Convolutional Neural Networks for Scene Recognition

Bavin Ondieki
Stanford University
CS 231N Final Project
ondieki@cs.stanford.edu

Abstract

This project is an attempt to apply the power of neural networks to detect scenes. The experiments make use of MIT Indoor 67 and SUN 397 datasets with an aim to see how far CNNs can match the current industry standards in scene recognition. Techniques such as fooling the CNN helped boost performance by reducing confusion between the most confused pairs of classes. In addition, the project compares the difference in properties between indoor-centric vs outdoor-centric datasets, which are responsible for the marked difference in performance on similar CNN architectures.

1. Introduction

A real time, robust scene recognition engine would be a valuable tool for digital marketers, given the increasing graphic media content on the Internet. For instance, digital marketers e.g. Pinterest/Instagram would be interested in knowing a consumer's favorite hangout spot e.g. a bar, bowling alley, bakery etc based on his/her Instagram photo uploads. This information would help an advertiser target his/her clients more accurately thus saving money.

1.1. Dataset

The project makes use of subsets of MIT Indoor67 and SUN 397 datasets.

1.2. Background

We have had state-of-the-art performance on a model built by MIT researchers based on the Places 205 Dataset, which comprises of 2.5 million images[3]. They used a hybrid of techniques to achieve about 70.8% accuracy. One of the important works in scene recognition was aimed at building a huge dataset, the Places dataset with over 7 million images [3]. They outperfomed prior achievements based on fine-tuning object-centric trained models (on ImageNet). The creation of scene-centric datasets has spurred interest and growth in scene recognition. This paper is attempts to uncover the differentiating factors between indoor and outdoor scenes. The SUN 397 and Indoor 67 models in these results are built from the scratch, as opposed to fine-tuning pre-existing models from out there. One of the biggest efforts involved understanding the nature of both outdoor-centric and indoor-centric datasets using supervised learning methods, and applying techniques to boost performance.

2. Technical Approach

The problem of identifying a scene can be approached in two ways [3].

1. Training the CNN on an object-centric dataset
Indoor scenes contain objects e.g. a casino would contain chairs, desktop, bottles and so on. Therefore, the task of indoor scene recognition is at least as hard as object recognition. Figure 8 shows the dataset under SVD using PCA.

2. Training the CNN on scene-centric dataset The SUN 397 Places dataset consists of 397 scenes. The SUN 397 subset consisting of outdoor scenes that was used for the experiments is also referred to as 'outdoor-centric dataset' in this paper.

I expected to find performance patterns similar to Bolei's work. This would imply that their results support our claim: that a pre-trained model of outdoor scene-centric data provides at least as good a performance on indoor scenes after fine-tuning - saving much training time. The second claim is that transfer learning from an indoor model to an outdoor scene-centric dataset does not yield an equally good match in performance. The results I found seem to match this described pattern, which also is evident in the results obtained by the MIT experiments (See MIT results on page 5).

2.1. Results from my experiments

CNN	MIT Indoor 67	SUN Dataset
Indoor 67 Trained CNN	43.89	58.33
SUN 397 Trained CNN	42.09	67.02

Figure 1. MIT Indoor67 Dataset. Graphic courtesy of Prof. Antonio Torralba, MIT [1]

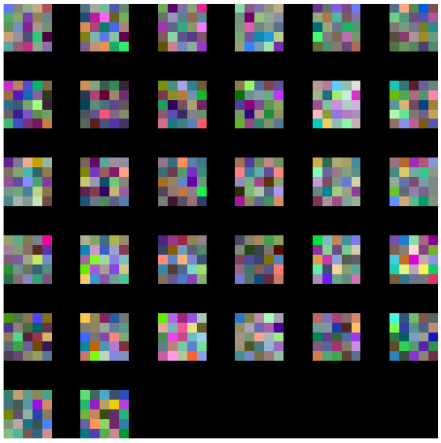

Figure 2. Using Convnets: Two Layer CNN on MIT Indoor 67

2.2. Convolutional Neural Networks for Scene Recognition

Convolutional neural networks help us simulate human vision, which is amazing at scene recognition. They enforce local connectivity between neurons in adjacent layers[2]. Thus, CNNs exploit spatially-local correlation. The neurons are replicated across the visual space so that features can be recognized regardless of position in visual field. Another effect is that the weights are shared. Weight sharing increases efficiency by decreasing the number of learnable parameters by gradient descent [2].

