On the Minimal Recognizable Image Patch

Mark Fonaryov
Technion
Israel, Haifa, 3200003
markf@campus.technion.ac.il

Michael Lindenbaum
Technion
Israel, Haifa
mic@cs.technion.ac.il

Abstract

In contrast to human vision, the common recognition algorithms often fail on partially occluded images. We propose to characterize, empirically, the algorithmic limits by finding the minimal image patch (MRP) that is by itself sufficient to recognize the image. A specialized deep network allows us to find the most informative patches of a given size, and serves as an experimental tool. A human vision study recently characterized related (but different) minimally recognizable configurations (MIRCs) [22]. The drop in accuracy associated with size reduction of these MIRCs was substantially sharp. Interestingly, such sharp reductions were found, for some measures, in our study as well.

1. Introduction

Deep neural networks (DNNs) provide the current state-of-the-art performance in many computer vision tasks, and especially in recognition [11, 7, 8, 9]. In contrast to the human recognition processes, which can rely on small and partial object regions to successfully recognize an object, the performance of neural networks quickly deteriorates when objects are partially occluded or cropped [16, 15].

This raises a natural question: how much information is needed for recognizing an object?

In this work we consider a special practical version of this question: what is the minimal size of a square subimage (patch) that is sufficient for recognition using a convolutional neural network (CNN)? The restriction to a CNN is not severe because currently CNN algorithms are at least as good as any other algorithms.

This is not the only way to formalize the question of sufficient information, and other works considered, for example, the minimal resolution required to recognize an image; see, e.g. [21] .

To achieve the proposed characterization, we design a special neural architecture that identifies the most informative patch and classifies the image relying on the information contained in it. Several variations of this patch based classification (PBC) architecture, corresponding to different patch sizes and different ways of accumulating the local information, are considered.

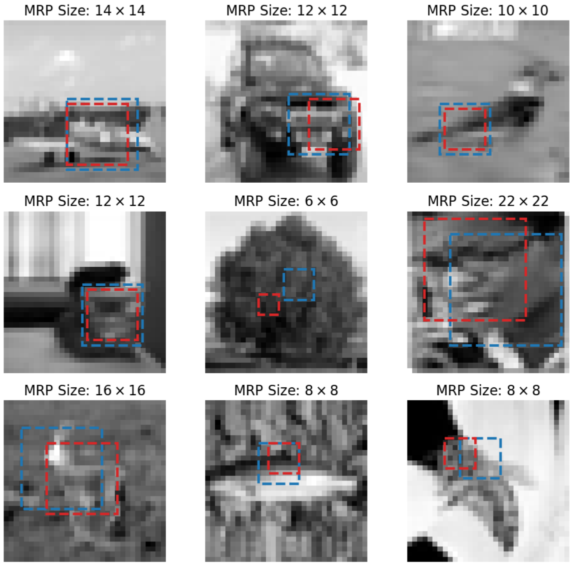

Figure 1. Minimally recognizable patches (blue) and best unrecognizable patches of slightly smaller size (red).

As expected, the minimal recognizable patches we found differ between the categories and within the categories, and increase for higher required accuracy. Interestingly, we also found that for the majority of images, the confidence associated with the patch-based recognition changed sharply with the patch size. This finding is consistent with some related observations on human vision, made recently [22].

Thus, this paper offers the following contributions:

1. The PBC neural architecture that finds the most informative patch and uses it to categorize the entire image.

2. A characterization of the minimal patch size sufficient for categorization.

3. An analysis showing some resemblance between the decisions made by patch based classification and observations on human vision.

2. Related Work

Local regions and features were used by many classical algorithms, and provided improved immunity to pose change, partial occlusion, and in-class variance [13]. Some examples are Visual-Bag-of-Words (BOW) models [4], constellation models [3] and the deformable-parts-models [5]. Likewise, convolutional neural networks (CNNs) effectively combine information available both on a global and local scale [11, 12]. While they represent local features, the recognition accuracy of CNNs decreases when faced with partial images, as happens, e.g. in the presence of occlusion, [15, 16, 23, 14]. Specialized dedicated modifications to the standard recognition architectures [15, 24] improve their partial image accuracy which is however still much lower that the accuracy obtained with full-image data.

This deterioration is expected, and yet, it stands in contrast to human vision, in which object recognition is attained remarkably well, even when seeing only part of the objects.

