Use visual concepts to handle occluded object

Abstract

In this paper, we gives a new understanding of the internal mechanism of DCNN and helps DCNN to acquire human's ability of recognizing objects by part. Motivated by the fact that human can easily recognize an object by part, we want to train a DCNN to have the similar ability. We firstly use occlusion paradigm to study the useful object parts in recognition tasks. An unsupervised clustering technique is applied to the hidden layer of VGG to extract part representations known as Visual Concepts. Then the informativeness of each visual concept is judged by the drop in object recognition performance when it is air-brushed away from the image. We also try to explain why some visual concept is more important in recognition both visually and theoretically. After that, we fine-tune VGG with attention windows which only show the receptive fields of useful visual concepts, helping the networks focus on the key information for recognition task and ignore the redundant parts. By testing original VGG and our fine-tuned version on well-designed datasets which include different kinds of apertures, we find that unlike humans, DCNN is weak to combine and make use of discrete local parts in recognition tasks, and single big-enough part is more welcomed than several discrete local semantic parts in recognition task. While our fine-tuned VGG greatly enhances the ability to recognize an object by part, which makes it more 'human'. Our work sheds light on the understanding of how DCNN works, makes DCNN more like a human and shows a possible solution to increase the recognition accuracy under different kinds of occlusion or distortion.

1 Introduction

Recent years, the developing of Deep Convolutional Neural Networks revolutionized Computer vision field. DCNNs can handle a lot vision tasks such as classification [6], detection[4] and segmentation[1] much better than the traditional techniques which is based on hand-crafted features. However, there are still many things that DCNNs cannot work very well. For example, It's hard for a DCNN trained by normal images to recognize objects under occlusion. In order to handle this problem, people use domain-specific methods in certain area like pedestrian detection[20] and face recognition [11], or use part-based methods like [2], both of which cost much time and need prior knowledge with supervision and poor transfer ability. While in the real world, human can easily make use of partial information and recognize an object under occlusion: A kid can recognize the panda hidden in the bamboos, and we can recognize our pet cats with only their tails or paws easily.

Human can recognize an object from partial information, which enables human to easily recognize an occluded object. However,[16] shows current DCNN models do not learn to recognize object part at a human level and [3] shows that some traditional part-based method works similar to humans in difficult tasks, corroborating the theory of part-based object representation in the brain. It seems that human use an object's semantic part to recognize the whole object but current DCNN models don't have this ability and they have another kind of mechanism.

Motivated by this, we want to train a DCNN which have human's ability to recognize an object just from partial semantic part. Our approach is composed of 2 parts, firstly, we need to find useful semantic parts for recognizing certain category of objects without supervision and prior human

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

knowledge. Secondly, we need to make DCNN remember these parts, so next time DCNN can recognize an object even it is occluded and just showing partial critical semantic parts.

In the first step, we propose a simple method based on occlusion paradigm to find which parts are necessary and sufficient for object recognition. We use DCNN itself, which are shown to be able to perform both object recognition and semantic part localization in a single forward-pass without ever having been explicitly taught the notion of semantic parts [21], to detect semantic parts. An unsupervised clustering technique is applied to the hidden layer of VGG [13] to extract part representations known as 'Visual Concepts' [1]. Then we apply occlusion paradigm on the hidden layer of each visual concept to detect which semantic part is more important in recognition task. The idea is very simple, once we occlude a visual concept, if it is important for recognizing certain object, a great drop in recognition probability will be observed. After finding these useful semantic parts for each category, we also investigate why they are important.

In the second step, we propose a new data augmentation technology. We fine-tune VGG with aperture images which only show the receptive fields of useful visual concepts, helping the networks focus on the key information for recognition task and ignore the redundant parts. After that, we well design some datasets which show different apertures as test set. Our fine-tuned network shows better performance on recognizing images with partial semantic parts, it can recognize an object from just one or several apertures opened.

Besides the better performance in recognizing partially occluded images, we also get some scientific findings about inner mechanism of DCNN during the testing process. In a series of experiments, original VGG was shown to already have the ability to combine and make use of ambiguous and discrete information in recognition tasks, which is also greatly enhanced in our fine-tuned network. We also demonstrate that although DCNN has the ability to make use of additional information to guarantee the accuracy of recognition, a confirmed semantic part is more welcomed in recognition task.

At last, we apply our network on a serious of different occluded object recognition tasks to show its robustness. Network fine tuned with useful visual concepts has better performance when the occluded area is big, and doesn't compromise when the occlusion is small.

Our work shows which parts are important in object recognition and why these parts are important. We also get some interesting scientific findings about the inner mechanism of DCNN. Besides, in application, we find Visual concepts a useful technique in data augmentation, and it's promising to apply visual concepts in occluded object recognition.

2 Related work

Besides clustering method, there are some other ways to get the semantic parts of certain category of object. [14] used hand-made hog features to study semantic parts, which cost too much time. [8] used CNN just as a feature extractor without exploring how mid-level visual concepts are captured. [12] assumed that a single filter from a hidden layer can detect parts, which is quite controversial among other scientists.

Visual concept is not a newly proposed term. In [9] , the authors calculate the distances between hidden layer feature vectors and the centers of their corresponding visual concepts to generate some regularization terms which are added to final objective function. In [18], visual concepts are regarded as semantic part detectors and the authors do a lot works to show the superiority of visual concepts in representing semantic parts.

There is also similar work [15] to us that also investigate the relative importance of visual concept in the recognition of certain category of objects. This paper introduces the 'concept activation vector' to represent each visual concept in hidden layer. Then the authors add or minus this 'concept activation vector' in hidden layer responses to see what happens to final recognition result. However, their 'visual concept' doesn't have the same definition as ours, they are more abstract things like colors or textures.

