Detecting avocados to zucchinis: what have we done, and where are we going? Supplementary material

Olga Russakovsky[1], Jia Deng[1], Zhiheng Huang[1], Alexander C. Berg[2], Li Fei-Fei[1]
Stanford University[1] , UNC Chapel Hill[2]

Abstract

This document contains additional details of the large-scale object detection analysis on ILSVRC2012 dataset.

1. Evaluation criteria: number of guesses

An important decision that was made in the analysis is how many guesses to allow an algorithm to make per image.

1.1. Top-5 evaluation of algorithms

Section 2.1 of the paper described the top-5 evaluation criteria used on the ILSVRC dataset: an algorithm is allowed to make up to 5 guesses per image without penalty. We briefly justify this choice.

Figure 1 plots the accuracy of the different methods as a function of the number of guesses. Here we consider just the state-of-the-art algorithms, SV and VGG (ignoring the black curves). As the algorithms are allowed to make between 1 and 5 guesses, the relative performance remains reasonably consistent: the difference in classification accuracy between the two methods ranges from 0.108 and 0.117, and the difference in classification+localization accuracy ranges from 0.140 to 0.160. Since these patterns are consistent, we follow the intuition of the ImageNet challenge evaluation (the images are not exhaustively labeled, so unannotated objects may be present and thus the algorithms should not be penalized for potentially predicting an unlabeled object as the top scoring detection) and use top-5 evaluation in our analysis. Similar conclusions can be drawn from the data when using just top-1 evaluation.

1.2. Top-10 evaluation of upper bound

Section 2.3 of the paper presented an upper bound of the two state-of-the-art methods, which combines the outputs of the VGG and SV on every image and considers the object to be correctly detected if any of the 10 proposed (class, location) pairs is correct. Here we provide some analysis and insight.

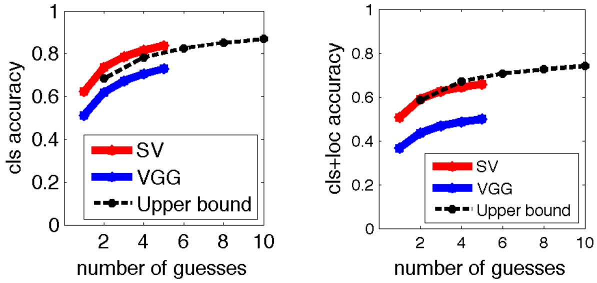


Figure 1. (a) Classification and (b) classification with localization accuracies of the three methods as a function of the number of guesses allowed during evaluation on ILSVRC2012.

The idea is to put an upper bound on how well a combination of the two systems might work – for example, given an oracle which would select the "best" box from the 10 boxes proposed by SV and VGG (in this setting, top-10 evaluation would actually the same as top-1).

Figure 1 shows the number of guesses versus accuracy for the upper bound (in black) by taking the top 1, 2, 3, 4, 5 guesses from each method, so 2, 4, 6, 8, 10 guesses. It is important to keep in mind that the scores between the two algorithms are uncalibrated, so this upper bound is in fact suboptimal. However, a few interesting trends are still worth noting:

- For classification, SV is an impressively strong algorithm: given a budget of only 5 guesses it's better to use the top-5 guesses from SV (accuracy of 0.838) rather than combining it with VGG (at least in the current setting of uncalibrated scores; upper bound top-6 accuracy is only 0.827).

- For cls+loc, taking the top 2 detections from SV (accuracy of 0.590) is also slightly better than taking the top detection from each algorithm (upper bound top-2 accuracy is only 0.586)

- For cls+loc, when considering top-5 detections the combination algorithm is in fact stronger than SV

	PASCAL (20 classes)	ILSVRC-sub (200 classes)
Num. instances	1.69	1.98
CPL	8.76%	6.11%
Clutter	5.90	6.15

Table 1. Comparing statistics on the PASCAL VOC 2012 validation set (20 classes) with a subset of ILSVRC 2012 validation set (200 classes; these classes have the highest level of clutter from the full set of 1000 classes).

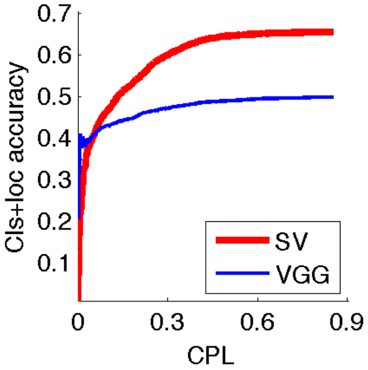


Figure 3. Cumulative cls+loc of SV (red) and VGG (blue) as a function of chance performance of localization (CPL). The height of the curve corresponds to the average accuracy of the object categories with equal or smaller CPL measures. SV outperforms VGG except when considering a subset of 225 object categories with lowest CPL.

alone: SV top-5 accuracy is 0.658, and upper bound top-4 accuracy is 0.671 (top-6 is 0.707)

