CNN FEATURES OFF-THE-SHELF: AN ASTOUNDING BASELINE FOR RECOGNITION

ALI SHARIF RAZAVIAN HOSSEIN AZIZPOUR JOSEPHINE SULLIVAN STEFAN CARLSSON

Umme Fatema

CONTRIBUTION

- This paper consolidates that features extracted from the convolutional neural networks (CNN) are very powerful descriptors in visual recognition tasks.

INTRODUCTION

- The paper answers the widely asked question in Computer Vision which is whether the features extracted from the CNN (trained with a diverse ImageNet dataset) can be a good descriptor for visual recognition tasks.
- Different recognition tasks are performed using publicly available code and model of the OverFeat network.
- OverFeat network is trained on ILSVRC13 [1] (ImageNet Large Scale Visual Recognition Challenge 2013).
- Recognition tasks that are analyzed are given below:
  - object image classification
  - scene recognition
  - fine grained recognition
  - attribute detection
  - image retrieval

1. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), pp.211-252.

INTRODUCTION

- Results are achieved using linear support vector machine (SVM) classifier (L2 distance in case of retrieval) where the feature is the feature vector of size 4096 extracted from the trained OverFeat network.
- The representation is further modified by using data augmentation. for example, jittering (flipping, cropping, color casting, distortion etc.)
- Results are compared with state-of-the-art systems for each classification task.
- The results proves that features extracted from the CNN are very good descriptor for visual recognition tasks.

INTRODUCTION

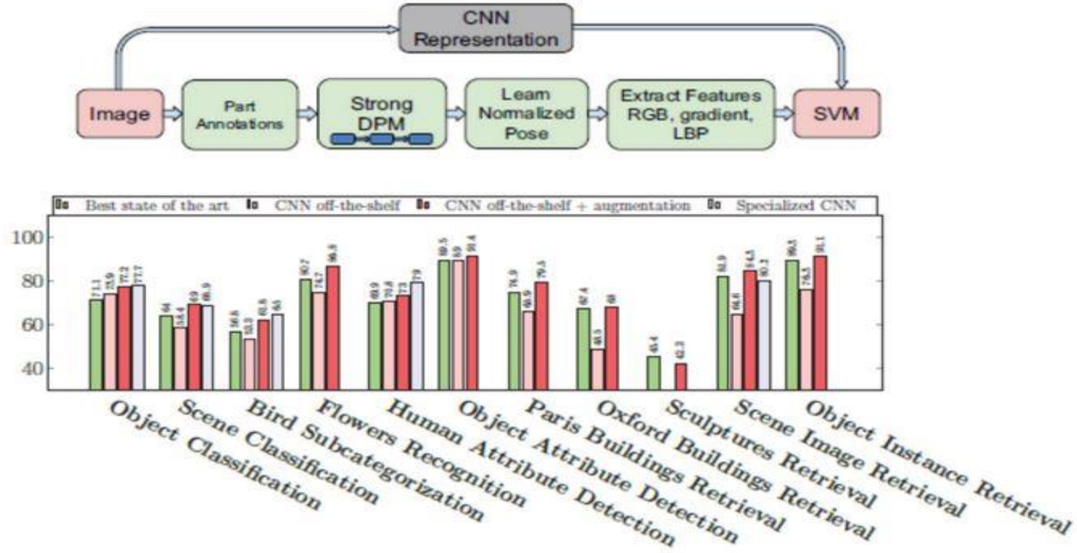
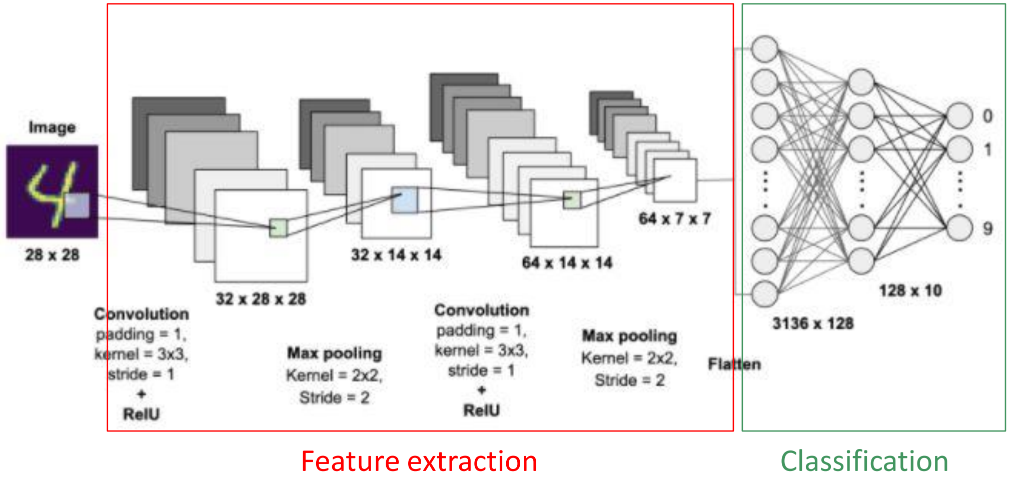