The human vision system is far from being fully understood. Several general theories have been offered to provide insight into human visual object recognition. The leading approaches to neural object representation can be divided to the viewpoint-invariant approach suggested in [2], and to the viewpoint-dependent approach used in [20].

While the human object recognition process remains debatable, it clearly works well even with partial data. First, low resolution is sufficient [21] and 32×32 color (or 64×64 gray-level) images are recognized well. Recently, a psychophysical study [22] showed that reliable human object recognition is possible even from small image-patches, and identified a special class of minimal image patches. These patches are minimal in the sense that sub-patches, smaller by 20%, or identical patches with 20% lower resolution, were unrecognizable. That is, such patches, denoted minimal recognizable configurations (MIRCs), are locally minimal. Interestingly, this study found that the (human) recognition accuracy associated with the sub-patches was significantly lower than that associated with the MIRC itself. Computer recognition tests, applied to the MIRCs and to their sub-images with computer-vision algorithms, did not find a similar accuracy drop. A model for local image interpretation [1], further dismantled these MIRCs into simple components (e.g. edges), and proposes an explanation of the sharp drops finding. A follow-up on this work have demonstrated that CNN classification of some patches, denoted fragile recognition images (FRIs), may be changed due to small translation or to small resolution reduction [19]. Our work relates to [22] because some of the minimal recognizable patches we find may be regarded as computer specified MIRCs and have similar properties.

3. Evaluating the Minimal Recognizable Patch

Our goal is to evaluate the minimal size sub-image (patch) required for successful categorization. We consider this general question in the context of a specific data set and in the closed set setting [17].

For successful categorization, the patch should be sufficiently informative, such that a categorization procedure accepting only this patch as its input will classify it to the correct category.

We are interested in the most informative patch in the image. Intuitively, the presence of this sub-image is sufficient for categorizing the (full) image successfully if the score associated with this patch and with the correct class is higher that all other scores associated with other classes and/or other patches.

Formally, let S_p^c denote the score of class c associated with the patch p. This score is provided by a classifier, denoted a single patch model, described below. Then the most informative patch suffices for categorization if the inferred class,

$$\hat{c} = arg\ max_c\ max_p\ S_p^c, \tag{1}$$

is correct.

Intuitively, a very small patch or a smooth one cannot be informative because typically, similar patches will be present in images of many categories. However, for categories that are not too similar, we expect to find, in each image, a sub-image of sufficient size and detail, that will be consistent only with its category.

By patch size we mean, approximately, the size as a fraction of the full object size, as seen in the image. For our study, we shall use a data set containing images of fixed size (32×32). We shall also assume that each image contains one object, which is tightly bounded by the image boundaries. This assumption approximately holds for most images in a variant of the CIFAR data sets, that we use. Under this assumption the patch size may be specified by its size in pixels.

3.1. The Patch Classification Model

The main tool developed for the study of minimal recognizable patch is the Patch Based Classification (PBC) model, which performs image-classification based on information included in the best, or most informative single patch of the full image.

The best patch is unknown and is not pre-specified, therefore locating it is part of the network's task. That is, the classification task is weakly supervised.

The general architecture of the network is composed of several parts: a. Splitting the input image into (overlapping) spatial patches and resizing each one to a standardized size. b. Independently analyzing each patch using a CNN, denoted the single-patch-network. For each patch, the single-

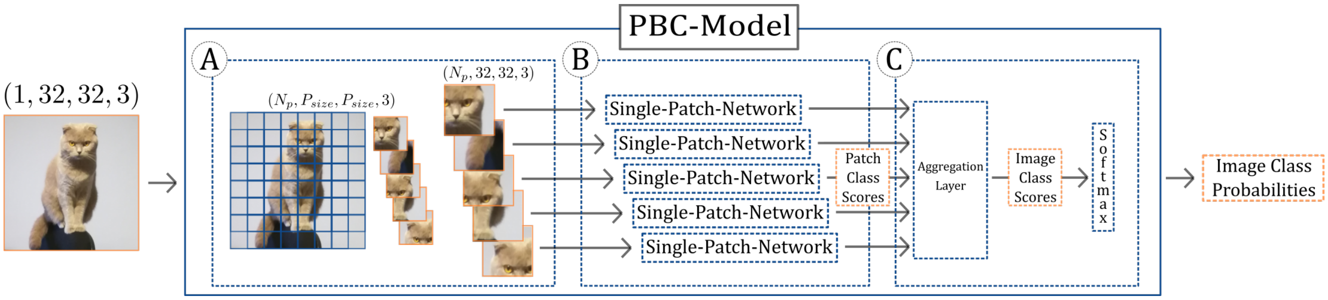

Figure 2. Patch-based-classification architecture - (A) The input image is split into overlapping, spatial patches, which are resized to a standardized size. (B) Each patch passes thru the single-patch-network. (C) The aggregation layer transforms patch-class-scores into image-class-scores and a softmax layer converts them into estimated image-level class-probabilities.