3. Supervised Learning methods for Data Understanding

Supervised learning methods turn out to be essential tools for data understanding. For this experiments, the simpler supervised learning methods used: KNN, Random Forest, SVM had a low accuracy of about 20%. Before classification, the data is first transformed to SIFT space. Each image has varying number of keypoints, that were averaged to return a single image descriptor. All images in SIFT space are mean-centered. SIFT keypoints do not pay attention to spatial and local correlation. Rather, the images are converted into visual words. But the results are still wanting, we can do more, by taking into account spatially-local correlation, therefore, CNNs are bound to outperform these simpler methods.

Supervised machine learning algorithms were crucial to data understanding. The insight gained helped guide the image pre-processing steps, and informed the decisions made for dataset augmentation to boost performance. (Noise perturbation improved performance - look at the last section to see this)

For instance, using the SVM, I was able to notice that indoor recognition dataset had a lot of interclass correlation.

Consider the confusion matrix in figure 3 above, obtained by first extracting the SIFT keypoints of the images and using an SVM classifier.

The data in SIFT space is projected into PCA space (Figure 6)

The outdoor scene dataset is quite noisy (check out the mean-centered image in figure 4) and correlated as well [fig. 5]. These factors are contributing causes to loss in accuracy/class differentiation.

Consider also the plot of the SIFT features after dimensionality reduction, projected onto PCA space.

The data seems quite correlated, and it was quite distinct from the outdoor dataset under the same transformation i.e. svd and projection in PCA space

3.1. Feature Importance Using Random Forests

Understanding relative feature importance was crucial to understanding the problem much better. To do this, I extract the SIFT features and obtain images representation in SIFT space. Fitting 250 trees to the transformed dataset produces a ranking of the most important features that contribute to the classification of the dataset. The relatively uniform importance of all the top 20 features in the indoor dataset is

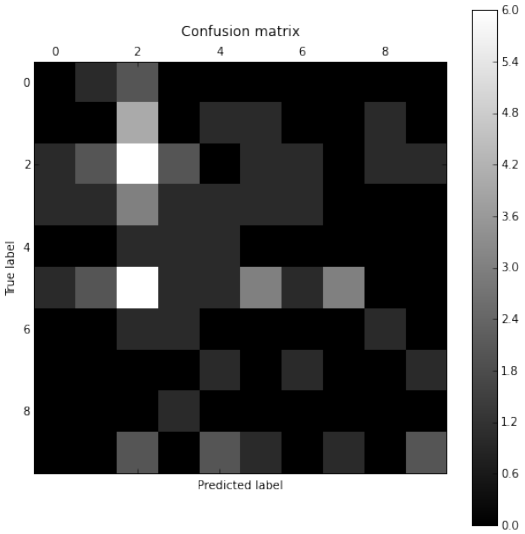

Figure 3. Confusion Matrix from Running an SVM on data in SIFT space

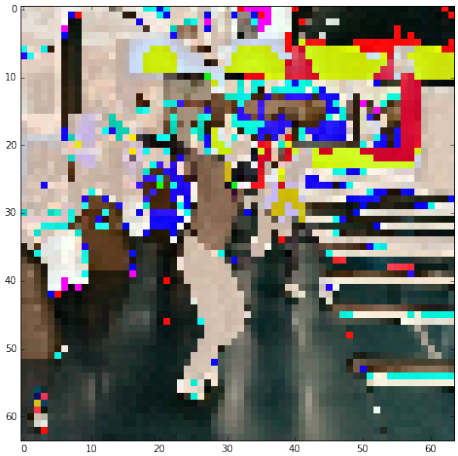

Figure 4. Mean centered image: MIT Indoor 67

distinct from the fairly non-uniform importance of features in the outdoor-scene data. Look at figures 7 and 10 to see the difference.

I fit 250 trees to both datasets so as to get an unbiased comparison of the two datasets.

4. Classification Pipeline

The first part of the classification pipeline consisted of a two layer convolutional neural network, with dropout, max pooling for shift invariance, and ReLu to ensure that the network learns the weights well and the other layers - fully connected and input layers.