Figure 1: **top**) CNN representation replaces pipelines of s.o.a methods and achieve better results. e.g. DPD [50].
bottom) Augmented CNN representation with linear SVM consistently outperforms s.o.a. on multiple tasks. Specialized CNN refers to other works which specifically designed the CNN for their task

BACKGROUND

<u>CNN Architecture:</u>

- CNN takes the image as input.
- The convolution layers apply convolution operation to the input and pass the result to next layer.
- Each convolutional layer is followed by an activation function for non-linearity.
- The pooling layer (max/average) subsample input in order to reduce computational load and number of parameters.
- CNN consists of stacks of convolutional layers and activation function followed by pooling layers.
- Number of filters increases at higher layers.
- Flatten gives a feature vector that is passed to the fully connected layers.



- Fully connected layers are regular feedforward neural network.
- Fully connected layers have more parameters to train.
- The final output layer is a prediction layer Softmax for classification probabilities.

BACKGROUND

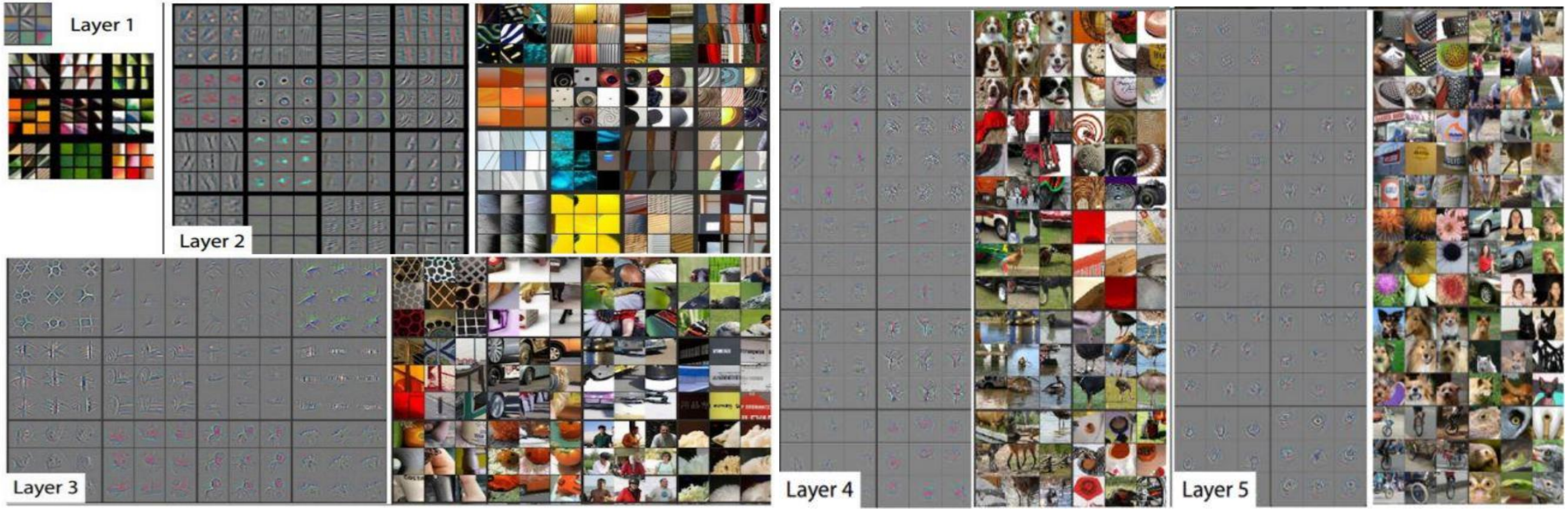
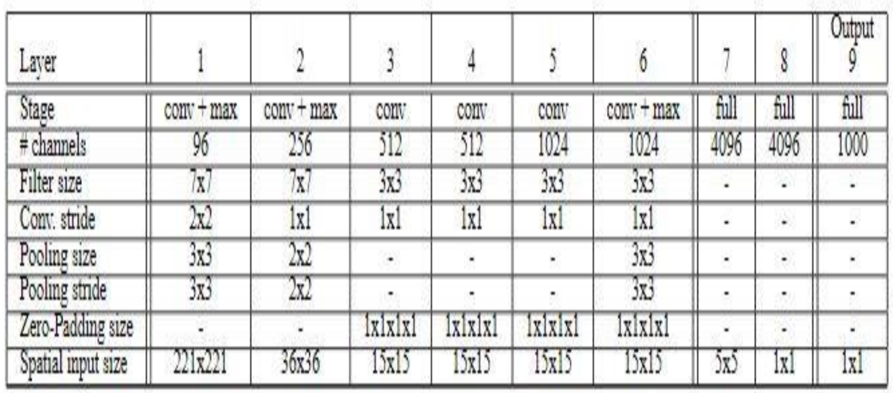


figure: Features extracted from each layer of a CNN

BACKGROUND AND OUTLINE

- Publicly available trained CNN called OverFeat [2] is used.
- The structure of this CNN is similar to Krizhevsky [3].
- The convolutional layers each contain 96 to 1024 kernels of size 3x3 to 7x7.
