Deep Learning for Image Analysis

Kyle Kastner
University of Texas-San Antonio

***Abstract*—Deep learning is the broad term for the recent development and extensions of neural networks in the machine learning community, which has allowed for state of the art results in speech, image, and natural language processing tasks. This paper will explore a deep neural network architecture for classifying objects in images for the CIFAR10 and Asirra dataset**

***Index Terms*—Convolutional neural network, ZCA, dropout, maxout, CIFAR10, Asirra, stochastic gradient descent, image processing, machine learning**

I. Introduction

NEURAL networks have a rich history in the A.I. and machine learning communities, starting with the perceptron algorithm in 1957 [1]. The perceptron algorithm is considered by many as the direct precursor to modern neural networks, though effective computation of the neural network was not achieved until 1975 with the invention of the backpropagation algorithm [2]. Most recently, neural networks have returned in several A.I. related fields: "machine learning", "big data", and "data analytics" among them. The massive volumes of data available for modern research, coupled with algorithmic improvements and massively increased computing power, have allowed the once purely theoretical "neural network" to achieve state of the art results in real world image, speech, and text processing tasks. Outpacing more traditional feature oriented techniques in each respective field, neural networks are a topic of interest in both academia and industry.

II. Dataset

The datasets used for these experiments were the CIFAR10 dataset [3] and the Asirra dataset [4], both of which are comprised of 3 channel RGB images, and a single label which describes the contents of the image. CIFAR10 images are 32x32 pixels by default, and the larger Asirra images were rescaled to 32x32 for this experiment. The CIFAR10 dataset features 60000 images representing ten different classes of objects, as listed below.

- automobile
- airplane
- bird
- cat
- deer
- dog
- frog
- horse
- ship
- truck

The Asirra dataset used for these experiments features 25000 images, containing objects representing the classes listed below.

- cat
- dog

III. Preprocessing

The key preprocessing step taken for both of these datasets was to apply zero phase component analysis (ZCA) to each training set. Mathematically, this technique can be described by Eq. 1 [3]. By centering new data (subtracting the column-wise mean of X), and applying W as shown in Eq. 2, the whitened result X_w is obtained. This ZCA whitened result has been shown to resemble the effects of the human vision system [5], and improves classification results in the CIFAR10 task by a wide margin [3]. It is crucial that the W matrix obtained during the training phase is applied to the test set, rather than calculating a new W using the test data.

$$W = (XX^T)^{\frac{1}{2}} = ED^{\frac{1}{2}}E^T \tag{1}$$

$$X_w = XW \tag{2}$$

IV. Network Architecture

Neural networks for image processing often utilize a special layer called a convolutional layer for the first few stages [6], in order to capture the spatial relationships between pixels and colors, as shown in Fig. 1. This allows for a more effective representation of the image during the learning phase, and also allows for some interesting visualizations of the layer one filters. This will be explored more extensively in the next section.

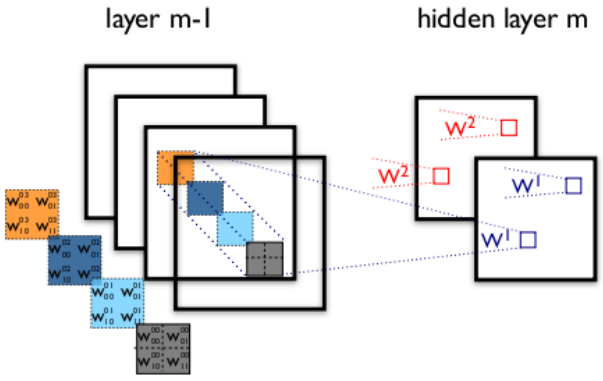

Fig. (1): Convolutional layer visualization [7]

Another technique used in this architecture is a maxout layer, which is a new "universal approximator" layer built

specifically to take advantage of dropout [11]. Using maxout units in place of rectified linear units typically provides fair improvement in existing dropout networks [8].

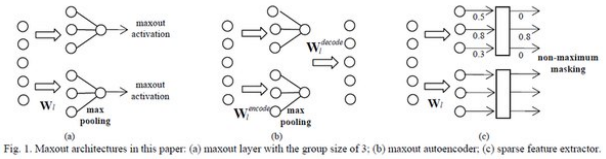

Fig. 1. Maxout architectures in this paper: (a) maxout layer with the group size of 3; (b) maxout autoencoder; (c) sparse feature extractor.

Fig. (2): Maxout layer visualization [10]

The overall network architecture and hyperparameter choices for the CIFAR10 tests were chosen from a draft version of [8]. The authors have subsequently released a higher performing configuration, which requires more computational resources. A nearly identical network and hyperparameters were chosen for the Asirra tests, with the main modification being that the Asirra results used a fused training set of both the CIFAR10 data and the Asirra data, in the hopes that the Asirra data would give discriminative power between cats and dogs, while the CIFAR10 data would provide a baseline image recognition capability. This idea is very similar to the concept of the recently developed DeCAF image processing technique [9]. A summary of the hyperparameter choices are shown below.