patch-network provides C patch level scores, one for each category. c. An aggregation layer converting the patch-level scores of all patches into C image-level scores. d. A softmax layer getting the image level scores and providing C image-level class-probabilities; See fig. 2 for a visualization of the model. We elaborate on these network parts below.

3.2. Single Patch Network

The patch-network could be any straightforward classification network. The network used in our experiments follows the All-Convolutional-Net model proposed in [18]. This model was slightly modified by replacing the dropout regularization layers with batch-normalization and the 6×6 global-averaging layer with a more generalized 6×6 convolutional layer. The softmax layer was moved out of the single-patch-network, to be placed after the aggregation stage; see section 3.4. A detailed summary table of the architecture, can be found in the supplementary material.

We chose to use a uniform size (32×32), interpolated, patch, as an input to the single-patch-network. The classifier is still learned for every patch size independently, and yet, this choice enables us to work with a uniform architecture. We experimented with other interpolated input sizes and found, as expected, that smaller interpolated patches work somewhat better with smaller original patch sizes, for which the interpolation is less extreme. However, the differences in the results (less then 5%) were not significant for this study.

3.3. Patch Score Aggregation

The manner by which the patch-level scores are compared and aggregated into image-level scores influences the choice of the best patch and its associated confidence. We considered two types of max score aggregation:

Location independent max - This maximum score, denoted by $S^c_{max-ind}$, is evaluated over all patches, separately for each class. For this aggregation, the score for each class is commonly taken from a different patch.

Winner directed max - This maximum score is denoted by $S^c_{max-dir}$. Here, the score of all classes is taken from a single patch, the one associated with the overall maximum score.

Note that both aggregation methods determine the winner according to the best overall score, as specified in eq. (1). The first uses other patches for evaluating the scores associated with other classes, and hence the confidence. The second aggregation takes the other classes' scores from the same patch, ignoring possibly higher scores from other patches. Formally,

$$S^c_{max-ind} = \max_p \{S^c_p\}, \tag{2}$$

$$S^c_{max-dir} = S^c_{p*}\ , where\ p* = argmax_p max_c S^c_p \tag{3}$$

It seems that an intelligent agent, wishing to categorize the object(s) in a scene, will scan it and will try to extract the best evidence for each category, no matter where. In this context, calculating confidence using the first aggregation method is justified. On the other hand, in experimental conditions, when only one patch is observed, the second aggregation method describes the available information better. Moreover, in scenes containing several objects, the second aggregation methods allows the detection of multiple categories. For such scenes, it also helps the learning process because the presence of an object from one category on one place does not indicates that responses to other categories in other locations should be suppressed. Empirically, the two aggregation methods give similar results with some advantage to the first.

3.4. Placing the Softmax Layer

We placed the softmax score normalization at the final layer, acting on the image-level scores provided by the aggregation layer. The alternative option, of applying softmax normalization to every patch, before choosing the maximal one, would let the response to other classes influence the maximum. A substantial, but not maximal response to some class, for example, would lower the normalized response to the winning class, which could otherwise be larger than the response to this class in all other patches. We also found experimentally that a softmax layer at the patch-level hurts the model's generalization ability.

4. Experiments and Results

4.1. The CIFAR10* Dataset

We started our experiments with the CIFAR10 dataset [10]. This dataset contains 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. With the exception of the ”birds” and ”frogs” classes, this dataset can be divided into pairs of related and similar categories: ship-plane, car-truck, dog-cat and horse-deer. We observed that for small patches, the learned model often preferred one of two similar categories, and “gave-up” on the second one. It seems that at small patch sizes, informative areas of related categories (e.g. the wheels in automobile and trucks) were indistinguishable for the classifier to be separated effectively. To achieve better mean performance over the dataset, it chose the class with more, or clearer, appearances of this informative area. This phenomenon interferes with finding minimal recognizable patches for the non-preferred categories. Therefore, we experimented with a CIFAR10 variant which is easier in the sense of containing less inner-similarities. This variant, denoted CIFAR10*, was based on classes from the CIFAR10 and CIFAR100, and consisted of 3000 samples from each of the following classes: airplane, automobile, bird, cat, deer, frog, fish, tree, person and insect. The data set was divided into a training set of 25,000 images and a test set of 5,000 images, both well-balanced between the 10 classes.