The first thing that I did was to resize all my images to 64 x 64. The two layer network was effective for extracting

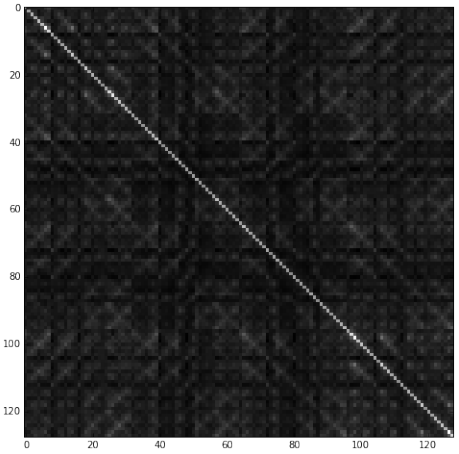

Figure 5. Covariance Matrix: MIT Indoor 67

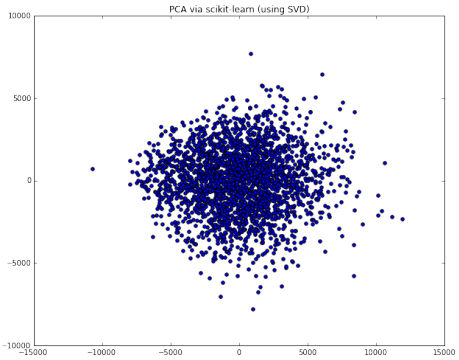

Figure 6. MIT Indoor 67 SIFT keypoints in PCA space

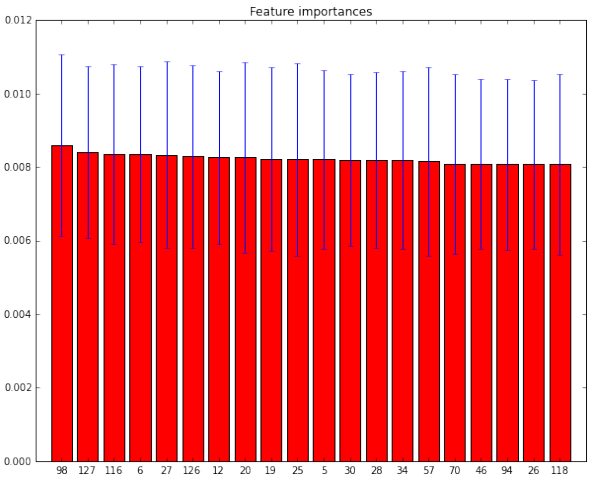

Figure 7. Random Forests Top 20 Features: MIT Indoor 67

some of the key components of the images, even though it was not as fine grained as possible due to the noisy nature of the dataset. This was evident from the noisy/grainy look-