- Layer sizes: 48 - 128 - 128 - 240 - 10
- Layer types: Convolutional (C) - C - C - Maxout - Softmax
- Initial learning rate .1, decay .01 per epoch for 250 epochs
- Initial momentum .5, ramping to .6 over 250 epochs
- Stochastic gradient descent, batch size 128
- Dropout .8, scale 1 for layer 1, each layer after has dropout .5, scale 2
- Weights initialized between (-.005, .005), random uniform distribution
- Kernel shape, pool shape, pool stride for layer 1: (8, 8), (4, 4), (2, 2)
- Kernel shape, pool shape, pool stride for layer 2: (8, 8), (4, 4), (2, 2)
- Kernel shape, pool shape, pool stride for layer 3: (5, 5), (2, 2), (2, 2)
- Maxout group size: 5

V. Results

After training, the CIFAR10 network slightly exceeds the published state of the art for unmodified training data, scoring 86.25% on the test set. In Fig. 3, the training and validation error are shown decreasing over time, and the learned first layer filters are shown in Fig. 4.

The network trained on Asirra and CIFAR10 data combined seems to have very similar results to the network trained on only CIFAR10 data. In fact, it appears the mixing of the two datasets has let the CIFAR10 training dominate the Asirra results, which means the data mixing experiment is largely a failure. This is likely due to excess shrinkage or distortion of the Asirra images in the conversion to 32x32 .png files from much larger .jpg files. The results of this experiment are shown in Figs. 5 and **??**.

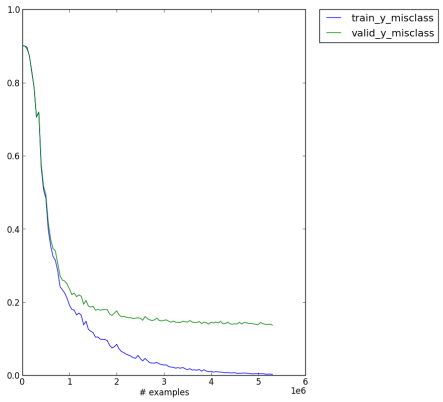

Fig. (3): CIFAR10 error

Fig. (4): CIFAR10 first layer filters

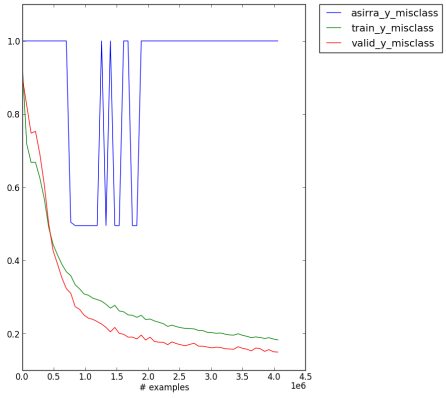

Fig. (5): Asirra + CIFAR10 error

Fig. (6): Asirra + CIFAR10 first layer filters

VI. Conclusions

The applicability of deep convolutional networks for image processing has been well established in both academic literature and industry, with this paper providing further backing. Using a deep convolutional network for the CIFAR10 dataset allows for recognition of objects in these images without feature engineering. This strongly hints that the convolutional network is able to learn effective hierarchical representations of complex data, at least in this particular case. The results of [8] have also been repeated and verified, confirming the improvement provided by maxout layers in dropout networks for image recognition. Unfortunately, the neural network trained on a mix of Asirra and CIFAR10 data did not appear to provide any additional discriminative power for recognition of cats and dogs in the Asirra dataset. The author hopes to

continue work on similar problems throughout this course, expanding the application of neural networks to time-series and less established image datasets.

REFERENCES

[1] Rosenblatt F., *The perceptron, a perceiving and recognizing automaton*. Report 85-460-1, Cornell Aeronautical Laboratory, 1957.

[2] Werbos P.J., *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, 1975.

[3] Krizhevsky A., *Learning Multiple Layers of Features from Tiny Images*, 2009. Retrieved from http://www.cs.toronto.edu/ kriz/cifar.html, 2013.

[4] Elson J., Douceur J.R., Howell J., Saul J., *Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization*, in Proceedings of 14th ACM Conference on Computer and Communications Security (CCS), Association for Computing Machinery, Inc., Oct. 2007. Subset retrieved from http://www.kaggle.com/c/dogs-vs-cats.

[5] Bell A.J., Sejnowski T.J., *The "Independent Components" of Natural Scenes are Edge Filters*, 37(23): 3327-3338, Vision Research, 1997.

[6] LeCun Y., Bengio Y., *Convolutional Networks for Images, Speech, and Time-Series*, The Handbook of Brain Theory and Neural Networks, MIT Press, 1995.

[7] LISA Lab, University of Montreal, *LeNet Tutorials*. Retrieved from http://deeplearning.net/tutorial/lenet.html, 2013.

[8] Goodfellow I., Warde-Farley D., Mirza M., Courville A., Bengio Y., *Maxout Networks*, JMLR WCP 28 (3): 1321-1327, 2013.

[9] Donahue J., Jia Y., Vinyals O., Hoffman J., Zhang N., Tzeng E., Darrell T., *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition*, arXiv 1310.1531, 2013.

[10] Y. Miao, Metze F., S. Rawat *Deep Maxout Networks for Low-Resource Speech Recognition*, in Proceeding of ASRU, Dec. 2013.

[11] Hinton G., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinhov R., *Improving neural networks by preventing co-adaptation of feature detectors*, arXiv 1207.0580, 2012.