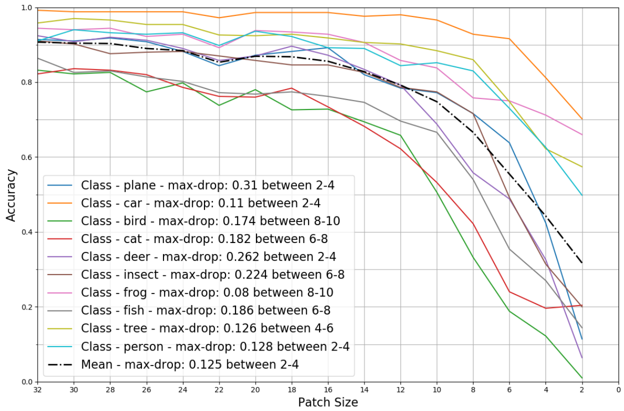

Figure 3. Mean and class specific accuracy with color images.

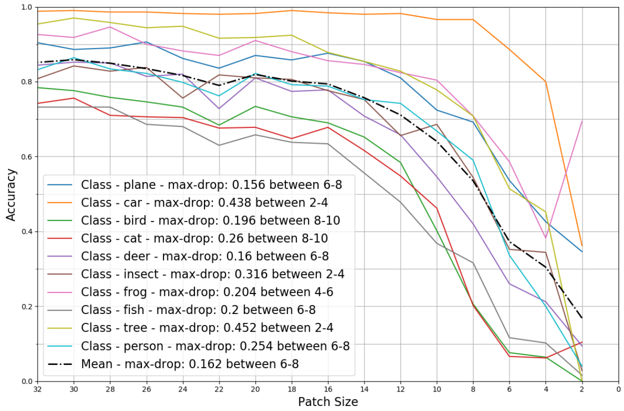

Figure 4. Mean and class specific accuracy with gray level images.

4.2. Implementation

For training we used a categorical cross-entropy loss function with an Adam optimizer. Training was conducted with a batch size of 50 images, for 150 epochs, with an initial learning rate of 0.001, reduced by a factor of 2 every 30 epochs. The weights were regulated with ridge regression, with an 1e-3 coefficient. All input images are normalized with the mean and variance of the training set. The network was implemented with a Keras-TensorFlow neural-network library, using a Geforce Titan X GPU.

During training, the spatial stride taken while splitting each image was set to be half the patch size, except the smallest, 2 × 2 patch, for which a 2 × 2 stride was used due to GPU memory limitations. During evaluation, the stride was always set to be a single pixel.

We trained the single patch model with the same hyper-parameters for all patch sizes. We checked experimentally that optimizing the hyper-parameters depending on the patch size had only a small effect.

4.3. Categorizing Color Images

We trained the 16 patch-based model with the first aggregation method (location independent max) for 16 square patch sizes, $d_i \times d_i$ pixels, where $d_1 = 32, d_2 = 30, \dots, d_{15} = 4, d_{16} = 2$. We refer to the models simply as ”model of size d”. The model of size 32 corresponds to the full image.

We then applied the learned models to classifying the test sets of CIFAR10*, and estimated the accuracy as the fraction of images classified to the correct category. As expected, smaller patches are associated with reduced accuracy. Remarkably all categories are classified correctly with 50% accuracy with 10 × 10 patches, corresponding to roughly 0.1 of the image area.

4.4. The Basic Gray Level Experiment

Interestingly, some categories may be identified from very small patch sizes, which may correspond to either distinct small features (e.g. a wheel or an eye) or to texture (tree or frog), but is probably due mostly to color. Color can be very discriminative, especially for the close-set context. A single blue pixel, for example, can hint to one of 3 classes: fish, birds or airplanes. See [21] for a study revealing the advantage of color in low resolution images.

