**Blending Mode Data Augmentation in CNN Classification**

**Abstract**

Model Architecture and Dataset size both directly impact the ability of a Convolutional Neural Networks to learn generalized patterns contained in training data. Overfitting occurs when a network learns too specifically to the training data and not general enough for the true scope of the testing data. A limited dataset in both size and breath produces overfitting in the model. Data Augmentation is used to create new training data from a limited dataset and relieve overfitting. Classical techniques such as flipping, rotating, translating, and transforming color channels are ways the data can be augmented while not effecting the data’s corresponding label. In this paper I propose a technique utilizing blending modes to highlight features within the dynamic range of training image information. The Multiply and Screen blending modes are used to focus the data on the shadows (dark areas with Multiply) and highlights (light areas with Screen) respectively. Blending modes are used along in conjunction with classic data augmentation techniques to increase classification accuracy and reduce overfitting.

**Acknowledgments**

Thank you to my advisor, Professor Min Chen, for her guidance and friendship.

For Damien & Colette.

**Introduction**

1. CNN

Convolutional Neural Networks (CNN) have revolutionized computer vision over the past decade. Previously computer vision was dominated by expert systems which are constructed by human explanations of the features contained in each object in an image. The CNN removed the need for human explanation and automated the task of building the feature maps, which objects are classified from.

The basic convolution neural network model is centered around a two-part neural network model. The first part being the convolution step where a kernel is moved over the image data creating a set of feature maps. These maps describe the different aspects of the learned image such as edges, textures, and colors. These features maps are fed to a traditional neural network containing full-connected dense layers. The full-connected dense layers then classify the set of feature maps into the user provided classes.

CNN architectures have progressively been optimized for both parameter size and accuracy. The main trajectory has been in constructed architectures that allow for deeper networks. The deeper the network the more features can be described within.

AlexNet(2012) is the first modern (2010 and after) innovation in CNN design to highly perform on the ImageNet dataset with a 16.4 % error rate.[1] AlexNet’s design features convolution layers, pooling, and normalization layers on top of a full-connected classifier with a layer depth of 8. AlexNet also introduced flipping, cropping, and PCA color augmentation to the data.[1]

In the following years model architecture followed the format of AlexNet while increasing the depth of the network.

VGG-16(2014), a deeper network with 16 layers. These layers each have smaller filters, allowing for finer features to be detected. A 3x3 filter, the smallest convolutional filter that can be implemented, is used all the down the network.

ResNet 50 layer network. [3]

As networks got deeper, they encounter the issue of vanishing gradients. The deeper the network goes features produce less activations on their corresponding weights. As the weights are updated through backpropagation, they start to zero out and no further learning is accomplished. ResNet employed residual connections to counteract this issue. The residual connections skip over layers of the network and feed higher level features back into the feature maps. The residual connection is added via matrix addition to the layer immediately below, bypassing the layers convolution step.

DenseNet 121 layer network

Dense Block. Convoltution, pooling, batch normailization. All layer blocks are concatonated to all layers below them.

AlexNet [1] to VGG-16 [2], ResNet [3], Inception-V3 [4], and DenseNet [5].

[1] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1106–14.

[2] Karen S, Andrew Z. Very deep convolutional networks for large-scale image recognition. arXiv e-prints. 2014.

[3] Kaiming H, Xiangyu Z, Shaoqing R, Jian S. Deep residual learning for image recognition. In: CVPR, 2016.

[4] Christian S, Vincent V, Sergey I, Jon S, Zbigniew W. Rethinking the inception architecture for computer vision. arXiv

e- prints, 2015.

[5] Gao H, Zhuang L, Laurens M, Kilian QW. Densely connected convolutional networks. arXiv preprint. 2016.

1. Data Augmentation

Data Augmentation is the technique by which training data is manipulated to produce new examples for the network. Data Augmentation is a, “data-space solution to a limited data set”[1]. Networks must overcome issues inherent in the training data such as viewpoint, lighting, occlusion, background data, and scale. The augmentations try to alleviate the issues from the data which will result in better performing models.

Geometric transformations build the base of Data Augmentation techniques. These transformation are spatial and involve moving the pixel image data within the frame.

Flipping an image on the horizontal axis is one of the most common augmentation methods. The flipping method was first used in AlexNet CNN created by Krizhevsky et  al.[2], on the ImageNet Dataset. In most real-world images, the horizontal flip retains the image’s corresponding label and is easily implemented. In contrast a vertical flip is used more sparingly. Real-world images flipped upside down rarely retain the same

Along with flipping, cropping, rotating, and translating image data also provide geometric augmentation. Cropping an image reduces the size of the input (for example 200 x 200 -> 100 x 100) but allows the network to train on specific areas of the image. The idea is to crop in on the signal in the image while removing the noisy information that surrounds it. Rotating and translating retain the image size and perform a similar feature removal and attention process as cropping.