- Half-wave rectification is used as the nonlinear activation function.
- Max pooling kernels of size 3x3 and 5x5 are used at different layers to build robustness to intra-class deformations.
- The input size for the CNN is 221x221x3



Layer	1	2	3	4	5	6	7	8	Output 9
Stage	conv + max	conv + max	conv	conv	conv	conv + max	full	full	full
# channels	96	256	512	512	1024	1024	4096	4096	1000
Filter size	7x7	7x7	3x3	3x3	3x3	3x3	-	-	-
Conv. stride	2x2	1x1	1x1	1x1	1x1	1x1	-	-	-
Pooling size	3x3	2x2	-	-	-	3x3	-	-	-
Pooling stride	3x3	2x2	-	-	-	3x3	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	221x221	36x36	15x15	15x15	15x15	15x15	5x5	1x1	1x1

Basic Structure of OverFeat network

2. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.
3. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, pp.1097-1105.

BACKGROUND AND OUTLINE

- OverFeat was trained on ImageNet ILSVRC 2013 and obtained very good results for the classification task of the 2013 challenge and won the localization task.
- ILSVRC13 contains 1.2 million images which are hand labelled with the presence/absence of 1000 categories.
- The images are mostly centered, and the dataset is considered less challenging in terms of clutter and occlusion than other object recognition datasets such as PASCAL VOC.
- In the paper, experiments are conducted on different recognition tasks.
- The tasks and datasets were selected such that they gradually move further away from the task the OverFeat network is trained to perform.