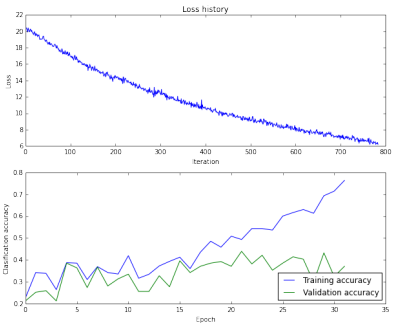

Figure 8. Accuracy and Loss History: Training Convnet on MIT Indoor 67

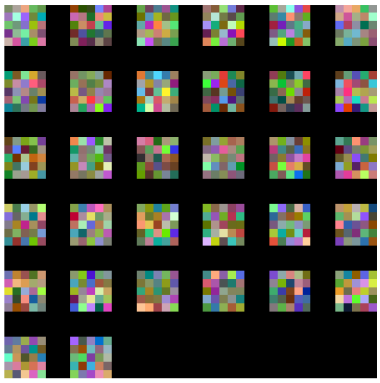

Figure 9. SUN 397 Subset: Filter Plot

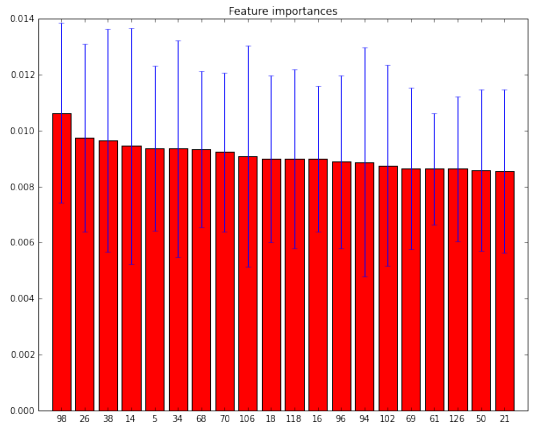

Figure 10. Fitting a Random forest of 250 trees to SUN 397 dataset

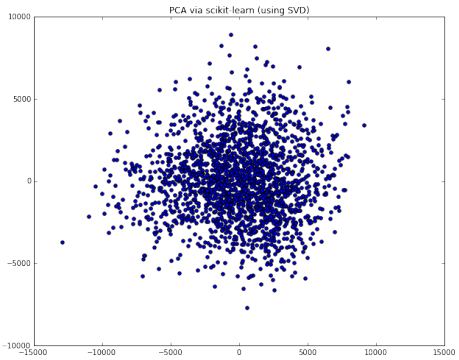

Figure 11. SUN 397 features in PCA space after dimensionality reduction using SVD

ing filters that were obtained after training for about 16-20 epochs.

Loss History, Training Accuracy vs Validation Accuracy Plot

4.1. Results from my experiments

CNN	MIT Indoor 67	SUN Dataset
Indoor 67 Trained CNN	43.89	58.33
SUN 397 Trained CNN	42.09	67.02

5. Analyzing a Subset of Outdoor Scene-Centric Dataset from SUN 397

How different are outdoor datasets from indoor datasets, so that they perform better?

The histogram in figure 10 shows the relative feature importance of the outdoor-scene centric dataset drawn from SUN 397.

The SUN-397 dataset showed better segmentation after mean centering the data [fig 12] than the indoor dataset as you can see in figure 4.

Consider the plot of the SIFT keypoint features of SUN 397 dataset, under SVD and projection into PCA space [fig 11], we see a much better de-correlation in space, which means that a multi-class SVM classifier would do better on it. In terms of CNNs, it we can say that the better segments obtained after mean centering the image make it easier to learn robust weights and features for classification in the case of outdoor scenes.

The PCA plot shows that the data is less correlated compared to the plot of the indoor feature dataset undergoing the same transformation [fig 6]. But again, there isn't much difference in terms of the feature importances, when we fit both datasets, of about the same size, with 250 trees. The slightly visible difference is that the indoor dataset top 20 features have about the same importance relative to the top 20 features in the outdoor dataset.

Training a two layer CNN on the SUN dataset using the same architecture as before. A much better accuracy of over 67% is achieved within a few epochs of training [figure 13].

6. Transfer Learning

Having obtained great results from training on SUN 397, I sought to attempt boosting performance on the MIT Indoor

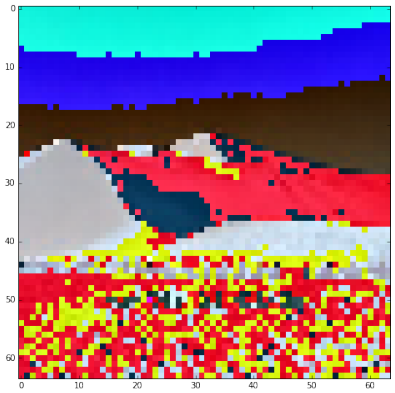

Figure 12. SUN 397 sample image

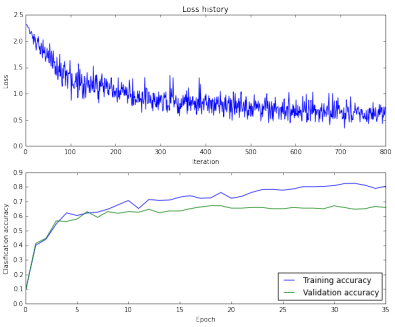

Figure 13. SUN 397: Loss History and Accuracy

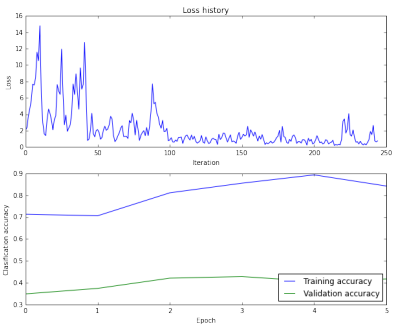

Figure 14. Transfer Learning: From outdoor-centric to indoor-centric datasets

dataset, or at least, reducing the training time by initializing the first layer weights using weights from the scene-centric trained CNN model. I discovered that the training time reduced significantly leading to equally good results compared to the model that was trained on the MIT Indoor dataset. This showed that a model trained on a good outdoor dataset provides great highlevel features from which to fine-tune in order to build a classifier for indoor scene recognition.