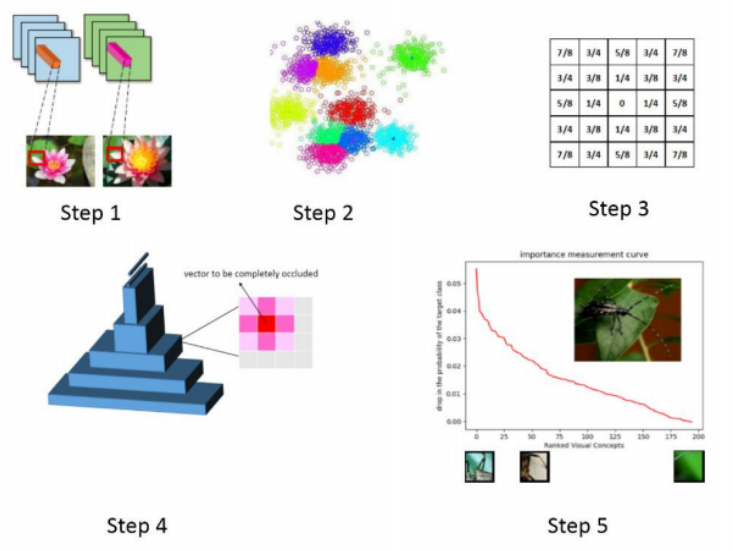


Figure 1: Steps to find useful semantic part in recognition task. **Step 1:** Apply K-means++ to cluster the features at pool3 in pre-trained and fine-tuned VGG network, yielding a dictionary of VCs (Visual Concepts). **Step 2**: Use greedy merging method to prune the VCs. **Step 3:** Apply VGG to an image. Attenuate pool3 unit response corresponding to a visual concept within a Gaussian window. **Step 4:** a) Calculate relative drop in probability of the target class with and without the presence of certain VC. Large relative drop means the VC is more important. b) Repeat with pair, triplet and quadruplet VCs and compare the difference between them. **Step 5:** Rank order the VCs or sets of VCs according to its importance. Try to answer the following question: for each class, what make some VCs more important than the others?

3 Learning useful semantic part for recognition task

This section describes our method to find useful semantic parts in the recognition task of certain class of objects [Figure 1]. We use method mentioned in [18] to find semantic parts in certain category of objects, and then we judge the usefulness of each semantic part in recognition task by air-brushing its hidden layer feature and investigating the drop in object recognition performance. Furthermore, we try to find out the factors that make some parts more important than the others in recognition task.

To make it easier, we fine tune the pre-trained VGG to classify 100 classes instead of 1000. We randomly sampled 100 classes from ImageNet, and extracted 34K bounding box images distributed equally from these class as training images. In order to eliminate the interference of background information, we cropped all images so that only pixels in the bounding boxes were left. The fine-tuning process refers to [17]. We fine tune the last 2 fully connected layer and the top-1 accuracy on 11K test images increases from 50% to 87%. We call this fine-tuned 100-classes classification network Baseline Network in the later part of this paper.

3.1 Extract visual concept from pool3 layer

We refer to the method mentioned in [18], where an effective unsupervised clustering technique is applied to cluster the feature responses at the hidden layer of VGG, yielding a dictionary of part representations known as Visual Concept, which is introduced to find the parts of a certain category of objects without semantic part annotations and prior knowledge of that category. Visual Concepts are shown to be semantically and visually coherent in that they represent semantic parts and the corresponding image patches who serve as the receptive fields look similar. Also, Visual Concepts provide full spatial coverage of the parts of an object. They can be regarded as auto semantic part detectors.

More precisely, in our work, for each category of objects, we add some white paddings to bounding box images to make them square, and then resize them and take them as the input of Baseline

Network. Then, in the hidden layer L (here we choose pool3 because the receptive field of pool 3 is 44*44, a suitable size for the whole 224*224 input) with spatial resolution $W_l \times H_l$ and N_l filters, we extract the population responses $P \in R^N$ from all big-enough bounding box images (the original longer side of the bounding box must be bigger than 100 pixels in order to guarantee the resolution of image patches corresponding to certain visual concept) which come from the same category in the training set. We cluster these population responses (ignoring the spatial locations of the features) using K-means++ method. These clusters are known as Visual Concepts.

The number of clusters is initially set to be N_l. Then a greedy cluster merging algorithm is applied to reduce the number of clusters. After that, we apply a threshold with some magic numbers to merge clusters whose centers are too close to each other and abandon clusters whose members are all from one or two rare cases. Here we show a number of the best image patches for one of the 189 visual concepts (i.e. the best 20 whose feature vectors are closest to the visual concept center) of white stork in [Figure 2].

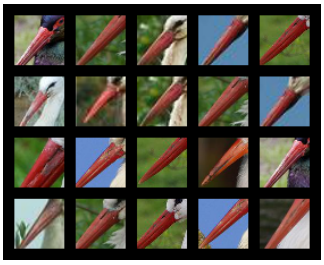
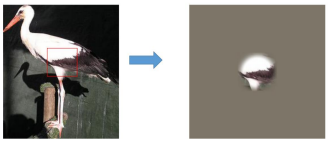


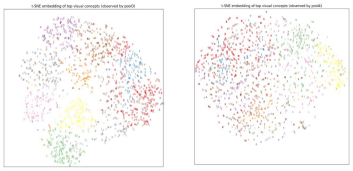
Figure 2: The visualization of original image patches corresponding to pool3 hidden layer features in one Visual Concept of white stork

3.2 Further decrease redundancy of visual concepts by utilizing higher layer's view

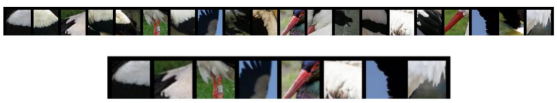
We can find that there are still some similar visual concepts even after using greedy merging algorithm. In order to further decrease the redundancy of extracted visual concepts, additionally, I propose a new visual concept merging algorithm which utilizes higher layer's view to merge lower layer's visual concepts. Firstly, for feature vector A from certain pool3 visual concept, move its corresponding receptive field to the center of the image and leave the remaining part grey [Figure 3a]. In order to eliminate the abrupt change of pixel values at the edge of receptive field, I use a "bubblize" method to soften the edge and transform the receptive field from square to circle (This trick is also used in later part and I will introduce it more clearly in later section). Then, I extract pool4 feature vectors of this 'receptive field centered' image, and call the corresponding pool4 feature vector at the center of that image as pool3 feature vector A's 'higher view feature'. The feature vectors in higher layer



(a) generate bubble aperture image



(b) t-SNE result of 10 pool3 visual concepts and their higher view vectors



(c) white stock's pool3 visual concepts before and after using this trick

Figure 3: Use higher layer's view to decrease redundancy

are closer to each other, and at the top layer, there's only one 100-dimensions vector that tells the category. From this observation, we can deduce that the 'higher view feature' of feature vectors from similar visual concepts are much closer to each other, and those from unsimilar visual concepts will be more separated. So naturally, we can calculate different pool3 feature vectors' 'higher view vector' and use greedy merging algorithm and magic number threshold again to decrease the redundancy of visual concepts. This method comes from higher layer's stronger classification ability and efficiently merges redundancy visual concepts. In [Figure 3b], 10 white stork's visual concept clusters and their corresponding 'higher view vector' are shown by t-SNE visualization method. [Figure 3c] shows 20 visual concepts of white stork and the result after using 'higher view vector' to merge them. We can find that similar visual concepts like the wings visual concepts are merged together.