The dependence of color on size is weak, and here we are more interested in the size dependent information. Therefore, we converted our data sets into grayscale and trained a single-channel-input version of the patch based model with the same aggregation method and 16 patch sizes. The results, shown in fig. 4, demonstrate that the accuracy associated with small patches significantly decreases.

4.5. The Aggregation Effect

We trained a single channel input version of the patch based model with the second, winner-directed, aggregation method, and compare the accuracy of the resulting classifier to that learned with location independent aggregation; see fig. 5. The differences are small. First, note that the difference is due only to possibly different training, because once the classifier is given, the class assigned to the images is determined according to eq. 1 for both aggregations methods. The small difference is remarkable because training with location independent aggregation uses the best patches for each incorrect class for suppressing the score to this class. Using these patches and not the particular winner directed patch is much more informative and leads to better SGD steps and faster convergence. Yet, we see that even with the less informative winner directed patches, learning is almost as good. Note, however, that with smaller patches the overlap between these best patches and the winner patch is smaller, the difference is potentially larger, and the disadvantage of learning with winner dependent aggregation is more significant. The difference in accuracy is correspondingly larger, but it is still small.

4.6. Is There a Sharp Drop Effect?

Following the observations on human vision [22], which address a different but related problem, it would be interesting to see if there is a sharp drop effect here. A sharp effect exists if some measure of recognition is changed significantly between two consecutive patch sizes.

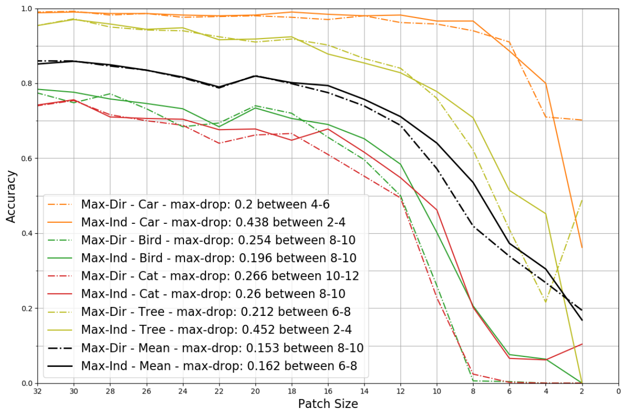

Figure 5. Difference in accuracy between the two aggregation methods.

The results of our investigation are mixed. For some recognition measures there is no sharp effect, but for one there is.

For the first experiment (color images), we measured individual accuracy drops for different classes and for each patch size step. The maximal accuracy drop, calculated independently for each class, is given in the legend of fig. 3. Clearly, for all classes, the maximal accuracy drop is moderate and the accuracy curve is rather smooth. For some patch size steps the accuracy drop is higher than for other but the difference is not large and for no patch size step is the accuracy drop larger than half of the accuracy range (i.e. than 0.5). The maximal value is 0.26.

The color information contributes to the accuracy curve smoothness, because color information, as available from, say, a color histogram, depends only weakly on the patch size. Indeed, the accuracy drops observed for gray level image for larger, and the maximal accuracy drop went up to 0.45; see the legend in fig. 4.

While the compared patch pairs and the classifier range of responses, are different here and in [22], it is clear that the accuracy drop found here is significantly lower than the average value of 0.71 ± 0.05 reported in [22].

4.7. Single-Image Experiments

A major difference between our experiments and the psychophysical experiment described in [22] is that in [22] the ability to classify an image from a patch was evaluated independently for each particular patch, using the responses available from an ensemble of subjects. Each subimage was observed by many participants and the recognition accuracy was estimated as fraction of the participants that classified it correctly. In our experiments, and essentially in most image classification processes, the situation is the opposite. We usually have only one algorithmic observer but many images of the same object or category. The recognition accuracy is estimated as the fraction of images (in a category) that are correctly classified from the most informative fixed size patch.

Images of objects from within a given category differ a lot due to the intra-class variability and the uncontrolled object pose. In particular, the minimal size of a sufficiently informative patch differs as well. Thus, for a particular patch size, some images in the ensemble contain a patch that is sufficient for successful recognition, but other images do not. As patch size increases, the fraction of images that contain a sufficiently informative patch also increases, sometimes slowly. Hence the smooth accuracy curve and the small accuracy drops.

