The Effectiveness of Data Augmentation in Image and Video Classification using Deep Learning

Jason Wang
Stanford University
450 Serra Mall
zwang01@stanford.edu

Luis Perez
Google
1600 Amphitheatre Parkway
nautilik@google.com

Abstract

In this paper, we explore and compare multiple solutions to the problem of data augmentation in image classification. Previous work has demonstrated the effectiveness of data augmentation through simple techniques, such as cropping, rotating, and flipping input images. In this paper, we artificially constrain our access to data to a small subset of the ImageNet dataset, and compare each data augmentation technique in turn. We also compare the performance of pretrained models on a classification task, both with and without our augmented data for each technique. Finally, we propose a novel technique which makes use of AGNs to transfer day/night styles on our input corpus, thereby augmenting our data size. We evaluate all of the above methods on our subset data and discuss their effectiveness.

1. Introduction

We propose exploring the problem of data augmentation for image and video classification, and evaluating different techniques. It is common knowledge that the more data an ML algorithm has access to, the more effective it can be. Even when the data is of lower quality, algorithms can actually perform better, as long as useful data can be extracted by the model from the original data set. For example, text-to-speech and text-based models have improved significantly due to the release of a trillion-word corpus by Google [4]. This result is despite the fact that the data is collected from unfiltered Web pages and contains many errors. With such large and unstructured data sets, however, the task becomes one of finding structure within a sea of unstructured data. However, alternative approaches exist. Rather than starting with an extremely large corpus of unstructured and unlabeled data, can we instead take a small, curated corpus of structured data and augment in a way that increases the performance of models trained on it? This approach has proven effective in multiple problems. Data augmentation guided by expert knowledge [5], more generic image augmentation [7], and has shown effective in image classification [6].

The motivation for this problem is both broad and specific. Specialized image and video classification tasks often have insufficient data. This is particularly true in the medical industry, where access to data is heavily protected due to privacy concerns. Important tasks such as classifying cancer types [5] are hindered by this lack of data. Techniques have been developed which combine expert domain knowledge with pre-trained models. Similarly, small players in the AI industry often lack access to significant amounts of data. At the end of the day, we've realized a large limiting factor for most projects is access to reliable data, and as such, we explore the effectiveness of distinct data augmentation techniques in image classification tasks.

The initial input to our algorithm is a randomly selected subset of the tiny-imagenet-200 data set. The original data set consists of 100k training, 10k validation, and 10k test images of dimensions 64x64x3. There are a total of 500 images per class with 200 distinct classes. To artificial restrict our data, we focus on a sample of 16k training, 2k validation, and 2k test images, randomly sampled. We will maintain 200 distinct classes. Let N_c be the number of items per class, and by uniform sampling without replacement, we expect to have:

$$\mathbb{E}[N_c] = 80$$

The above is a relatively small amount of data per class, matching what can sometimes be expected from new fields.

With the above data, we will train off-the-shelf models such as VGGNet and AlexNet, which we will tune for optimal performance. We will then proceed to use typical data augmentation techniques, and retrain our models. Finally, we will make use of CycleGAN [8] to augment our data by transferring styles from one image to another, such as doubling our input data by transforming it from day to night. For all the above, we will measure and record training performance and compare the different data augmentation techniques, with specific emphasis on our new ”style”-transfer techniques.

2. Related Work

This section provides a brief review of past work that has augmented data to improve image classifier performance. Additionally, we provide references for recent advances in AGNs and their effectiveness in transferring ”style” from image to image.

The field of data augmentation is not new, and in fact, various data augmentation techniques have been applied to specific problems. The main techniques fall under the category of *data warping*, which is an approach which seeks to directly augment the input data to the model in *data space*. The idea can be traced back to augmentation performed on the MNIST set in [2].