3.3 The differential contributions of different semantic parts in certain class of objects

Each feature vector in certain visual concept corresponds to a patch from original images as its receptive field just as [Figure 2] shows. So we can use visual concept as semantic part detector. The informativeness of certain visual concept is accordant with the usefulness of its corresponding semantic part. In the later parts of this paper, we may use visual concept to represent its corresponding semantic part.

Here we propose a method to find the informativeness of different visual concepts. The informativeness of visual concepts of certain category of objects is judged by the drop in object recognition performance when certain visual concept is air-brushed away from the image. Large relative drop means the visual concept is very important for recognition task. The air-brush process is to attenuate hidden layer (here we choose pool3) unit response corresponding to a visual concept within a Gaussian window. In other word, in inference process, we set the pool3 feature vector which belongs to certain visual concept to 0 and decrease the feature vectors surrounding it according to a Gaussian template. It seems like the feature vector of certain visual concept is 'occluded' in hidden layer.

The whole process for measuring the importance of visual concept V_m^k of certain category m is defined as:

$$S(V_m^k) = \frac{\sum_{I_k^i \in M_k, P_{original}(I_k^i)>0.3}(P_{original}(I_k^i) - P_{occluded}(I_k^i))}{|M|} \tag{1}$$

Where $S(V_m^k)$ is a score to represent the importance of visual concept V_m^k in the recognition of m. M is a set of images which are used to extract visual concepts of m. $M_k \subset M$ are the images which have patches as the receptive field of feature vectors in visual concept V_m^k . $P_{original}(I_k^i)$ is image I_k^i 's recognition probability of m in Baseline Network, and $P_{occluded}(I_k^i)$ is image I_k^i 's recognition probability after air-brushing visual concept V_m^k. Here,

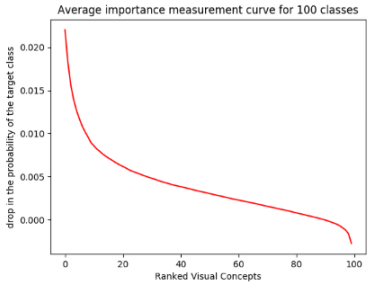
$$P_{original}(I_k^i) = F(f_l)$$

$$P_{occluded}(I_k^i) = \frac{\sum_{P_{x,y} \in P_{k,l}^i} F(f_{l,x,y}^{air-brushed})}{|P_{k,l}^i|}$$

F is the forward propagation process. $f_{l,x,y}^{air-brushed} \in R^{W_l \times H_l \times N_l}$ is generated by air-brushing a feature vector $p_{x,y} \in R^{N_l}$with spatial grid (x, y) from $I_{k,i}$'s intermediate layer response $f_l \in R^{W_l \times H_l \times N_l}$,where $W_l \times H_l$ is the spatial resolution of the intermediate layer L and N_l is the number of filters of L. $P_{k,l}^i$ is a set of feature vectors that belong to certain visual concept V_m^k as well as act as feature population responses in $I_{k,i}$'s intermediate layer response f_l. For the air-brushing process , $f_{l,x,y}^{air-brushed}$ and f_l have the same entries except that $f_{l,x,y}^{air-brushed}[x-2:x+2, y-2:y+2,] = f_l[x-2:x+2, y-2:y+2,] \circ G(0,1)$. $G(0,1)$ is a 55 Gaussian template whose center is 0.

By using this method, we can get the most useful (the largest drop when air-brushed) as well as the most meaningless visual concepts of certain category of objects. We also draw an importance measurement curve to rank order visual concepts of certain category of objects according to their importance in recognition task. We do this for all 100 classes of objects. [Figure 4a] shows the most and least useful 10 visual concepts of longicorn and the importance measurement curve, [Figure 4b] shows the average importance measurement curve of all 100 classes of objects.

(a)



(b)

Figure 4: Importance measurement curve and some ranked visual concepts.

Visually, the most useful visual concepts are most characteristic and recognizable semantic parts lying in the body of the target object, while the most meaningless visual concepts are background patches or patches that can make some confusion, which also explains why the recognition probability even increases when these visual concepts are air-brushed away (see the importance measurement curve). In the next section, we will theoretically study the factors that make some visual concepts more important than the others.

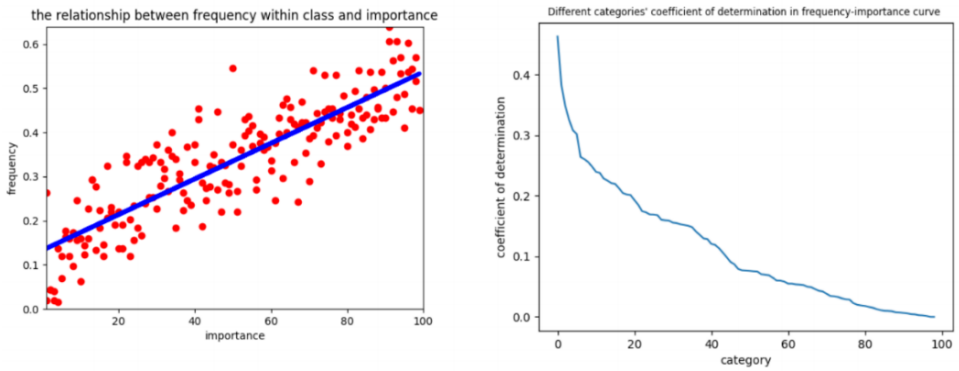
By using our method, the learning of useful parts for recognition can be end-to-end and doesn't need any prior knowledge as well as well-designed training data or semantic annotation. Our work shows the DCNN can detect most useful part in recognition automatically without being explicitly taught the notion of object's parts, which also demonstrates the conclusion in [3]. Our work also implies a method to increase the object recognition accuracy——just air-brush some meaningless visual concepts in recognition process.

We also do similar things for pair visual concepts, triplet visual concepts and quadruplet visual concepts and choose the most useful pair, triplet and quadruplet visual concepts.

3.4 Why some visual concept are more important in recognition task?

This section we will discuss the factors that make certain semantic parts critical for recognition task. Naturally, people will think those unique and typical parts as the key to recognize certain category of objects, but how to define 'unique' and 'typical'? We have 2 approaches to solve this problem: image instance retrieval and categorical retrieval. In the image instance retrieval, we calculate the occurrence frequency of certain semantic part among images from certain category, where 'typical' semantic part should have higher occurrence frequency. In the categorical retrieval, we calculate the occurrence frequency of certain semantic part among 100 classes of objects, where 'unique' semantic part should have lower occurrence frequency. We then compare one semantic part's 'typicalness' and 'uniqueness' (occurrence frequency) with its importance score calculated in section 3.3.

