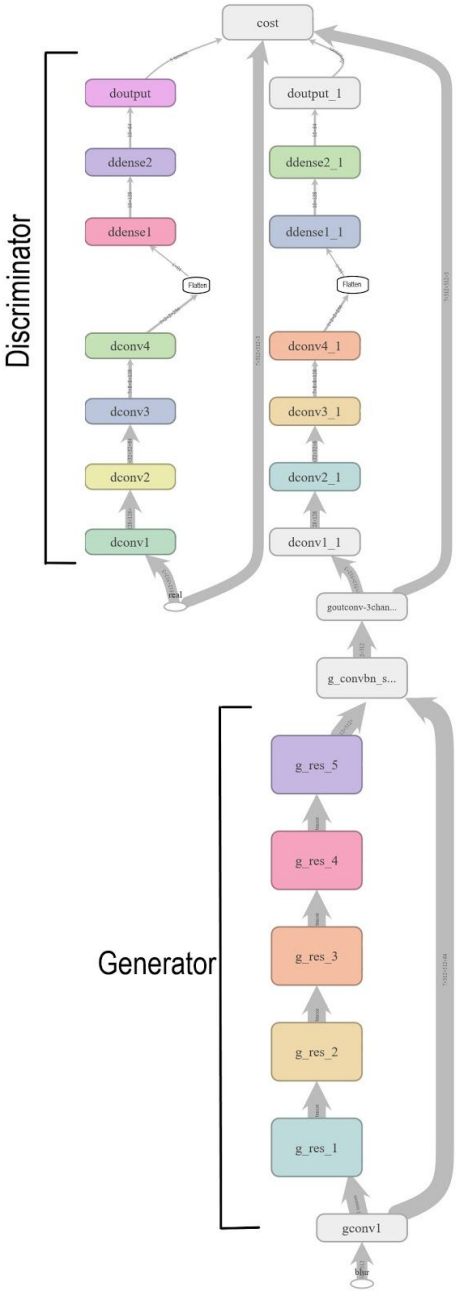
Not having constant image sizes can pose a problem while training. This problem can be solved with zero padding the images to make them have constant sizes .However Zero Padding images is basically adding noise to the images .Since images were already of varying sizes , this caused the padding affixed to be of varying sizes making it very tedious for our model to learn and adapt. Another method to resolve this problem is spatial pyramidal pooling[8]. Finally we observed that resizing the images to a fixed size helped the model learn better and tackle this problem most efficiently .The loss function used in our network involves two functions which are the content loss and the adversarial loss. We have borrowed this idea from Image Super Resolution using a Super Resolution Generative Adversarial Network (SRGAN) proposed by Ledig et al[2].The content loss is a pixel wise mean as follows :

$$L_{MSE} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (I_{x,y} - G_{\theta G}(I_{LR})_{x,y})^2$$

The adversarial loss contains a generative component which is added to the perceptual loss function. This loss function is based on [7] and is as follows:

$$L_{GAN} = \sum_{n=1}^{N} -logDD_{\theta_D}(G_{\theta_G}(I_{LR}))$$

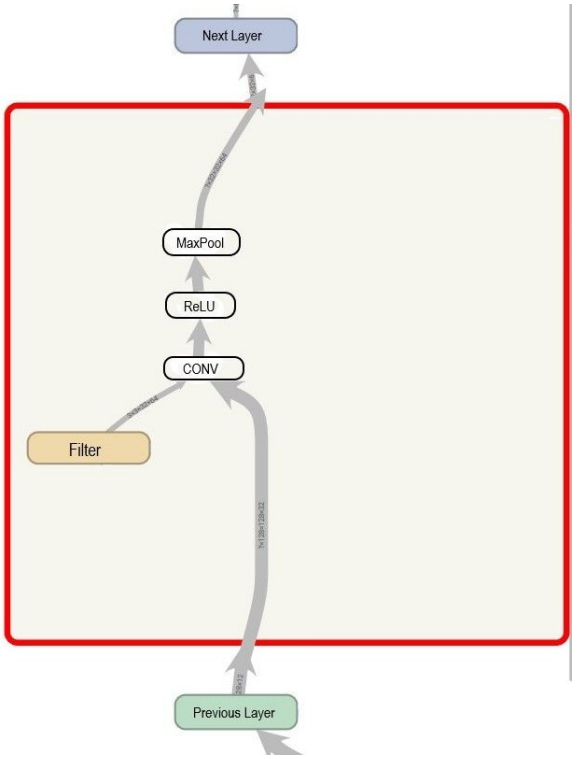
2.1.3 Adversarial Network Structure



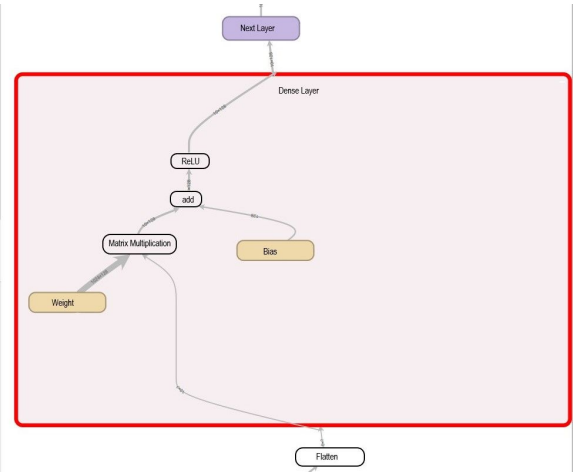
All generator elements are prefixed with g and all discriminator elements are prefixed with d.