A very generic and accepted current practice for augmenting image data (performed by) is to perform geometric and color augmentations, such as reflecting the image, cropping and translating the image, and changing the color palette of the image. All of the transformation are affine transformation of the original image that take the form:

$$y = Wx + b$$

The idea has been carried further in [3], where an error rate of 0.35% was achieved by generating new training samples using data augmentation techniques at each layer of a deep network. Specifically, digit data was augmented with elastic deformations, in addition to the typical affine transformation.

3. Methods

The first method will consist of measuring the improvement demonstrated by normal data augmentation techniques, and this will be our baseline.

The main object of the paper will be to analyze the effectiveness of our new method for data augmentation. The first task will be to learn style transfer from a large data set or from a pretrained model. We will follow [8] closely for this purpose. For example, we hope to learn transfers:

1. Day to Night
2. Summer to Winter

The above makes use of generative adversarial networks. We will use the ImageNet dataset for this training as well as other datasets used by CycleGAN. We will then compare the results of training an off-the-shelf CNN for image classification using our original and augmented data sets. We will attempt to replicate current world-class image classification models, and differentiate them only on the input data. Finally, we will prototype a successful method by testing it with real-world data sets for video classification.

4. Datasets and Features

As discussed above, we will use: ImageNet data for style transfer learning. Multiple style transfer data sets linked by CycleGAN Video data collected from with Jake Lussier for video classification tasks.

For the milestone task, we wanted to explore if a few simple data augmentation tricks would improve our model's ability to generalize to data. The task we want to solve is image classification just like the ImageNet challenge. However to force the data into a scenario where lack of data poses a challenge, we only feed our CNN a very small portion of the dataset.

We extracted 3 classes from TinyImageNet data, namely goldfish, dog, and cat. From each class, we took 64 random images to add to the training set and 64 random images to add to the validation set.

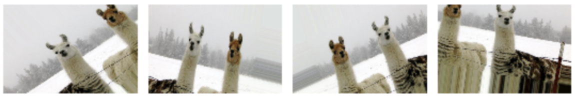

Figure I: Example of Transformations

5. Experiments/Results/Discussion

The task is to explore how small datasets affects training of CNNs. The models we tackle for the milestone are AlexNet and VGG [1]. We vary the number of layers we allow the weights to be trained. By allowing more layers to be trained instead of initialize by pretrained weights, we can observe a larger detrimental impact of having a small dataset.

Alexnet is trained on a dataset of 64 fish, 64 cats, and 64 dogs. Each image is 64 by 64 so we scaled them to 227 by 227. The learning rate is 0.001. Batchsize is set to 32. We tested dropout at 0.15. We train for 20 epochs. Finally, we initialize the weights for various parts of net to the pretrained weights from the ILSVRC submission. One modification was to initialize all the convolution layers while allowing the fully connected layer to be trained. The other was to allow the last convolution layer to be trained as well. Data is then validated on a dataset of 64 fish, 64 cats, and 64 dogs.

VGG is trained/validated on the same dataset. Each image is scaled to 224 by 224. The learning rate is also 0.001. Batchsize is 32 and dropout is 0.15. Training is ran for 20 epochs. We test two modification. The first imports weights for all convolution layers and allows the fully connected layers to be trained. The second iteration allows the weights of the last convolution block to be trained as well.

Finally, in data augmentation, we try various methods to generate 3 images for each image in the training set. This creates a training set of size 64*3*4 = 768. The transformation are a random rotation of up to 40 degrees, width and height shifts of 20 percent, a zoom of up to 20 percent, shear, and some distortion. All images are then flipped randomly. Because the augmented data is 4 times larger, we train for 5 epochs so that all variations go through the same number of batches.

Results indicate that simple data augmentation produces better results than none when we allow the last conv layer (conv5) weights to be trained. In the below plot, we show the validation accuracy against the number of batches trained. Validation accuracy is as high as 70% when data is augmented but only around 50% when not.