[Figure 5a] shows the examples of the result of image instance retrieval, which indicates that there is a positive correlation between one semantic part's importance score and its occurrence frequency among different objects from the same categories. However, some people may argue that our method to calculate the importance score tends to give those occurring many times in target category a higher score since the denominator is fixed to be the total number of images, and more occurring times means more pluses in the numerator. So we change the denominator of formula (1) from M_k to M:



(a) Left: the relationship of importance score and occurrence frequency of longicorn's visual concepts. Right: 100 categories' coefficient of determination in frequency-importance scatter gram. 83 of 100 classes of objects are shown to have a significant positive correlation between visual concepts' importance score and visual concepts' intra class occurrence frequency

(b) The relationship of importance score and occurrence frequency of longicorn's visual concepts when computing in formula (2)

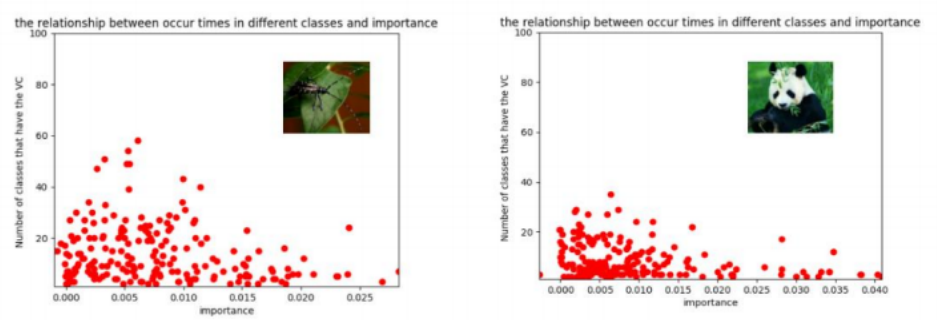
Figure 5: Typicalness analysis

$$S(V_m^k) = \frac{\sum_{I_k^i \in M_k, P_{original}(I_k^i)>0.3}(P_{original}(I_k^i) - P_{occluded}(I_k^i))}{|M|} \tag{2}$$

which only care about the average performance drop, and then draw similar things in [Figure 5b]. Though with a weaker correlation, it still demonstrate our hypothesis that there is a positive correlation between one semantic part's importance and it's 'typicalness'. In fact, we investigate the ranks of top visual concepts from both measuring methods (1) and (2), and find out that the top 10 useful visual concepts got from method (1) are always among the top 20 when using method (2), which implies the consistency of these two measuring methods and demonstrates again that critical semantic parts in recognition are always the 'typical' ones. Anyway, in order to emphasize that the useful semantic parts we extract before have universal significance in recognizing certain category of objects, we continue to use measuring method (1) in the later section of this paper.

[Figure 6] is the example of categorical retrieval. In order to judge whether 2 semantic part are similar and thus could be regarded as 'being shared' by 2 different categories, we calculate the distance between the centers of their corresponding visual concepts and apply threshold method with a magic number. [Figure 6a] gives an example of similar semantic parts in different categories, and these 2 visual concepts can be regarded as the same. By using this this method, we get [Figure 6b], which implies that there is a negative correlation between one semantic part's importance score and it's occurrence frequency among different categories. In other word, critical semantic parts in recognition are always the 'unique' ones.

6a



6b

Figure 6: (a) Corresponding semantic patches of 2 similar visual concepts in different categories. (b) Relationship between certain visual concept's importance and its occurrence frequency among different categories. Here we give 2 examples of long-horned beetle's visual concepts and giant panda's visual concepts. Each dot represents a visual concept in certain category, X axis is visual concept's importance score in certain category, and Y axis is its number of occurrences among 100 categories.

3.5 Why choose pool 3 visual concepts?

Some people may argue why we choose visual concepts extracted from pool3 to represent semantic parts instead of pool2 or pool4. Here we show some corresponding semantic parts of hyena's visual concepts extracted from pool2, pool3 and pool4, whose receptive field size is respectively 16*16, 44*44, 100*100 [Figure 7] . (the whole input image is 224*224)

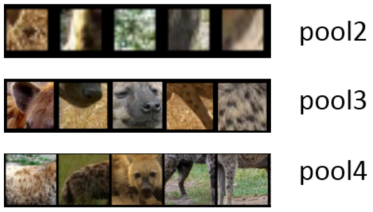


Figure 7: Corresponding semantic patches of visual concepts in pool2, pool3 and pool4

We can find that pool2 visual concepts are too ambiguous due to its low resolution, while pool4 is more sensitive to detect 'big' parts, or the complete shape of certain category of objects. Our target is to detect which semantic parts is more critical in recognition, pool2 is too small and pool4 is more likely to find the global properties, so pool3 is the best choice.

Moreover, by using the method in 3.3, we investigate different hidden layer's visual concept's occurrence frequency in different categories of objects. Pool3 visual concept is shared by 10 different categories among all 100 on average while pool2 visual concept is shared by 46 different categories

(a) The process to generate the 'Bubble' of corresponding aperture of certain visual concept.

(b) Example images from 4 sets in training dataset

(c) Example images from different sets in testing dataset. Top: Original, single and open-all set used in basic test. Middle: Combination set. Bottom: Enlargement set

Figure 8: Datasets in fine tuning process

on average, and pool4 visual concepts are always unique for certain category. Pool2 visual concepts are always common patches and pool4 visual concepts are too unique and related to the complete object. So in order to fit our motivations and ambitions, we choose pool3 visual concepts to continue our experiments.

4 Fine-tune DCNN with apertures of useful visual concepts

In above sections, we get the useful visual concepts of each category without supervision and we explain why they are key parts in recognition task. [21] shows that simplified image sometimes can also get classified as the same category as the original image if it shows critical semantic parts. Inspired by this, we use useful visual concepts to generate some simplified images which only show the receptive fields of these visual concepts, and fine-tune the Baseline Network with these aperture images, helping the networks focus on the key information for recognition task and ignore the redundant parts. Besides, this process makes the network sensitive to critical semantic part, giving a possible solution to increase the recognition accuracy under different occlusions and other difficult situations. We also have some comparative testing experiments on Baseline Network and our fine-tuned networks, disclosing some inner mechanism of DCNN.