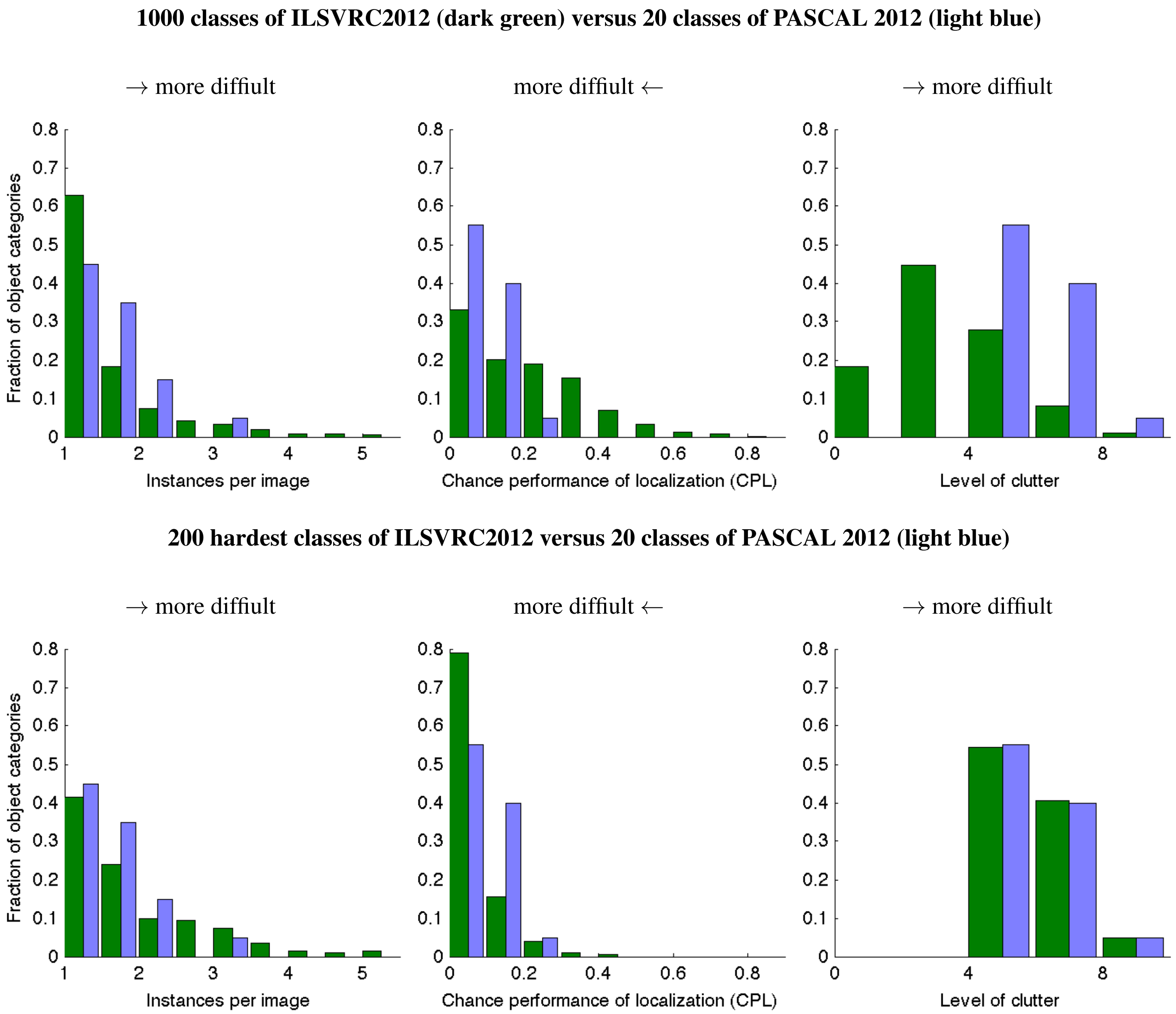
Further investigation is outside the scope of this work but may yield more interesting insights and stronger combined detection algorithms.

2. PASCAL to ILSVRC comparisons

Section 2.2 of the paper defined three measures of localization difficulty (number of object instances per image, chance performance of localization, and level of clutter) and presented some summary statistics comparing PASCAL2012 validation dataset to ILSVRC2012 validation dataset. Figure 2 **(next page)** shows the distribution across these measures on both datasets. Note that taking a subset of 200 classes from ILSVRC which have the highest level of clutter we can obtain a detection dataset which is an order of magnitude larger than PASCAL VOC while being competitive on all three metrics (summarized in Table 1).

3. Effect of object scale on localization on accuracy

Referring to Section 3.2 of the paper we plot chance performance of localization (CPL) versus the accuracy of the algorithms in Figure 3. VGG actually outperforms SV when considering up to 225 object categories with lowest CPL. This implies that the VGG system is currently stronger than SV at localizing small objects.



1000 classes of ILSVRC2012 (dark green) versus 20 classes of PASCAL 2012 (light blue)

200 hardest classes of ILSVRC2012 versus 20 classes of PASCAL 2012 (light blue)

Figure 2. Distribution of various measures of localization difficulty on the ILSVRC2012 (dark green) and PASCAL VOC 2012 (light blue) validation sets. The plots on top contain the full ILSVRC2012 validation set with 1000 classes; the plots on the bottom contain 200 ILSVRC classes with the lowest chance performance of localization. All plots contain all 20 classes of PASCAL VOC.