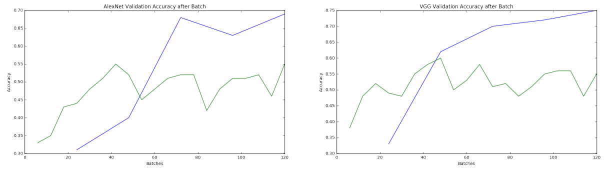

Figure II: Validation Accuracy (Blue is augmented)

The same results are not evident when we only train the fully connected layers. AlexNet achieves 92% accuracy after after 5 epochs without data augmentation and also 92% with augmentation. We hypothesize that the convolution layers learned all the features so the FC layers just aggregate predictions into the desired number of classes.

Validation Accuracy		
Model-layers trained	Original Data	Augmentation
AlexNet-FC	**0.928**	0.924
Alex-C5+FC	0.552	**0.689**
VGG-C5+FC	0.601	**0.748**

Figure III: Best Val Accuracy for all Models

Finally, we see evidence that suggests data augmentation prevents overfitting. When we train weights to the last convolution layer (conv5) without data augmentation, the training accuracy reaches 1.0 really quickly. However with data augmentation, the model overfits at a much slower pace. We run the AlexNet on augmented data again and the results show that training accuracy only slightly beats validation accuracy within 9 epochs. Furthermore, the small discrepancy suggests that there is still more learning to do.

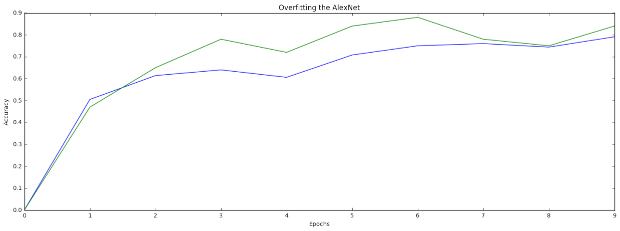

Figure IV: Overfitting the AlexNet with Augmented Data (Blue is val acc)

A good place to explore would be to see what kind of features from the augmented data that the model selected to learn. We hypothesize that simple transformations account for patterns seen in the validation set due to variance in the images. Hence, we anticipate that the CNNs pick up these traits too.

6. Conclusion/Future Work

We've explored multiple common techniques of data augmentation and will compare them to our own, novel, ”style” transfer technique enabled by CycleGAN and other similar networks. Given the plentifulness of data, we would expect that such data augmentation techniques might be used to benefit not only classification tasks lacking sufficient data, but also help improve the current state of the art algorithms for classification. Furthermore, the work can be applicable in more generic ways, as ”style” transfer can be used to augment data in situations were the available data set is unbalanced. For example, it would be interesting to see if reinforcement learning techniques could benefit from similar data augmentation approaches. We would also like to explore the applicability of this technique to videos. Specifically, it is a well known challange to collect video data in different conditions (night, rain, fog) which can be used to train self-driving vehicles. However, these are the exact situations under which safety is the most critical. Can our style transfer method be applied to daytime videos so we can generate night time driving conditions? Can this improve safety?

7. Appendices

Optional. More details to be added later.

References

[1] Finetune alexnet with tensorflow 1.0. https://github.com/kratzert/finetune_alexnet_with_tensorflow. Accessed: 2017-05-19. 2

[2] H. S. Baird. Document image analysis. chapter Document Image Defect Models, pages 315–325. IEEE Computer Society Press, Los Alamitos, CA, USA, 1995. 2

[3] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber. Deep big simple neural nets excel on handwritten digit recognition. *CoRR*, abs/1003.0358, 2010. 2

[4] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, Mar. 2009. 1

[5] C. N. Vasconcelos and B. N. Vasconcelos. Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. *CoRR*, abs/1702.07025, 2017. 1

[6] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: when to warp? *CoRR*, abs/1609.08764, 2016. 1

[7] Y. Xu, R. Jia, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin. Improved relation classification by deep recurrent neural networks with data augmentation. *CoRR*, abs/1601.03651, 2016. 1

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. 1, 2