When decided on which Data Augmentation techniques to employ you must decide if the augmentation is “safe” for the training data. An augmentation is considered safe if it transforms the data without changing the corresponding label. An example of a safe augmentation would be horizontally flipping and image of a dog. An unsafe augmentation would entail flipping the digits in the Minsit Dataset, for example flipping a “6” would change it’s corresponding label. The safety of an augmentation is domain dependent and must be chose specially for the dataset it is used on.

You must perform a search on your dataset to choose the optimal set of augmentations to be layered together.

Offline or Online Data augmentation. Offline is when you augment the dataset beforeband and store it in memory. Online is when you augment on the fly as your model runs. Online disadvantage is the increase in training time and Offline is the increase in memory needed to store the dataset.

*An interesting question for practical Data Augmentation is how to determine post-augmented dataset size. There is no consensus as to which ratio of original to final dataset size will result in the best performing model. However, imagine using color aug-mentations exclusively. If the initial training dataset consists of 50 dogs and 50 cats, and each image is augmented with 100 color filters to produce 5000 dogs and 5000 cats, this dataset will be heavily biased towards the spatial characteristics of the original 50 dogs and 50 cats. This over-extensive color-augmented data will cause a deep model to overfit even worse than the original. From this anecdote, we can conceptualize the existence of an optimal size for post-augmented data.*

[1] Survey of Data Aug

[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1106–14.

1. **Description of Problem**

Overfitting is a major issue in Deep Learning and one of the main causes is the lack of sufficient training data. As network architecture advanced overfitting continued to be an issue. Data Augmentation is one of the main tools used to fight overfitting. In this thesis, I implement a Data Augmentation method which creates new training examples by implementing Blending Modes on training images. The Blending Modes allow the network to search the gamma space of the images. Revealing hidden shadow data while obscuring highlight data or the inverse, revealing hidden highlight data while obscuring shadow data. Blending modes are easily implemented and their layered structure allows for numerous combinations of stacks.

**Thesis Outline**

This Thesis is organized as follows. Chapter 2 describes the background and working details of the chosen Network architecture, Dataset, and Data Augmentation Method. Each was chosen to highlight a clear path from a base structure to a final implementation which in turn highlights the increased accuracy of the network. Chapter 3 describes my proposed architecture of the Blending Mode augmentation layer and the full Data Augmentation Pipeline. The Pipeline and Augmenation layer where constructed using the Keras CV library.

Chapter 4 will summarize my testing results.

Chapter 5 will summarize where the problem solution is currently and propose further sets of testing.

**Background**

1. Model: Densenet

Concatenation in DenseNet. You add the feature maps back in for more layers.

ResNet. Matrix addition. Add the feature map to the next layers feature map.

****

1. Dataset: Tiny Imagenet

ImageNet has become a standard benchmark for computer vision classification. The most used subset of has been ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012-2017 image classification and localization dataset. This dataset contains 1000 classes with 1,281,167 training images, 50,000 validation images and 100,000 test images. Each image in the training, validation, and testing datasets is sized to 256x256 pixels.

Tiny ImageNet was created at Stanford for their Deep Learning for Computer Vision course. (<http://cs231n.stanford.edu/>). It is a smaller version ILSVRC-2012-2017 classification dataset with only 200 classes with 100,000 training images, 10,000 validation images, and 10,000 testing images. The images are evenly distributed between the classes so no class is over represented. The images are down sampled to 64x64 pixels. The lower resolution makes classification a harder task even to the human eye. The benefit of the reduced image number and size is in training time and preprocessing.

The Tiny ImageNet challenge was setup in 2015 at Sandford and has been used to test model architectures and data augmentation pipelines before training on the larger ImageNet dataset.

[1] Ya and Xuan S. Yang. “Tiny ImageNet Visual Recognition Challenge.” (2015).

[2] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

1. Augmentation: Blending Modes

**Proposed Approach**

1. Blending Mode Data Augmentation layer

Adding layers add to the gamma search space. Lighter and Darker.

1. Full Data Augmentation Pipeline

**Experiments & Results**

1. CNN Accuracy for different Data Augmentation layers
2. Training time.

**Conclusion & Future Work**

Since the advent of photography practitioners have been searching for process to maximize the detail in their images. The photographic image is inherently a limited representation of our visual reality. Each image sacrifices certain elements to produce a generalized view of the scene.

Photography in its earliest form was a practice in capturing brightness values. It's invention in the 1800's as a Black and White medium was our first semi-permanent (all physical prints fade over time) process to capture our visual existence. Photography democratized the image creation process away from the artist and allows all of us the ability to curate our visual world.

The digitalization of the image created the opportunity to . . The discrete nature limits the dynamic range and compresses the visual relationships.

Overfitting is a major issue with a limited dataset. The best CNN models come from big data. The more images available the better the ability of the model to form a more generalized view of the relationships in the data.

Image issues: Limited size, lighting, exposure, viewpoint, occlusion, background, scale, \dots

My Thesis will focus on lighting and exposure issues.

Maximize the information in the dataset by creating a more generalized representation by training on the full dynamic range of the image.