BACKGROUND AND OUTLINE

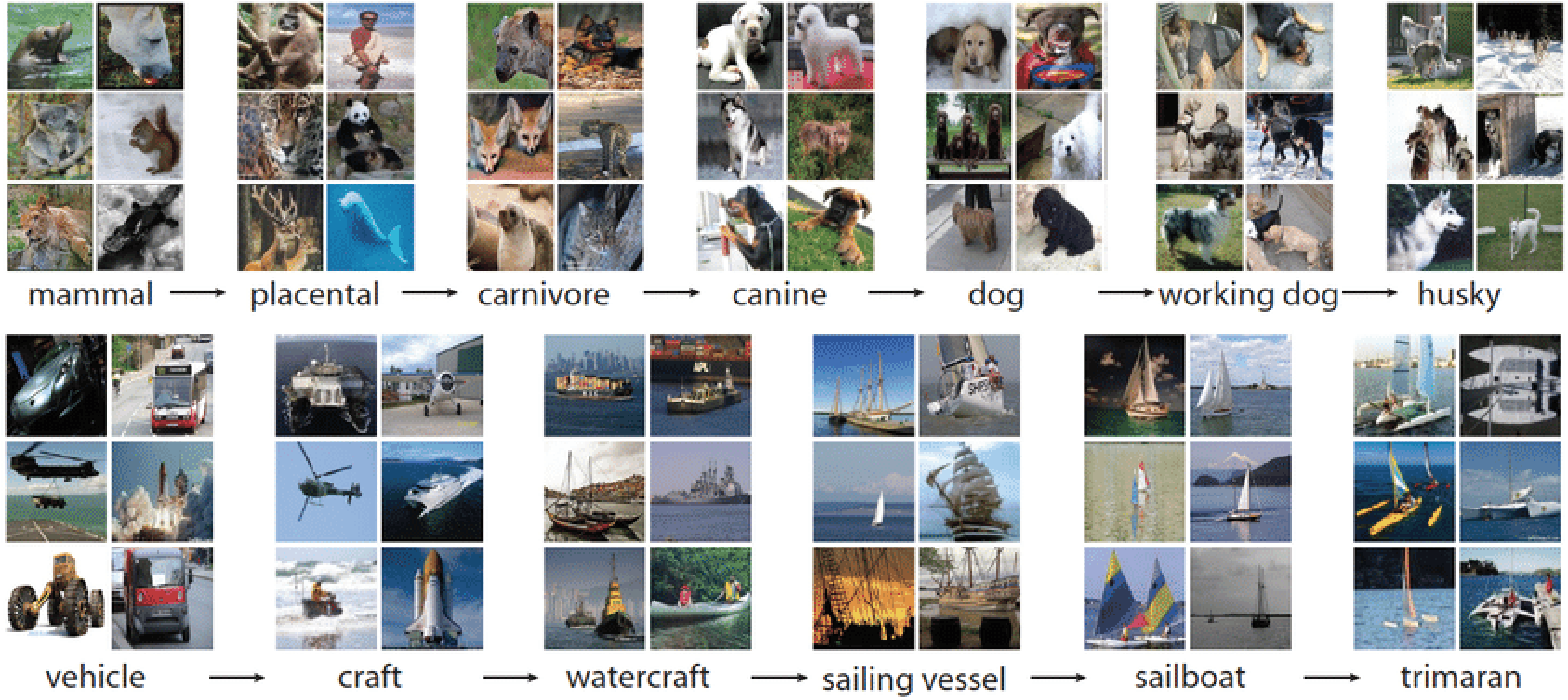

mammal → placental → carnivore → canine → dog → working dog → husky
vehicle → craft → watercraft → sailing vessel → sailboat → trimaran

figure: ImageNet

VISUAL CLASSIFICATION

- For the experiments, the first fully connected layer (layer 22) of the network is used as the feature vector.
- The input for the OverFeat network is images resized to 221x221 and the feature vector dimension of the extracted feature from the network is 4096.
- Here two settings have been used:
  - The feature vector is further L2 normalized to unit length for all the experiments. Then the 4096-dimensional feature vector in combination with a linear Support Vector Machine (SVM) is used for classification tasks (CNN-SVM).
  - The training set is augmented by adding cropped and rotated samples and doing component wise power transform and is used with a SVM for classification tasks. (CNNaug+SVM).
- For the classification scenarios where the labels are not mutually exclusive a one-against-all strategy I used. In the rest of the experiments, one-against-one linear SVMs is used.
- For all the experiments a linear SVM is used