We therefore proceed to experiments on single images,

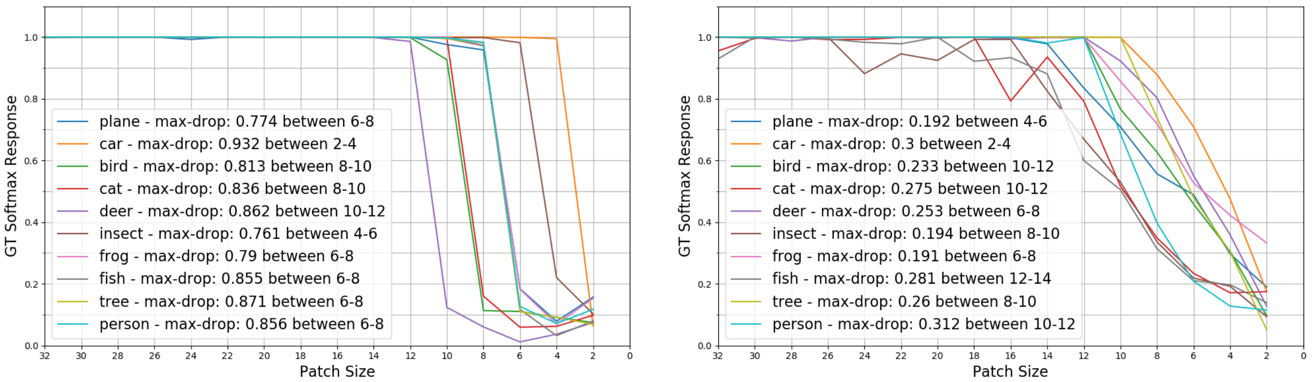

Figure 6. Confidence (Softmax-response of the ground truth class) in single-images. (Left): Images with largest max-drops per class. (Right): Images with smallest max-drops per class.

and would like to estimate the recognition accuracy as a function of patch size. For a particular image, the accuracy cannot be estimated empirically and is replaced by a single image accuracy estimate, or confidence. In networks trained with cross entropy loss, the softmax response approximates the posterior probability for this category, and may be used as a simple, and yet reasonably accurate, confidence [6].

The confidence curves for specific images reveal sharp, significant, accuracy drops in many images. The curves associated with the maximal and the minimal drop in each category are given in fig. 6. A small part of the images were associated with a smooth uniform confidence drop starting at some patch size; see fig. 6(right). Some other images were difficult to classify even as a full-image, and were associated with low, smooth, confidence curves.

For most images, however, the maximal confidence drop is substantial, as revealed the histogram of the maximal drop; see fig. 7. Clearly, for the majority of images, a maximal drop larger than 0.5 was found.

As expected, the critical patch sizes associated with the maximal confidence drop are not fixed. (Otherwise the evaluation of accuracy over an ensemble of images would be characterized by a sharp drop as well.) To show this variation, we plotted a 2D histograms of the maximal drop size and the patch size change; see fig. 8. Clearly, the size of this critical patch varies significantly over the set of images associated with each category, and for some categories there are even several dominant sizes. The drop size varied as well and was typically larger when it occurred with larger patches. These variations fully explains the lack of significant accuracy drops, when the accuracy is evaluated as an average over a data set.

This behavior was reproduced for both grayscale and color images. There was a small difference in the maximal accuracy drop. The average (over all images) of this drop was 0.624 for grayscale images and 0.608 for color. This difference was smaller than expected, considering the variances observed in sec. 4.4. There was, however, a clear gap in the patch sizes associated with the maximal accuracy drop: the gray level images required larger sizes for recognition.

The second aggregation method revealed even more substantial drops, with average maximal confidence drop being 0.72. The number of images, with a maximal drop larger than 0.5 somewhat increased (3683 vs. 3492).

5. The Minimal Recognizable Patch

As described above, for a majority of the images, there is a sharp confidence drop which, for most images, is larger than 0.5. This implies that there is a patch of size $d* = d_i$, for which the confidence is higher than 0.5 and is substantially larger than the confidence of the best smaller patch, of size d_{i+1}. We call to this critical patch the minimal recognizable patch (MRP).