The residual block contains
Conv, BN, relu, Conv, BN, element wise sum k3n64s1, skip connection between each other.

The deconv start with k3n32s2 and go upto k3n256s2. They all have maxpool with k2s2.



Expanded view of a convolution layer with ReLU activation and maxpool



Expanded view of a fully connected layer with ReLU activation

The output from last dconv gets flattened and passed through some fully connected layers(ddense).

2.2 Image Captioning Model

Due to the recent improvements made in the image captioning domain there are a large number of architectures based upon which one can construct a model. We have gone for the Recurrent Neural Network(RNN) + Convolutional Neural Network(CNN) model.

2.2.1 Deep Convolutional Network

The network we have used is composed of a standard VGG16 model architecture.We have implemented this architecture as this serves as an excellent tool while working with datasets containing a large number of images. The VGG16 model was pre-trained on the ImageNet dataset for classifying images. But instead of using the last classification layer, we have redirected the output of the previous layer. This gives us a vector with 4096 elements that summarizes the image-contents. This vector has to mapped down to a vector containing 512 elements , to solve this problem we use a fully connected layer or a dense layer. We have implemented embedding layers[14] which have proven to outperform the traditional bag - of - words model encoding schemes.In word embedding layers we see that words are represented by dense vectors and these vectors represent the projection of the word into a continuous vector space.

2.2.2 Recurrent Neural Network

This network is trained to map the vectors with transfer-values from the image-recognition model into sequences of integer-tokens that can be converted into text. This recurrent network consists of one embedding layer along with three Gated Recurrent Units(GRU)[4]. Since neural networks cannot work directly on text-data. We use a two-step process to convert text into numbers that can be used in a neural network. The first step is to convert text-words into so-called integer-tokens.The second step is to convert integer-tokens into vectors of floating-point numbers using an embedding-layer.

3. Related Work

3.1 Image super resolution

Recent works in the field include Moeslund[10], J Yamanaka[11], Ledig[2]. These papers show the accuracy boost of DNN over traditional dehazing techniques like bayesian probabilistic model Nishino et. al.[12] and prediction based techniques for upsampling.

3.2 Generative Adversarial Network

Since the introduction of GAN's in 2014 by Goodfellow et al. [1] they have come a long way. Present day GAN's are capable of producing desired results at a speed unparalleled to its previous counterparts while maintaining high accuracy. Our model utilises aspects from the architecture proposed by Christian Ledig et al. [2] and improves upon it to remove noise from the image.

3.3 Image Captioning

There are numerous models for image captioning as mentioned in survey Hossain et al.[13]. We focus on using efficient model that train fast with less hardware resources with VGG16 and GRU.

4. Future work

There are still many tests, adaptations and hyperparameter tuning which have been left for future due to lack of time and resources.
There are some ideas that we wanted to test during the development of the model like spp Kaiming et al.[8] , improving performance by using skip models Dieng.[9] which helps avoid possible mode collapses.

5. Results

During training at first we observed that the GAN was primitive and couldn't identify definite edges or corners in the images. Upon continued training the network was able to detect the edges and outlines of objects in the images to a certain extent. The model will be able to generate promising results if it is allowed to train for a longer period of time.

6. References:

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative Adversarial Networks. 2014.

[2]Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017.

[3]Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014.

[4]KyungHyun Cho, Junyoung Chung, Caglar Gulcehre, Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 2014.

[5]Sergey Ioffey, Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015.

[6]Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. 2015.

[7]A. Emin Orhan, Xaq Pitkow. Skip Connections Eliminate Singularities. 2018.

[8]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. 2015

[9]Adji B. Dieng, Yoon Kim, Alexander M. Rush, David M. Blei. Avoiding Latent Variable Collapse With Generative Skip Models. 2019.

[10]Nasrollahi, K., & Moeslund, T. B. Super-resolution: A comprehensive survey.2014.

[11]Jin Yamanaka , Shigesumi Kuwashima and Takio Kurita. Fast and Accurate Image Super Resolution by Deep CNN with Skip Connection and Network in Network. 2017.

[12]K. Nishino, L. Kratz, and S. Lombardi. Bayesian defogging.2012

[13]Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. 2018.

[14]Zhongliang Li , Raymond Kulhanek ,Shaojun Wang ,Yunxin Zhao ,Shuang Wu. Slim Embedding Layers for Recurrent Neural Language Models. 2017.