IMPLEMENTATION DETAILS

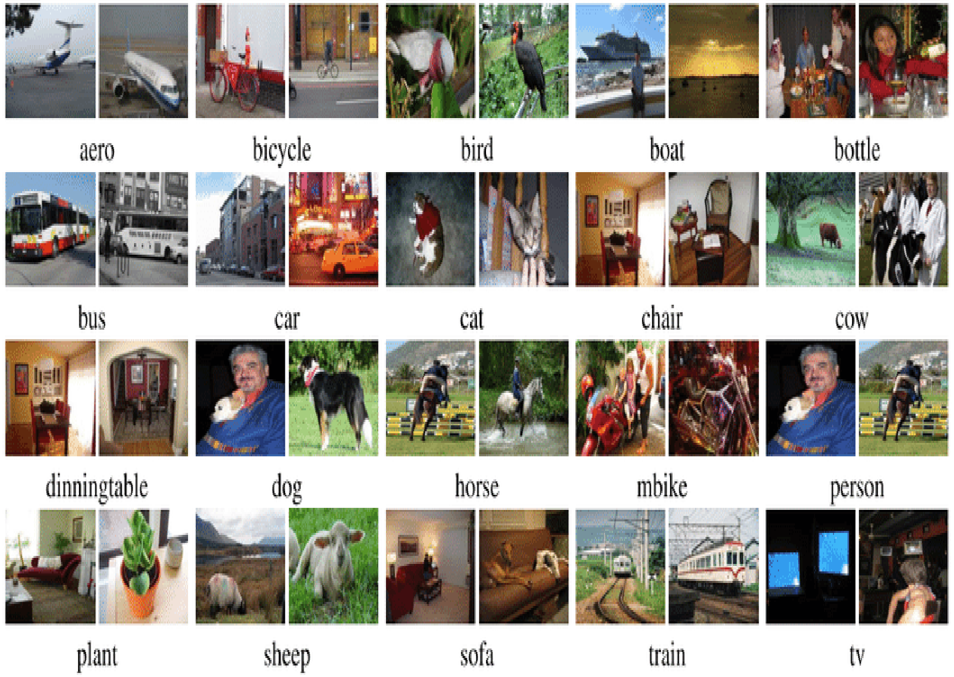
- Precomputed linear kernels with libsvm is used for the CNN-SVM experiments and liblinear for the CNNaug- SVM with the primal solver (#samples#dim).
- Data augmentation is done by making 16 representations for each sample (original image, 5 crops, 2 rotation and their mirrors).
- For CNNaug-SVM, signed component-wise power transform is used by raising each dimension to the power of 2.
- For this case, one-vs-one approach SVM works better than structured SVM for multi-class learning.

IMAGE CLASSIFICATION

- The first recognition task that is tested is the image classification of objects and scenes.
- The CNN network is trained with ILSVRC.
- Two different image classification datasets are used for the task. They are below:
  - **Pascal VOC 2007 : It** contains 10000 imagesof 20 classes including animals, handmade and natural. . The objects are not centered and in general the appearance of objects in VOC is perceived to be more challenging than ILSVRC.
  - **MIT-67 indoor scene :** It contains 15620 images of 67 indoor scenes. The dataset consists of different types of stores , residential rooms, public, and working places. The similarity of the objects present in different indoor scenes makes MIT indoor an especially difficult dataset compared to outdoor scene datasets.