Using our methods, that specify only one patch as the best one for every given size, there is only one MRP in every image. However, other images patches, associated with the

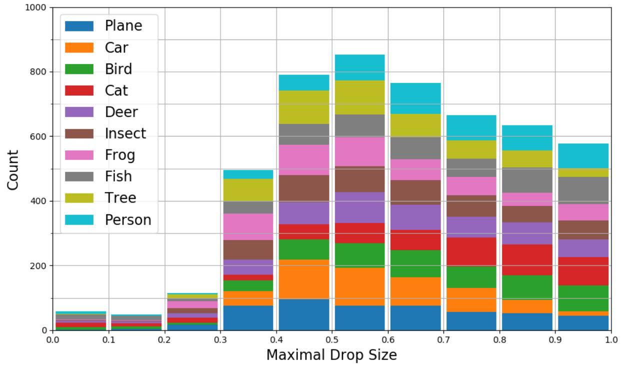

Figure 7. Maximal-drop size of softmax-response in gray level images. A majority of images (4328 out of 5000) displayed maximal drops consistent with the MIRC requirement in [22].

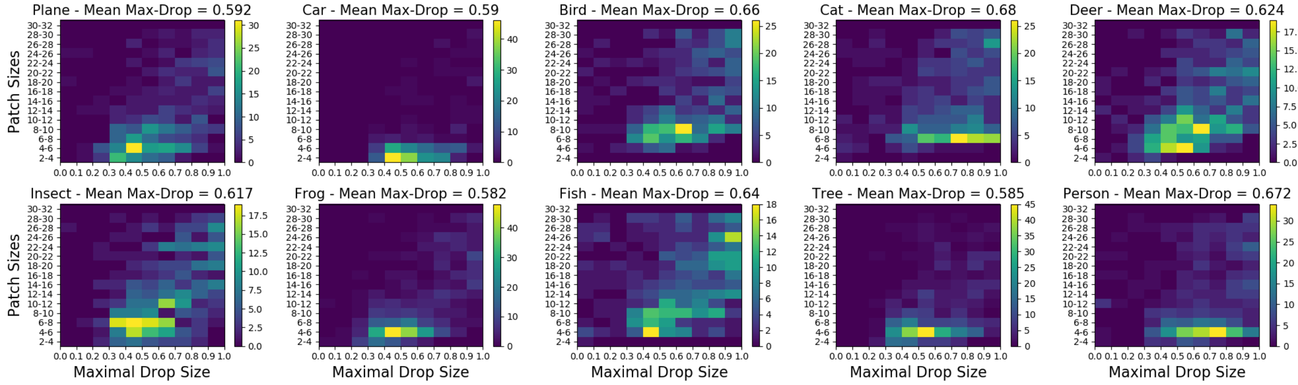

Figure 8. 2D histograms of maximal-drops in softmax-response. Each histogram shows drop-size (x-axis) and between which patch sizes it occurred (y-axis) for the 500 test images of a specific category.

same size and with somewhat lower scores than the best one are often present in the image. These patches are associated with similarly large confidence drop, and could be regarded as MRPs as well.

This term MRP is somewhat misleading because, in principle, the classifier may give the correct classification even if the confidence is lower than 0.5. Yet, due to the large typical drop, this is rarely the case.

Some MRPs are shown in fig. 9. Some of them are consistent with human judgment which can identify the category from the MRP but not from the smaller patch. For some MRPs however the consistency is weaker and humans either cannot classify the image from the MRPs, or can do it even from the smaller images.

6. Conclusions

This work empirically characterizes the minimal sub-image required to categorize an image successfully. A specialized deep network was designed for this task, and was used to find the most informative sub-image in each image. We show that the size of this minimal sub-image takes, on average, is a small fraction of its full area, but also that it varies significantly within each category.

The single image experiments (sec 4.7) seem related to the human vision study described in [22]. Both share a common finding: there are image regions, that are sufficiently informative for recognition, but stop to provide the required information due to a small size reduction. Moreover, the reduction in region informativeness is sharp and substantial. Remarkably, in both studies, this sharp reduction was not part of the demands but was found, empirically, as a byproduct.

Interestingly, earlier work did not succeed to computationally reproduce the perceptual sharp reduction effect [22]. See, however, [19]. We intend to further study this difference. One possible advantage of our algorithm is that finding the (unknown) most informative patch of a given size, is a weakly supervised auxiliary task, which uses all image patches for training, and supports the classifier.

References

[1] Guy Ben-Yosef, Liav Assif, and Shimon Ullman. Full interpretation of minimal images. *Cognition*, 171:65–84, 2018.

[2] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115, 1987.