The receptive field of certain visual concept is a square part (for pool3 it's 44 by 44) in original bounding box image. However, if we just show the theoretical receptive fields in original bounding box image with black background, the changes of pixel values at the edge of the apertures are abrupt, which may lead to misunderstanding on neighboring hidden layer feature vectors whose receptive fields include a stark straight line. In order to remove the sharp change of pixel values at the edges of apertures, we apply a technique called 'Bubbles' [5][10][19] to smooth the edges of apertures with a gray (mean pixel value of Image net images) background [Figure 8a]. This technique is applied to every aperture images of our training and testing datasets.

All the apertures in our training and testing datasets are processed by using this 'Bubble' technique and each aperture corresponds to one of the 10 most useful visual concepts in the image category.

4.1 Training process

There are 4 sets of images in our training dataset: 34K original bounding box images from 100 classes which are used to train Baseline Network, 213K images with single aperture (generated from 34K original bounding box images), 183K images with double apertures from two different visual concepts (generated from 34K original bounding box images), 106K images with double apertures

that don't have overlap with each other and from two different visual concepts (this set belongs to the previous 183K set) [Figure 8b]. We trained 5 different networks by fine tuning Baseline Network with different training strategies and different training sets. The training process was accomplished by minimizing the multinomial logistic loss by back-propagation [7]. Mini-batch gradient descent was used with 128 batch size and 0.9 momentum. The learning rate was initially 0.001 and decayed by a factor 0.1 for every 5000 iterations. The weight decay was 0.001.

Network fine-tuned with single apertures: We only fine tune the last 2 fully connected layer with 34K original bounding box images and 213K images with single aperture for 20 epochs. Training accuracy is 0.6205.

Network fine-tuned with single and double apertures: We only fine tune the last 2 fully connected layer with 34K original bounding box images, 213K images with single aperture and 183K images with double apertures for 20 epochs. Training accuracy is 0.6587.

Network fine-tuned from pool3 with single apertures: We fine tune all the layers after pool3 with 34K original bounding box images and 213K images with single aperture for 20 epochs. Training accuracy is 0.9317.

Network fine-tuned from pool3 with single and double apertures: We fine tune all the layers after pool3 with 34K original bounding box images,213K images with single aperture and 183K images with double apertures for 20 epochs. Training accuracy is 0.9162.

Network fine-tuned from pool3 with single and no-overlap double apertures: We fine tune all the layers after pool3 with 34K original bounding box images,213K images with single aperture and 106K images with no-overlap double apertures for 20 epochs. Training accuracy is 0.9265.

4.2 Test sets

Original test set: 11K original bounding box images which are used to test Baseline Network[Figure 6(c)]. All the following test sets are generated from this set.

Single aperture set: 104K images with single aperture [Figure 6(d)] just like those in training set.

Open-all set: Test images with all available apertures opened. For each original bounding box image, open all the apertures corresponding to 10 most useful visual concepts in the image category (if one visual concept has more than 1 corresponding aperture in an image, just open them all). We have 9.5 K images which at least have one aperture corresponding to 10 most useful visual concepts. [Figure 6(e)]

Combination set: This set includes 5 subsets which respectively show 1,2,3,4,5 apertures corresponding to different top visual concepts [Figure 6(f)]. Due to the fact that one top visual concept may have more than 1 corresponding aperture in an bounding box image, in order to control the number of aperture combinations, we just choose one representative aperture (the one whose pool3 feature vector is closest to the cluster center) for each top visual concept and then combine them with each other to generate images in Combination set. The number of images for each subsets is 47K, 117K, 191K, 220K and 182K. We also generate a no-overlap version (apertures in one image don't have overlaps) of this set, called No-overlap Combination set, whose subsets have 47K, 66K, 40K, 13K and 2256 images respectively. Test on this set will allow to see the level of synergistic contribution of the different top visual concepts.

Enlargement set: This set includes images with single aperture but with different aperture sizes. This set includes 5 subsets: the first subset is the same with that in Combination set, and then keep the center of aperture unchanged, but increase the diameter to enlarge the aperture to 2,3,4,5 times the size of original aperture in area [Figure 6(g)]. Each subset has 47K images. Theoretically, Enlargement set and No-overlap Combination set have the same aperture area size in corresponding subsets.

5 Experiments on aperture images

5.1 basic experiment

Firstly, we test Baseline network and our fine-tuned networks on Original test set, Single aperture set and Open-all set to see some basic performance of these networks [Table 1].

	Original test set	Single aperture set	Open-all set
Baseline Network	0.8726	0.1351	0.5762
NSA	0.8373	0.5015	0.7125
NSDA	0.8269	0.4967	0.7258
NSA-pool3	0.8555	0.7362	0.8594
NSDA-pool3	0.8376	0.7315	8665
NSDA-pool3-nooverlap	8432	0.7346	0.8655

Table 1: Top 1 accuracy of 6 different networks on 3 test sets. 'NSA' refers to network fine-tuned with single aperture (just fine tune the last 2 fully connect layer);'NSDA' refers to Network fine-tuned with single and double apertures;'NSA-pool3' refers to Network fine-tuned from pool3 with single apertures; 'NSDA-pool3' refers to Network fine-tuned from pool3 with single and double apertures; 'NSDA-pool3-nooverlap' refers to Network fine-tuned from pool3 with single and no-overlap double apertures.

It's not surprising that Baseline Network works best on Original test set because of the high radio of bounding box images in its training dataset. Also, NSA works better than NSDA, NSA-pool3 works better than NSDA-pool3 and NSDA-pool3-nooverlap due to this reason. Furthermore, we find the networks fine-tuned from pool3 (networks whose name includes 'pool3') significantly outperforms Baseline Network and the networks that only fine-tune the last 2 fully connected layers (NSA,NSDA) with an increased top 1 accuracy of at least 23% on Single aperture set and 13% on Open-all set. This improvement is due to that when fine-tuning from pool3 layer, the network can have an understanding of visual concepts from hidden layer, once it find a certain feature vector in pool3, this is enough for it to recognize the complete object. However, the network fine-tuned on last 2 fully connected layers just regards the whole aperture image as the target object, which leads to misunderstanding when aperture place changes in testing data. This also explains why even with the same radio of bounding box images in training dataset, network fined-tuned from pool3 does better than network only fine-tuned on last 2 fully connected layers on Original test set.(exp. NSA and NSA-pool3, NSDA and NSDA-pool3)

5.2 enlarge the size of apertures

When the size of aperture increases, the network's performance increases as well [Figure 9]. This shows that the networks fined tuned with useful visual concept apertures is valid not only for the trained small apertures, but can also be transferred to apertures of bigger size.

