Building a Neural Network that Understands Paintings

Chen Liu
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Convolutional neural networks (CNNs) have become a great tool for brain research because of its similarity to the biological processes in visual systems. However, CNNs' performance on abstract paintings, e.g. cubist paintings, is far below human performance. In this paper we will investigate the possible reasons behind such disprepancy between the performances on natural images and cubist paintings of CNNs. We propose that, although able to recognize low level features, CNNs fail to combine the features into higher level concepts. However, experiments have shown that CNNs may not even be able to recognize low level features in cubist paintings. We then propose that Bayesian inference and recurrent connections can be used to solve such problems.

1 Introduction

Convolutional neural networks (CNNs) are a type of feed-forward neural network inspired by the biological processes in animals' visual systems. Each CNN has multiple layers of neurons, each looking at a small part of the image called receptive fields, with overlaps between neighboring neurons. CNNs usually consist of convolution layers, pooling layers and fully connected layers. In a convolution layer, filters of small sizes are convolved with the input image of that layer to extract some specific features. The deeper the layer is in, the more abstract and global the features extracted by the convolution layer are.

CNNs are shown to be similar to how the neurons are organized in the brain; specifically, they are similar to what Hubel and Wiesel found as simple cells and complex cells, and are even shown to explain well the fMRI brain activity [1]. CNNs perform exceptionally well on image classification tasks, and can even beat human performance in the ImageNet classification challenge [2]. CNNs have become a great tool to help us understand how brain works.

Although CNNs work very well on natural images, it still has shortcomings. On cubist paintings, for example, humans are still able to recognize the objects in the image; CNNs, however, will not correctly classify the image as a whole or objects in the image. The top 5 classes for the cubist cat painting in Figure 2(a) classified by an AlexNet are electric fan, comic book, pinwheel, tank suit and stethoscope. On the other hand, as long as the painting is not too abstract, CNNs are still able to recognize the objects despite some odd features. In Figure 2(b), the cat eyes are duplicated, but AlexNet is still able to classify the image as an Egyptian cat.

In the next sections, we will provide some hypotheses that could explain such phenomena, and experiments on how to tweak the CNN so that it works more like brains.

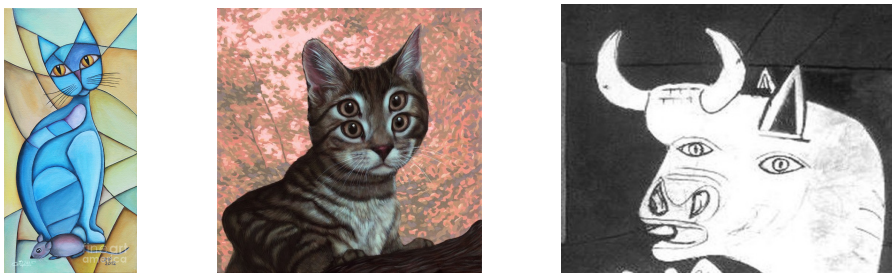

Figure 1: (a) Cubist cat painting; (b) Cat painting with odd features; (c) An ox in Picasso's Guernica

1.1 Previous Work

Ginosar, et al. [3] presented an evaluation of CNN and human performance on recognizing people in cubist art. They found that although human perception significantly outperforms the CNN, they both exhibited a similarly graceful degradation in performance as the objects become more deformed and abstract.

2 Method

As the goal is to build a CNN that works more similar to how brain works, we may want to investigate whether humans are able to recognize objects in a cubist painting on first sight, even s/he has never seen paintings of such style before, or have to be instructed first. In the first case, we may similarly fine-tune the CNN on images of cubist paintings and objects. In the second case, we have to tweak the architecture of the CNN because it means that there is some mechanism that human brains have but CNNs do not.

Due to lack of related research support, we assume the second case, that it is human's innate ability to recognize objects in cubist paintings. Note that even the first case is true, most people see far fewer cubist cats and dogs paintings compared to natural images of cats and dogs while still able to recognize those cubist objects, but CNNs will not perform very well on few training examples. Therefore we want to investigate the reason behind the poor performance and tweak the architecture accordingly.

We propose that the reason behind the poor performance is that although the CNNs are able to recognize low level features in the image, it fails to combine the features to form higher level concept. For example, in Figure 2(c), the picture has all low level features an ox may have: eyes. noses, mouth etc. However, one eye of the ox is misplaced on the neck instead of on the face, and the concept formed in CNN may enforce the requirement that eyes have to be on the front of the face for the object to be recognized.

To verify this hypothesis, we want to check if the CNN is able to detect the low level features. In this paper, we use AlexNet, one implementation of CNNs, for experimentation. Details of the AlexNet architecture can be found in [4]. To find if the AlexNet recognizes eyes, noses etc, we need to first find the neurons responsible for recognizing the objects in later layers of the network, e.g. the 5th convolution layer and the fully connected layers. Due to the fact that AlexNet uses distributed coding, it is hard to recognize a single neuron that will only activate when the input image is an eye or noses. We then train a support vector machine on training images of eyes, noses and other images to classify the later layer responses of the AlexNet into these three classes. With such a classifier, we can go back to the cubist paintings of objects to see if the network sees the low level features.

3 Experiments

We use Caffe with an AlexNet model pre-trained with the ILSVRC 2012 dataset for experiments. To find the eye neuron, we first feed an image of an eye to AlexNet and observe its response. Because the receptive fields of earlier layers of AlexNet are small, we shrink the eye images to images with a height of 67 pixels and the same height-width ratio, and inscribe the smaller images in a 400x400 white background.