IMAGE CLASSIFICATION

Pascal VOC 2007 dataset



MIT-67 indoor scenes dataset

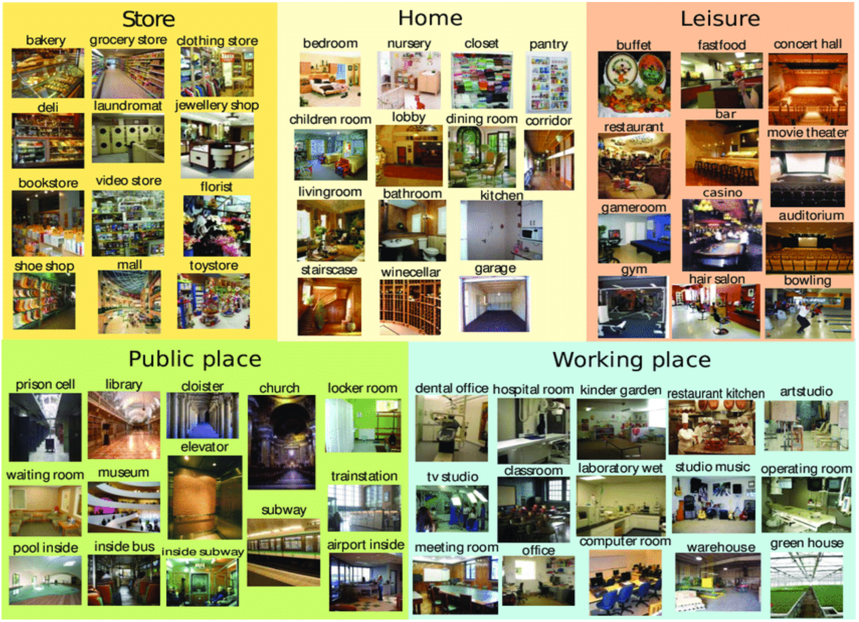
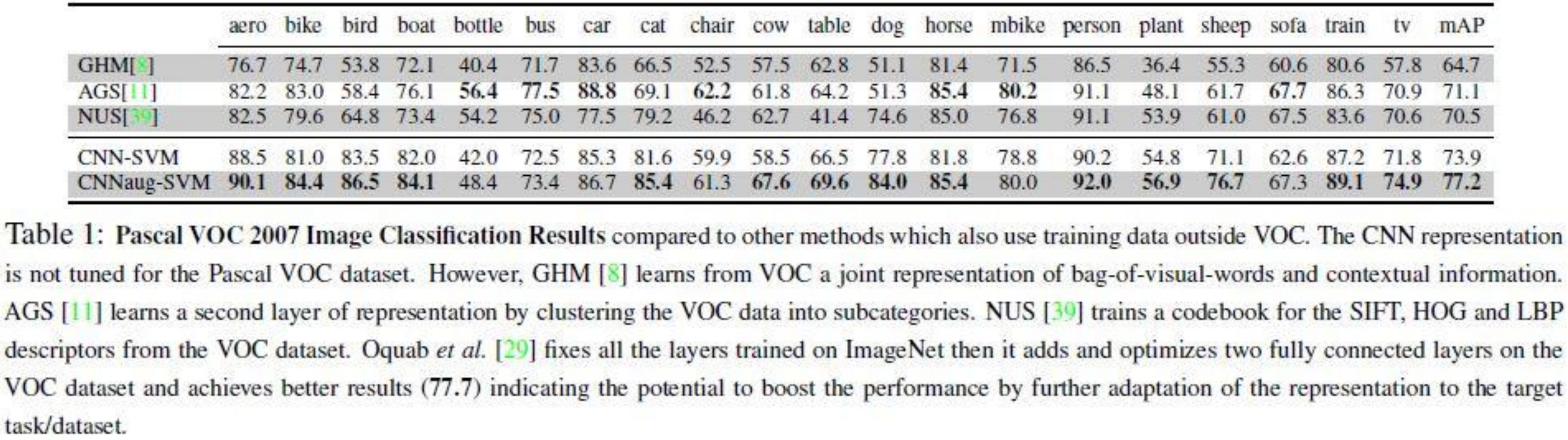


IMAGE CLASSIFICATION

- Table 1, shows the results of the OverFeat CNN representation for object image classification.
- The performance is measured using average precision (AP) criterion.
- The results are compared with methods which have used training data outside the standard Pascal VOC 2007 dataset.
- The method outperforms all the previous efforts by a significant margin in mean average precision (mAP).
- It has superior average precision on 10 out of 20 classes.