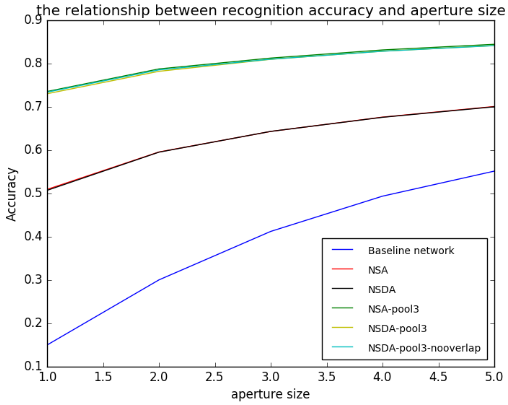


Figure 9: Top 1 accuracy for different networks on Enlargement set

Besides, what's interesting is that the curve of NSA almost coincides with that of NSDA (NSA-pool3 also coincides with NSDA-pool3 and NSDA-pool3-nooverlap), which show that there is no difference between network fine-tuned with just single apertures and network fine-tuned with single as well as double apertures. One explanation for this is that because the enlarged aperture has the same center

with the original aperture, the enlarged aperture may still just show one visual concept as the original aperture does.

In short, this experiment shows network can confirm its recognition result when it sees more of certain critical part. 'Understanding deeply' helps the network.

5.3 showing more apertures at a time

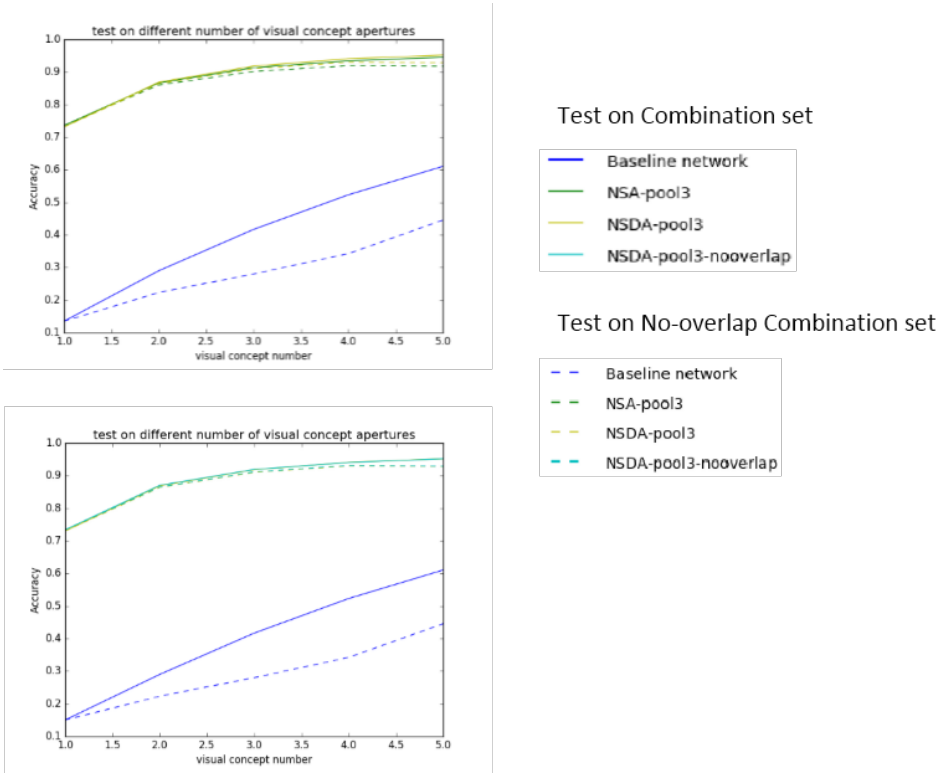
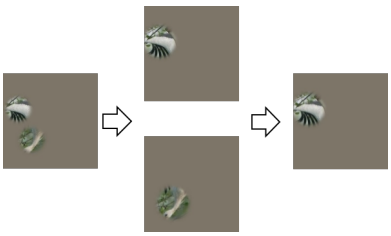


Figure 10: Top 1 accuracy for different networks on (No-overlap) Combination set. Top: comparison between NSA-pool3 and NSDA-pool3. Bottom: comparison between NSDA-pool3 and NSDA-pool3-nooverlap

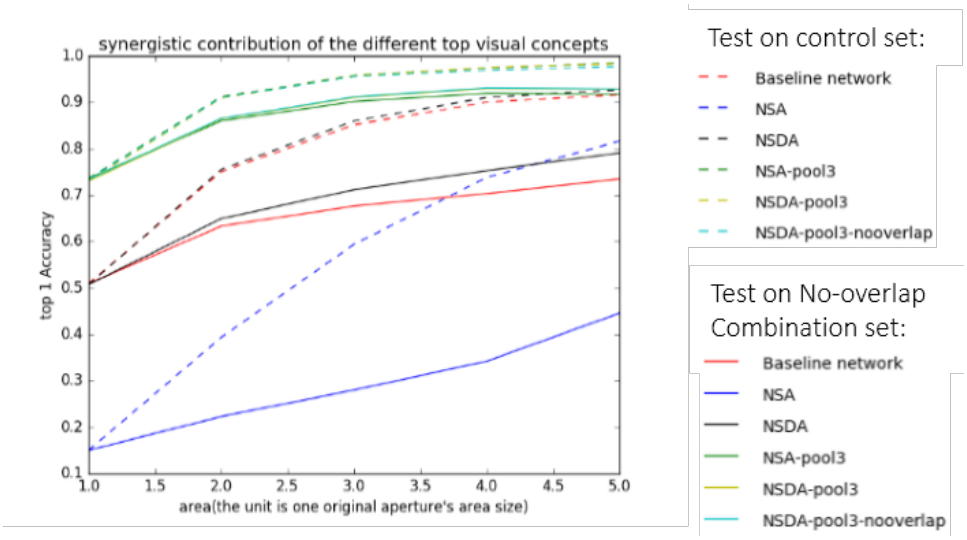
We test the networks on Combination set and No-overlap Combination set [Figure 10]. The recognition accuracy increases when more visual concept apertures are shown in the image, which means additional critical semantic part is helpful to the recognition. In other word, 'understanding widely' helps the network.