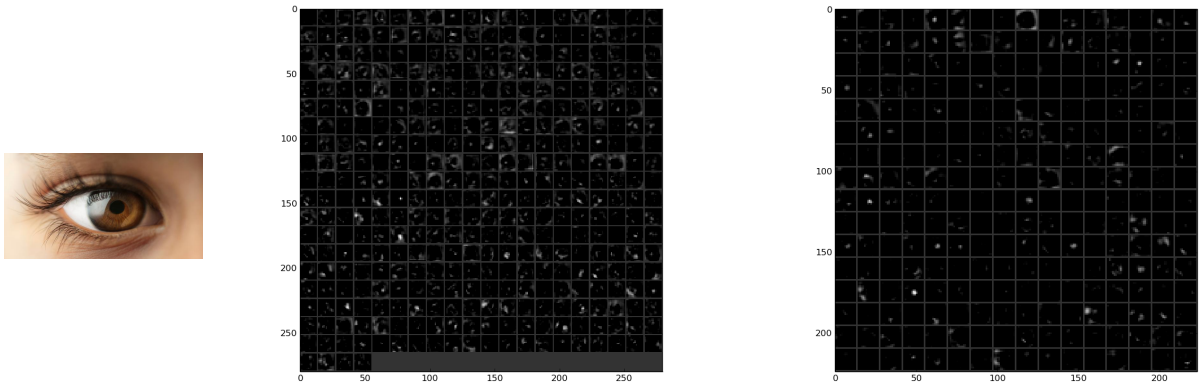

Figure 2: (a) An eye image; (b) 4th convolution layer response; (c) 5th convolution layer response

Notice that the firing pattern is rather noisy, and we are not able to determine a fixed set of neurons that are responsible for encoding the eye concept by inspecting the later layers responses of multiple eye images.

We then explore classifying the responses of AlexNet with a support vector machine (SVM). The dataset contains 20 eye images (6 human eyes, 14 eyes of other animals), 20 noses images (7 human noses, 13 noses of other animals), and 20 images of other categories (10 images of body parts, 10 images of whole objects, e.g. ox, house, airplane etc.). We use the support vector machine implementation in scikit-learn library for experiments. We found that among the later layers of AlexNet, classifying on the responses of fc6 (the first fully connected layer) provides the best classification results: the 60 training examples are all correctly classified to their labels after training, indicating that there exists a perfect decision boundary between the responses of eyes, noses and other images.

Despite the success in classifying later layer responses to three categories, the trained SVM still fails to recognize low level features in whole objects. For example, the SVM classifies the response to Figure 2(a) as "others" instead of recognizing that there are eyes or noses in the image. It also fails to classify small images of eyes and noses from cubist paintings. This could be because of low information carried on the cubist body parts, compared to the rich information on the images that the SVM was trained with.

As we fail to confirm the hypothesis that the CNN recognizes low level features, we have to investigate other ways to make CNNs recognize objects in cubist paintings.

4 Future Work

One way to deal with the problem of low information is to utilize context of the object Bayesian inference. Doersch et al [5] have shown that contexts of objects in an image can be helpful for visual concept learning. Jia

et al. [6] have shown that Bayesian inference can be incorporated into common computer vision algorithms for generalizing concept hierarchies. We may experiment next using objects around the one that the CNN is currently looking at for better inference. For example, since eyes and noses usually occur together, when the CNN indicates that it is possible for one part of the image to be an eye while the region near it may contain a nose, then the probabilities for the first area to contain an eye and the second area to contain a nose are both boosted.

Recurrent connections may also be helpful in that concepts in later layers of the network can be used to boost probabilities of seeing low level features in previous layers. Since CNNs are purely feed-forward, while human brains have been shown to have many recurrent connections, they could play a critical role in recognizing objects in cubist paintings.

5 Summary and Conclusions

In this paper, we explored the reason behind convolutional neural networks' incapability of recognizing objects in cubist paintings, despite amazing performance in classifying natural images. We proposed that failure to combine low level features into higher level concepts due to their odd way of organization could be the main cause, given that the CNNs are still able to recognize such low level features. However, by training a classifier on later layer responses of AlexNet with support vector machine, we found that AlexNet may not be able to recognize those low level features. Future work has been proposed to utilize the context and Bayesian inference to boost the probability of these features, and recurrent connections may also be used so that higher level concepts can be used to boost probability of low level features.

Acknowledgments

The author thanks Professor Taising Lee for advising, directed reading and providing valuable feedback throughout the research this semester. I also thank Yimeng Zhang for technical assistance regarding the use of deconvolutional neural networks and visualization.

References

[1] Ramakrishnan, Kandan, et al. "Convolutional Neural Networks in the Brain: an fMRI study." *Journal of vision* 15.12 (2015): 371-371.

[2] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *arXiv preprint* arXiv:1512.03385 (2015).

[3] Ginosar, Shiry, et al. "Detecting people in Cubist art." *Computer Vision-ECCV 2014 Workshops*. Springer International Publishing, 2014.

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[5] Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised Visual Representation Learning by Context Prediction." *arXiv preprint* arXiv:1505.05192 (2015).

[6] Jia, Yangqing, et al. "Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies." *Advances in Neural Information Processing Systems*. 2013.

[7] Deco, Gustavo, and Tai Sing Lee. "The role of early visual cortex in visual integration: a neural model of recurrent interaction." *European Journal of Neuroscience* 20.4 (2004): 1089-1100.