Results from Experiments:

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

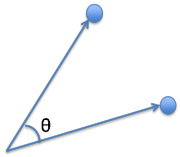

Figure 15. Cosine Similarity of feature vectors: Outdoor-centric vs indoor-centric dataset

CNN	MIT Indoor 67	SUN Dataset
Indoor 67 Trained	43.89	58.33
SUN 397	42.09	67.02

6.1. Feature Similarity Between Indoor and Outdoor Scenes

To find the similarity between the feature set in the two datasets, we first extract the SIFT keypoints from the images, and use random forests to obtain the relative importance of the SIFT keypoints/features in classifcation. Having obtained the most important features for classification in the form of feature indices, I used the cosine similarity as a measure of similarity between the top 20 features from both datasets (figure 15).

The value was 0.604, implying a high level of similarity between the two feature sets, but the difference, as we've seen, causes much difference.

This promixity in feature space is due to the largely similar high level structures that are common in both. The performance of the CNN with transfer learning does not exceed the previous bench mark at all, but helps the training to run much faster to converge to about the same result as before.

The MIT dataset also shows that the performance is just about the same when transfer learning from a less complex model to a more complex model.

Results from MIT CNN models:

CNN	MIT Indoor 67	SUN Dataset
Places CNN	54.32 (BM 47.2)	68.24 (BM 66.87)
ImageNet-CNN	42.61 (BM 47.2)	56.79 (BM 66.87)

(BM value) indicates the benchmark for each datset at the time that the MIT team performed their experiment. The architectures for the two convnets in the MIT experiment were the same, so the results are comparable: we cannot attribute superior performance to differences in architecture, rather, rather, it is a result of the nature of the weights learned.

6.2. Indoor Scene Recognition at least as hard as object recognition tasks

From the work done to understand the properties of the indoor scenes, it became clearer that indoor scenes are often charactized by relatively more clutter and are generally

quite noisy. Indoor scenes have a greater level of complexity than outdoor scenes that needs to be learned properly: not too specific - leading to overfitting, or too generic, leading to underfitting. This complexity, I'd argue, makes them at least as hard as object recognition. With this claim in mind, I'll quickly touch on the MIT experiment results to help clarify and provider further evidence, besides my own results, that indeed outdoor to indoor transfer learning provides about the same results, if not better, but the converse is not true. Look at the MIT results to the right of page 5 above.

The CNN trained on scene-centric data (Places CNN), is used for transfer learning to the MIT Indoor 67 dataset, achieving 2% higher than the benchmark. On the other hand, using the ImageNet trained model, which in this case represents a slightly simpler version of an indoor model - to the places dataset - yields a reduction in 5% from the bench mark. This implies that the hypothesis learned from an indoor model is not well generalizable and good enough model for the outdoor scene - which is simpler and less complicated[3].

Transfer learning yielded a 58.3 % accuracy, down from the original 67% obtained from training the model from the ground up using the same parameters. I used fewer epochs but final accuracy is about 8.7% lower than expected. This reduction is expected if we go by the observations from the results in the MIT paper.[3]

6.3. Understanding the Confusion Matrix for SUN 397 Dataset

Consider the confusion matrix of the SUN 397 dataset [fig. 16]. As you can notice, pairwise confusion counts are much lower than the Indoor 67 dataset. Classes 0 and 2 are the most confused classes with a count much lower than the most confused classes in Indoor 67 dataset. There is 4 times as much confusion in the Indoor case versus the outdoor case when you compare the most confused pair of classes in the two datasets.

7. 'Ethical CNN Fooling' for Performance Augmentation

This section is an attempt to take advantage of the vulnerability of CNN to fooling, so as to minimize interclass confusion. To do this, we find the pair of classes that are most confused, and build a fooling image, with the mean image of the misclassified images as the start. In this case, we want to fool the CNN to boost and not break performance, hence the use of the term 'ethical'.

From the subset of the data used, the bar and the casino were the most confused classes [figure 17].