The increase also happens when testing on No-overlap Combination set (all dotted curves), which suggests that DCNN has the ability to combine and make use of even discrete semantic parts in recognition tasks. This ability is greatly enhanced in our fine-tuned networks when networks can recognize an object from just one visual concept of it. The comparison between NSA-pool3 and NSDA-pool3, and the comparison between NSDA-pool3 and NSDA-pool3-nooverlap also demonstrate this finding. There is little improvement after training with combination of visual concepts, which implies that DCNN has already known how to make use of separate partial information without being trained to do so.

Furthermore, on average, when showing the same number of visual concepts, images in Combination set tend to show less area than images in No-overlap Combination set because of overlapping. However, the test accuracy on Combination set is always higher than that on No-overlap Combination set, even just a little in fine-tuned networks. It seems that although DCNN has the ability to make use of concrete information, DCNN doesn't like separate information, our fine-tuning process just helps it to strengthen the ability to cope with separate information. We will discuss about this phenomena more clearly in next section.



(a) The process to generate control image



(b) Tests on control set and No-overlap Combination set

Figure 11

5.4 understand deeply or understand widely?

From the above sections, we know that the network performs better when it sees more of critical part of an object ('understanding deeply') and it also benefits from seeing other additional critical parts ('understanding widely'). But what is more important in the mechanism of DCNN? Understanding deeply or understanding widely? When showing partial images to an DCNN, which will have a higher recognition accuracy? A big aperture with one critical visual concept inside or several small apertures of different useful visual concepts?

We design an experiment to explore this question. For each image that shows more than 1 visual concept in No-overlap Combination set, we enlarge every component of the showing part to the same area size of the original image in No-overlap Combination set (e.g. If original image opens 2 aperture, then we enlarge each aperture to twice the aperture size, and thus generating 2 enlarged single aperture images). For all these enlarged components, we choose the one with highest recognition probability as the control image of the original image in No-overlap Combination set [Figure 11a]. Thereby, we generate a control set with the same number of images with No-overlap Combination set. If top-1 recognition accuracy of No-overlap Combination set is higher than its control set in any subset, we then could conclude that synergistic contribution of different top visual concepts plays a big role in recognition, and several small visual concepts is more important than one confirmed big aperture of critical visual concept.

However, the result is just the opposite [Figure 11b], for all the networks, accuracy on control set is higher than that on No-overlap Combination, which shows that 'one big aperture' is more welcomed than 'several small apertures' even they have the same total area size. In other word, for DCNN, a confirmed big critical semantic part is the real key to recognize an object, rather than the combination several ambiguous small semantic parts. The effect of fine tuning is to improve DCNN's ability to make use of discrete cues (see the smaller gap between test on 2 sets after fine-tuning), but still, a solid evidence is more important in recognition.

Understanding deeply is more important than understanding widely, although DCNN has the ability to make use of additional discrete information to guarantee the accuracy of recognition, a confirmed semantic part is more welcomed in recognition task.

6 Experiments on occluded objects

6.1 bar test

In order to show the robustness of our fine-tuned network, we also apply it in different situation. In this experiment, test images are placed with several grey bars with different quantity, width and intervals, which seems that the images are put in jail. [Figure 12] shows some test results and sample images.

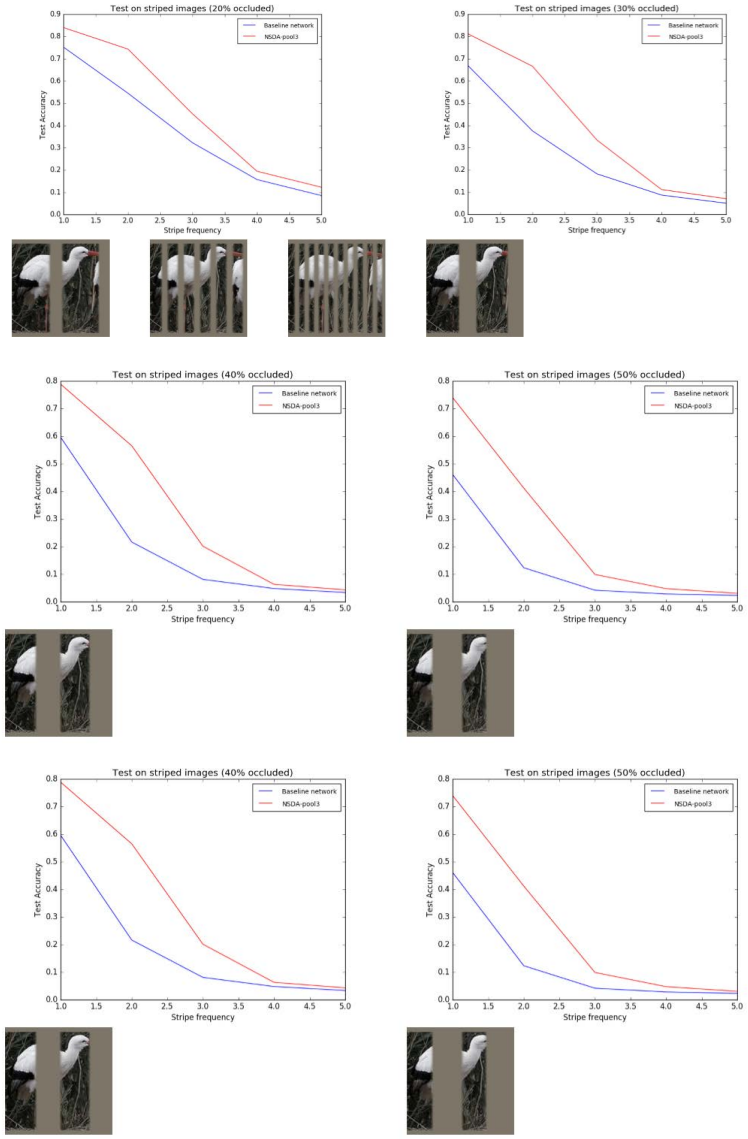


Figure 12: Bar test, each figure shows the recognition accuracy when there are different number of bars under the same occlusion area

We can find that in this test, NSDA-pool3 still shows significant performance improvement than Baseline network. This is because the gap between 2 bars may contain some visual concepts, and

these visual concepts can be recognized by NSDA-pool3, thus improving the recognition rate of the complete object. In [Figure 12], each figure shows the recognition accuracy of different networks when there are different number of bars under the same recognition accuracy. We can find that even under the same occlusion level, when increasing the number of bars, the recognition accuracy is continually decreasing, which may due to the fact that the width of exposed image between 2 bars are smaller and smaller and have less visual concepts.

