# Analyzing the Performance of Multilayer Neural Networks for Object Recognition

Pulkit Agrawal, Ross Girshick, Jitendra Malik
{pulkitag,rbg,malik}@eecs.berkeley.edu

University of California Berkeley

Supplementary Material

## 1 Effect of fine-tuning on CNN parameters

In the main paper we provided evidence that fine-tuning a discriminatively pre-trained network is very effective in terms of task performance. We also provided insights into how fine-tuning changes its parameters. Here we describe and discuss in greater detail some metrics for determining the effect of fine-tuning.

### 1.1 Defining the measure of discriminative capacity of a filter

The entropy of a filter is calculated to measure its discriminative capacity. The use of entropy is motivated by works such as [1], [2]. For computing the entropy of a filter, we start by collecting filter responses from a set of $N$ images. Each image, when passed through the CNN produces a $p \times p$ heat map of scores for each filter in a given layer (e.g., $p = 6$ for a conv-5 filter and $p = 1$ for an fc-6 filter). This heat map is vectorized (`x(:)` in MATLAB) into a vector of scores of length $p^2$. With each element of this vector we associate the class label of the image. Thus, for every image we have a score vector and a label vector of length $p^2$ each. Next, the score vectors from all $N$ images are concatenated into an $Np^2$-length score vector. The same is done for the label vectors. We define the entropy of a filter in the following three ways.

**Label Entropy.** For a given score threshold $\tau$, we define the *class entropy of a filter* to be the entropy of the normalized histogram of class labels that have an associated score $\geq \tau$. A low class entropy means that at scores above $\tau$, the filter is very class selective. As this threshold changes, the class entropy traces out a curve which we call the *entropy curve.* The *area under the entropy curve* (AuE), summarizes the class entropy at all thresholds and is used as a measure of discriminative capacity of the filter. The lower the AuE value, the more class selective the filter is.

**Weighted Label Entropy.** While computing the class label histogram, instead of the label count we use the sum of the scores associated with the labels to construct the histogram. (Note: Since we are using outputs of the rectified linear units, all scores are $\geq 0$.)

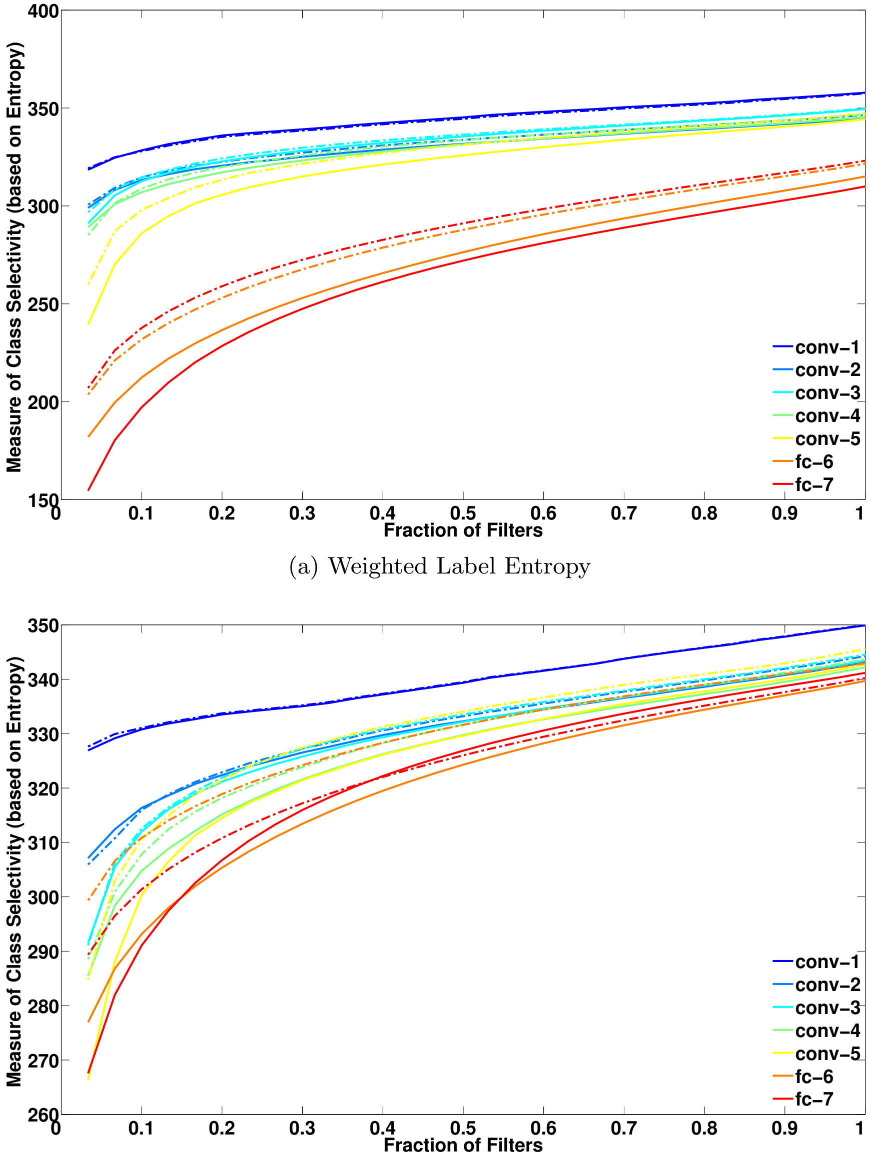
**Spatial-Max (spMax) Label Entropy.** Instead of vectorizing the heatmap, the filter response obtained as a result of max pooling the $p \times p$ filter output is associated with the class label of each image. Thus, for every image we have a score vector and a class label vector of length 1 each. Next, the score vectors from all $N$ images are concatenated into an $N$-length score vector. Then, we proceed in the same way as for the case of Label Entropy to compute the AuE of each filter.