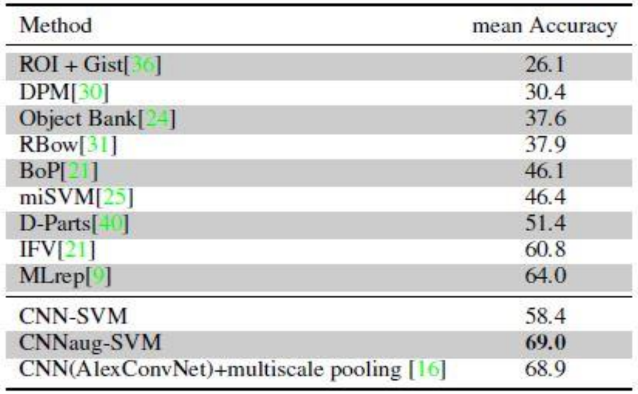


	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
GHM[8]	76.7	74.7	53.8	72.1	40.4	71.7	83.6	66.5	52.5	57.5	62.8	51.1	81.4	71.5	86.5	36.4	55.3	60.6	80.6	57.8	64.7
AGS[11]	82.2	83.0	58.4	76.1	**56.4**	**77.5**	**88.8**	69.1	**62.2**	61.8	64.2	51.3	**85.4**	**80.2**	91.1	48.1	61.7	**67.7**	86.3	70.9	71.1
NUS[39]	82.5	79.6	64.8	73.4	54.2	75.0	77.5	79.2	46.2	62.7	41.4	74.6	85.0	76.8	91.1	53.9	61.0	67.5	83.6	70.6	70.5
CNN-SVM	88.5	81.0	83.5	82.0	42.0	72.5	85.3	81.6	59.9	58.5	66.5	77.8	81.8	78.8	90.2	54.8	71.1	62.6	87.2	71.8	73.9
CNNaug-SVM	**90.1**	**84.4**	**86.5**	**84.1**	48.4	73.4	86.7	**85.4**	61.3	**67.6**	**69.6**	**84.0**	**85.4**	80.0	**92.0**	**56.9**	**76.7**	67.3	**89.1**	**74.9**	**77.2**

Table 1: **Pascal VOC 2007 Image Classification Results** compared to other methods which also use training data outside VOC. The CNN representation is not tuned for the Pascal VOC dataset. However, GHM [8] learns from VOC a joint representation of bag-of-visual-words and contextual information. AGS [11] learns a second layer of representation by clustering the VOC data into subcategories. NUS [39] trains a codebook for the SIFT, HOG and LBP descriptors from the VOC dataset. Oquab *et al.* [29] fixes all the layers trained on ImageNet then it adds and optimizes two fully connected layers on the VOC dataset and achieves better results (77.7) indicating the potential to boost the performance by further adaptation of the representation to the target task/dataset.

IMAGE CLASSIFICATION

- Table 2, shows the results of different methods on the MIT indoor dataset.
- The performance is measured by the average classification accuracy of different classes (mean of the confusion matrix diagonal).
- Using a CNN off-the-shelf with linear SVMs training significantly outperforms a majority of the baselines.
- The few relatively bright off-diagonal points are annotated with their ground truth and estimated labels. These labels could be challenging even for a human to distinguish.



Method	mean Accuracy
ROI + Gist[36]	26.1
DPM[30]	30.4
Object Bank[24]	37.6
RBow[31]	37.9
BoP[21]	46.1
miSVM[25]	46.4
D-Parts[40]	51.4
IFV[21]	60.8
MLrep[9]	64.0
CNN-SVM	58.4
CNNaug-SVM	**69.0**
CNN(AlexConvNet)+multiscale pooling [16]	68.9

Table 2: **MIT-67 indoor scenes dataset.** The MLrep [9] has a fine tuned pipeline which takes weeks to select and train various part detectors. Furthermore, Improved Fisher Vector (IFV) representation has dimensionality larger than 200K. [16] has very recently tuned a multi-scale orderless pooling of CNN features (off-the-shelf) suitable for certain tasks. With this simple modification they achieved significant average classification accuracy of **68.88**.