[3] Michael C. Burl, Markus Weber, and Pietro Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *European Conference on Computer Vision (ECCV)*, pages 628–641, 1998.

[4] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *European Conference on Computer Vision (ECCV)*, pages 1–22, 2004.

[5] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[6] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations (ICLR)*, 2019.

[7] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[10] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

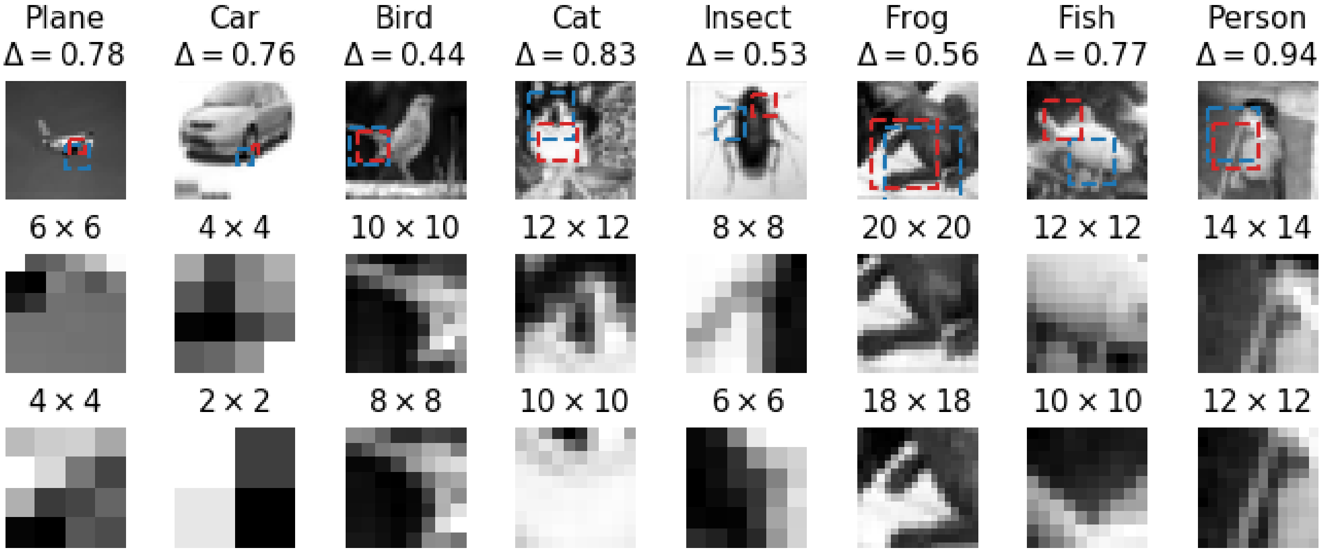

Figure 9. MRP examples. (Top): Original Images with their MRP (blue) and best patches of the next smaller size (red). Associated confidence drop, between the MRP and the best patch of smaller size is displayed above the image; (Middle): The MRP image. (Bottom): the best unrecognizable patch.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, 2012.

[12] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. In *International Conference on Learning Representations (ICLR)*, 2019.

[13] David G. Lowe. Object recognition from local scale-invariant features. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157, 1999.

[14] Michael Opitz, Georg Waltner, Georg Poier, Horst Possegger, and Horst Bischof. Grid loss: Detecting occluded faces. In *European Conference on Computer Vision (ECCV)*, pages 386–402, 2016.

[15] Elad Osherov and Michael Lindenbaum. Increasing CNN robustness to occlusions by reducing filter support. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 550–561, 2017.

[16] Bojan Pepik, Rodrigo Benenson, Tobias Ritschel, and Bernt Schiele. What is holding back convnets for detection? In *German Conference on Pattern Recognition GCPR*, pages 517–528, 2015.

[17] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[18] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR)*, 2015.

[19] Sanjana Srivastava, Guy Ben-Yosef, and Xavier Boix. Minimal images in deep neural networks: Fragile object recognition in natural images. In *International Conference on Learning Representations (ICLR)*, 2019.

[20] Michael J Tarr and Heinrich H Bülthoff. Is human object recognition better described by geon structural descriptions or by multiple views? comment on biederman and gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, 1995.

[21] Antonio Torralba, Robert Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[22] Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 113(10):2744–2749, 2016.

[23] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3039–3048, 2017.

[24] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In *European Conference on Computer Vision (ECCV)*, pages 657–674, 2018.