### 1.2 Defining the measure of discriminative capacity of a layer

The discriminative capacity of layer is computed as following: The filters are sorted in increasing order of their AuE. Next, the cumulative sum of AuE values in this sorted list is calculated. The obtained list of Cumulative AuEs is referred to as CAuE. Note that, the $i$-th entry of the CAuE list is the sum of the AuE scores of the top $i$ most discriminative filters. The difference in the value of the $i$-th entry before and after fine-tuning measures the change in class selectivity of the top $i$ most discriminative filters due to fine-tuning. For comparing results across different layers, the CAuE values are normalized to account for different numbers of filters in each layer. Specifically, the $i$-th entry of the CAuE list is divided by $i$. This normalized CAuE is called the Mean Cumulative Area Under the Entropy Curve (MCAuE). A lower value of MCAuE indicates that the individual filters of the layer are more discriminative.

Table 1: This table lists percentage decrease in MCAuE as a result of finetuning when only 0.1, 0.25, 0.50 and 1.00 fraction of all the filters were used for computing MCAuE. A lower MCAuE indicates that filters in a layer are more selective/class specific. The 0.1 fraction includes the top 10% most selective filters, 0.25 is top 25% of most selective filters. Consequently, comparing MCAuE at different fraction of filters gives a better sense of how selective the "most" selective filters have become. A negative value in the table below indicates increase in entropy. Note that for all the metrics maximum decrease in entropy takes place while moving from layer 5 to layer 7. Also, note that for fc-6 and fc-7 the values in Label Entropy and spMax Label Entropy are same as these layers have spatial maps of size 1.

| Layer | Label Entropy | | | | Weighted Label Entropy | | | | spMax Label Entropy | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.25 | 0.5 | 1.0 | 0.1 | 0.25 | 0.5 | 1 | 0.1 | 0.25 | 0.5 | 1.0 |
| conv-1 | −0.02 | −0.14 | −0.19 | −0.19 | 0.06 | −0.13 | −0.16 | −0.16 | 0.19 | 0.10 | 0.07 | 0.04 |
| conv-2 | −0.71 | −0.31 | −0.14 | 0.01 | 0.41 | 0.53 | 0.58 | 0.57 | −0.39 | −0.03 | 0.11 | 0.23 |
| conv-3 | −1.14 | −0.86 | −0.67 | −0.44 | 1.11 | 0.66 | 0.52 | 0.32 | 0.14 | 0.20 | 0.32 | 0.33 |
| conv-4 | −0.54 | −0.31 | −0.19 | −0.05 | −0.10 | 0.55 | 0.64 | 0.57 | 0.93 | 0.97 | 0.80 | 0.65 |
| conv-5 | 0.97 | 0.55 | 0.43 | 0.36 | 5.84 | 3.53 | 2.66 | 1.85 | 4.87 | 3.05 | 2.31 | 1.62 |
| fc-6 | 6.52 | 5.06 | 3.92 | 2.64 | 9.59 | 7.55 | 6.08 | 4.27 | 6.52 | 5.06 | 3.92 | 2.64 |
| fc-7 | 5.17 | 2.66 | 1.33 | 0.44 | 20.58 | 14.75 | 11.12 | 7.78 | 5.17 | 2.66 | 1.33 | 0.44 |



(a) Weighted Label Entropy

(b) Spatial-Max Label Entropy

Fig. 1: PASCAL object class selectivity (measured as MCAuE) plotted against the fraction of filters, for each layer, before fine-tuning (dash-dot line) and after fine-tuning (solid line). A lower value indicates greater class selectivity. (a),(b) show MCAUE computed using Weighted-Label-Entropy and Spatial-Max Label-Entropy method respectively.