This experiment demonstrates that the DCNN fine-tuned with important visual concepts(NSDA-pool3) is robust. However, we can still find some limitation of our fine-tuned network. When there are very dense bars, we human can still recognize the object because we can imagine the pixels behind the bar and make up a complete object, while neural network cannot do so, which is another important factor leading to the failure of recognizing occluded objects by DCNN.

6.2 rectangle test

Previous experiment shows that DCNN fine-tuned with important visual concepts have robustness under different occlusion situation, but we don't have comparison with other works. Furthermore, we don't know whether visual concept works in occluded object recognition, or just because of data augmentation.

Here we compare our network with Wang's network in his test set. The test images are bounding box images with grey rectangle occluder with different occlusion level, from 0% to 90%. The position of the rectangle is random and the size of it is in scale with the size of whole bounding box image.

In order to show the superiority of using visual concepts, we also train another random patch network, which use the same number of random apertures instead of top visual concepts apertures to fine tune the baseline network.

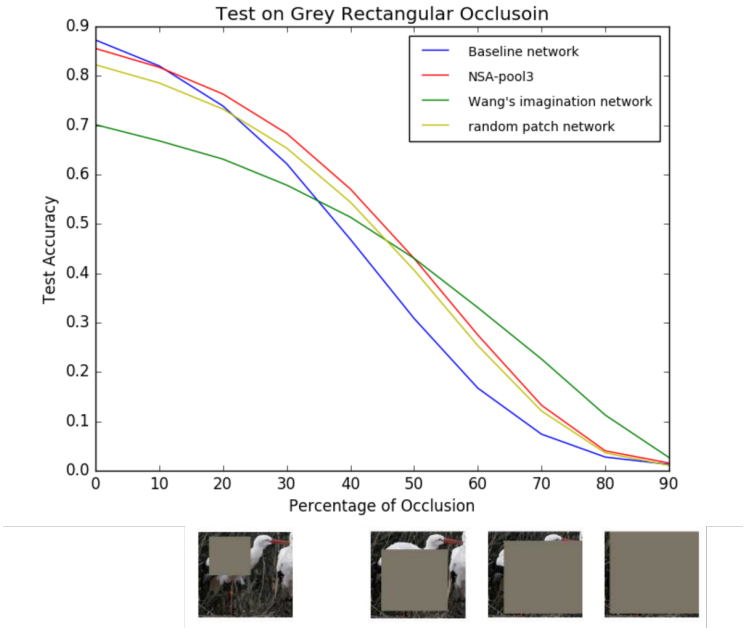


Figure 13: Rectangle occluder test.

The result [Figure 13] shows that DCNN fine-tuned with visual concepts performs better than that find-tuned with random patches, which means visual concepts itself is useful in occluded object recognition.Besides, we find that Wang's network has higher recognition accuracy than other networks when occlusion level is high, while it performs poorly when occlusion level is low. The training process of Wang's network implies that the network makes use of background information of an object, which leads to higher recognition accuracy when only background is exposed, while makes a compromise when occlusion level is low. Wang's network has grey rectangle occluder images as

its training set, so the test cannot show its robustness. While our training set doesn't include grey rectangle occluders, which shows the robustness of our network.

7 Conclusion

There are 3 major contributions of our work: firstly, we develop a method to automatically extract meaningful semantic parts of certain category; secondly, we find some inner mechanism of DCNN, although DCNN has the ability to make use of discrete object parts, a confirmed and big-enough semantic part plays a more important role in recognition process, thirdly, we make DCNN acquire human's ability of recognizing object by part, propose a new data augmentation method which can greatly improve the recognition accuracy when the object is partly occluded. Our work shed light on the understanding of DCNN and give a possible and accessible solution for recognizing partially occluded objects. In the future, we want to answer why a confirmed big semantic part is more welcomed than several small parts for DCNN. One hypothesis is that big aperture can make full use of higher layer's neuron to extract the inner meaning of it, however, small apertures have little influence on higher layer's neurons. Another hypothesis is that the surrounding part in big aperture helps to confirm the network's judgement. We will also explore the visual neurons in brain to help us get a deeper understanding of how human recognize things.

References

[1] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2018), pp. 834–848.

[2] Golnaz Ghiasi and Charless C Fowlkes. "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2385–2392.

[3] Shiry Ginosar et al. "Detecting people in cubist art". In: *Workshop at the European Conference on Computer Vision*. Springer. 2014, pp. 101–116.

[4] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[5] Frédéric Gosselin and Philippe G. Schyns. "Bubbles: a technique to reveal the use of information in recognition tasks". In: *Vision Research* 41 (2001), pp. 2261–2271.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[7] Y. Lecun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (2014), pp. 541–551.

[8] Yao Li et al. "Mid-level deep pattern mining". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 971–980.

[9] Renjie Liao et al. "Learning deep parsimonious representations". In: *Advances in Neural Information Processing Systems*. 2016, pp. 5076–5084.

[10] Adolphs Ralph et al. "A mechanism for impaired fear recognition after amygdala damage". In: *Nature* 433.7021 (2005), p. 68.

[11] Antonio Rama et al. "More robust face recognition by considering occlusion information". In: *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE. 2008, pp. 1–6.

[12] Marcel Simon and Erik Rodner. "Neural activation constellations: Unsupervised part model discovery with convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1143–1151.

[13] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[14] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. "Unsupervised discovery of mid-level discriminative patches". In: *Computer Vision–ECCV 2012*. Springer, 2012, pp. 73–86.

[15] "Tcav: Relative Concept Importance Testing with Linear Concept Activation Vectors". In: 2017.

[16] Shimon Ullman et al. "Atoms of recognition in human and computer vision". In: *Proceedings of the National Academy of Sciences* 113.10 (2016), pp. 2744–2749.

[17] Hao Wang et al. "Learning Robust Object Recognition Using Composed Scenes from Generative Models". In: *2017 14th Conference on Computer and Robot Vision (CRV)* (2017), pp. 232–239.

[18] Jianyu Wang et al. "Unsupervised learning of object semantic parts from internal states of CNNs by population encoding". In: *Computer Science* (2015).

[19] Keith Worsley et al. "Detecting connectivity between images: MS lesions, cortical thickness, and the 'bubbles' task in fMRI?" In: (Jan. 2007).

[20] Ying Wu, Ting Yu, and Gang Hua. "A statistical field model for pedestrian detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 1023–1030.

[21] Bolei Zhou et al. "Object Detectors Emerge in Deep Scene CNNs". In: *CoRR* abs/1412.6856 (2014).