In their paper on fooling deep neural networks, Nguyen, Anh et. al showed that it's quite easy to construct an image

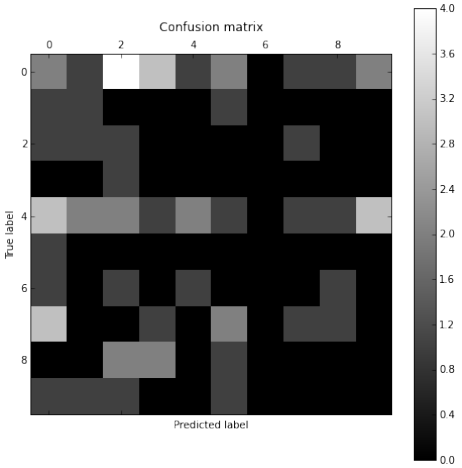

Figure 16. SUN 397 Dataset Confusion Matrix

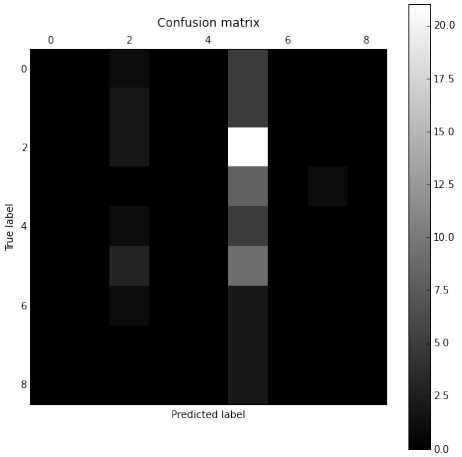

Figure 17. MIT Indoor 67: Bar and Casino most confused

that fools the neural network [5]. One of the experiments involved augmenting the confused image classes with some noise so as to reduce confusion and boost performance. It did result in improved results by only dealing with the two most confused classes - by adding some noise to the casino class images. Creating the fooling image is an optimization problem - done using gradient descent the starting image as the average of the misclassified images. The goal is to 'fool', or encourage the convnet to tweak the weights so that the starting image is correctly considered a casino with reasonably high confidence of 0.70. This value is arbitrary but convenient: we add just enough noise to gain enough confidence, not too little to produce little change, and not too much to visibly or excessively perturb the look of images.

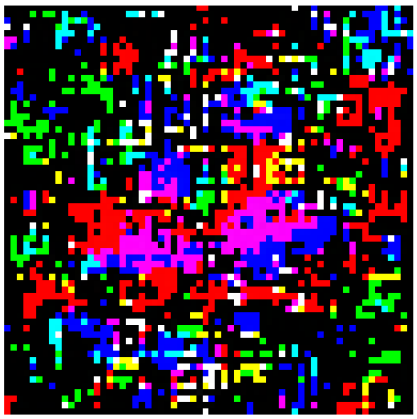

Figure 18. Noise Added to Minimize Confusion between Casino and Bar

Optimization problem [5]:

$$x_f = \arg\min_x \left(L(x, y, m) + \frac{\lambda}{2} \|x - x_0\|_2^2 \right)$$

(Nguyen et. al)

Suppose that $p(x = y \mid m)$ is the probability that the input x is assigned the label y under the model m. We then specify a desired *confidence threshold* t for the engineered image, and stop optimization when $p(x_f = y \mid m) >= t$ [5]

7.1. Results of 'Ethical CNN Fooling'

Adding the noise : 46.40% accuracy

Without the noise: 42.80% accuracy

Thus, 'ethical' fooling can be an essential tool to boosting performance on the most confused pairs of classes.

8. Conclusions

Transfer learning to an indoor dataset speeds up training and gives about the same performance. A model obtained from a less complex dataset works as a good starting point for the indoor scene dataset, converging quickly and giving equally good performance as full training(+/- 2%) difference in performance. But there is the possibility of achieving better results with more fine-tuning of the pre-trained model.

Ethical fooling of the CNN, using the optimization outlined by Nguyen et. al is a useful tool to reduce confusion and boost performance in indoor scene recognition. It helped boost performance by about 4%

Model trained on indoor-centric data causes reduction in performance on an outdoor scene - thus there is need to come up with inexpensive techniques to lessen the complexity of this model. MIT's result, as well as mine, were many points (8.7%) below the performance attained by training from the ground up.

9. Future Work

It'd be great to work on model ensembles, combining indoor, outdoor, as well as perturbing most confused classes with varying amounts of noise with an aim to augment performance.

10. References

[1] Antonio Torralba, MIT

[2] DeepLearning.net

[3] Learning Deep Features For Scene Recognition using Places Database, Bolei Zu et. al

[4] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint, 2013.

[5] Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images." arXiv preprint, 2014.