### 1.3 Discussion

The MCAuE measure of determining layer selectivity before and after fine-tuning is shown in Figure 1 for Weighted Label and Spatial-Max Label Entropy. Results for Label Entropy method are presented in the main paper. A quantitative measure of change in entropy due to finetuning, computed as percentage change is defined as following:

$$\text{Percent Decrease} = 100 \times \frac{MCAuE_{pre} - MCAuE_{fine}}{MCAuE_{pre}} \tag{1}$$

where, $MCAuE_{fine}$ is for fine-tuned network and $MCAuE_{untuned}$ is for network trained on imagenet only. The results are summarized in table 1.

As measured by Label Entropy, layers 1 to 5 undergo negligible change in their discriminative capacity, whereas layers 6-7 become a lot more discriminative. Whereas, the measures of Weighted Label and Spatial-Max Label Entropy indicate that only layers 1 to 4 undergo minimal changes and other layers become substantially more discriminative. These results confirm the intuition that lower layers of the CNN are more generic features, whereas fine-tuning mostly effects the top layers. Also, note that these results are true for fine-tuning for moderate amount of training data available as part of PASCAL-DET. It is yet to be determined how lower convolutional layers would change due to fine-tuning when more training data is available.

## 2 Are there grandmother cells in CNNs?

In the main paper we studied the nature of representations in mid-level CNN representations given by conv-5. Here, we address the same question for layer fc-7, which is the last layer of CNN and features extracted from this lead to best performance. The results for number of filters required to achieve the same performance as all the filters taken together is presented in Figure 2. Table 2 reports the number of filters required per class to obtain 50% and 90% of the complete performance. It can be seen that like conv-5, feature representations in fc-7 are also distributed for a large number of classes. It is interesting to note, that for most classes 50% performance can be reached using a single filter, but for reaching 90% performance a lot more filters are required.

## References

1. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (Oct 2001), http://dx.doi.org/10.1023/A:1010933404324
2. Geman, D., Amit, Y., Wilder, K.: Joint induction of shape features and tree classifiers (1997)

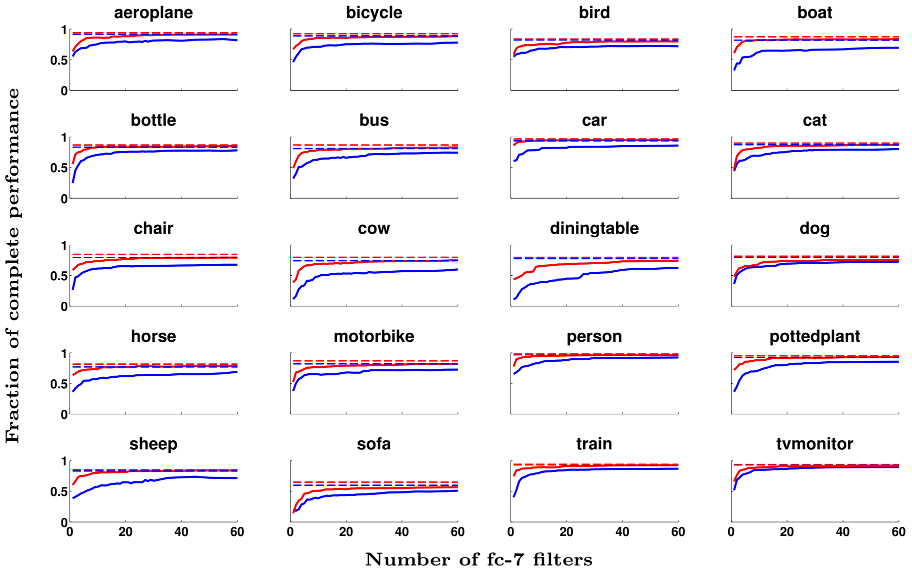


Fig. 2: The fraction of complete performance on PASCAL-DET-GT achieved by fc-7 filter subsets of different sizes. Complete performance is the AP computed by considering responses of all the filters. The solid blue and red lines are for pre-trained and fine-tuned network respectively. The dashed blue and red lines show the complete performance. Notice, that for a few classes such as person, bicycle and cars only a few filters are required, but for many classes substantially more filters are needed, indicating a distributed code.

Table 2: Number of filters required to achieve 50% or 90% of the complete performance on PASCAL-DET-GT using a CNN pre-trained on ImageNet and fine-tuned for PASCAL-DET using fc-7 features.

| | perf. | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pre-train | 50% | 1 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | 2 | 6 | 11 | 2 | 2 | 2 | 1 | 3 | 3 | 5 | 2 | 1 |
| fine-tune | 50% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 |
| pre-train | 90% | 33 | 40 | 40 | 40 | 17 | 32 | 32 | 31 | 40 | 40 | 40 | 35 | 37 | 37 | 17 | 29 | 40 | 40 | 17 | 8 |
| fine-tune | 90% | 6 | 7 | 11 | 4 | 5 | 10 | 2 | 8 | 19 | 27 | 32 | 18 | 9 | 16 | 2 | 7 | 7 | 40 | 3 | 4 |