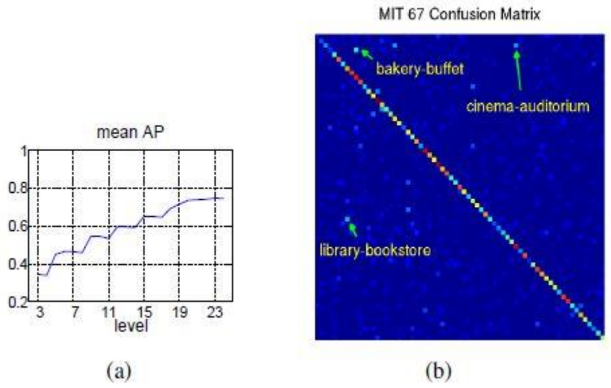


Figure 2: **a)** Evolution of the mean image classification AP over PASCAL VOC 2007 classes as we use a deeper representation from the `OverFeat` CNN trained on the ILSVRC dataset. `OverFeat` considers convolution, max pooling, nonlinear activations, etc. as separate layers. The re-occurring decreases in the plot is of the activation function layer which loses information by half rectifying the signal. **b)** Confusion matrix for the MIT-67 indoor dataset. Some of the off-diagonal confused classes have been annotated, these particular cases could be hard even for a human to distinguish.

OBJECT DETECTION

- For the task of object detection, authors of the paper did not conduct any experiments.
- It is mentioned that Girshick et al. [4] have reported remarkable numbers on PASCAL VOC 2007 using off-the-shelf features from Caffe code. Using off-the-shelf features, they achieve a mAP of 46.2 which already outperforms state of the art by about 10%.
- This adds to the evidences of how powerful the CNN features off-the-shelf are for visual recognition tasks.

4. R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arxiv:1311.2524 [cs.CV], 2013.

FINE GRAINED RECOGNITION

- Fine grained recognition is a process of recognizing subclasses of the same object class such as different bird species, dog breeds, flower types, etc.
- It requires fine detailed representation to recognize the subtle differences across different subordinate classes (as opposed to different categories)
- Because of this characteristic, fine-grained recognition is considered as a good test to identify of whether a generic representation can capture these subtle details.
- Datasets used for this method are:
  - **Caltech-UCSD Birds (CUB) 200-2011 dataset** : It contains 11,788 images of 200 bird subordinates. 5994 images are used for training and 5794 for evaluation. Many of the species in the dataset exhibit extremely subtle differences, even hard for humans to distinguish.
  - **Oxford 102 flowers dataset:** It contains 40 to 258 of images of 102 categories of flowers. The flowers appear at different scales, pose and lighting conditions. The dataset provides segmentation for all the images.

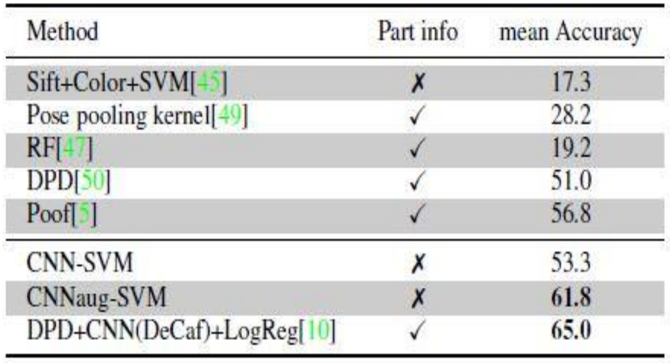
FINE GRAINED RECOGNITION

Caltech-UCSD Birds (CUB)

Oxford 102 Flowers

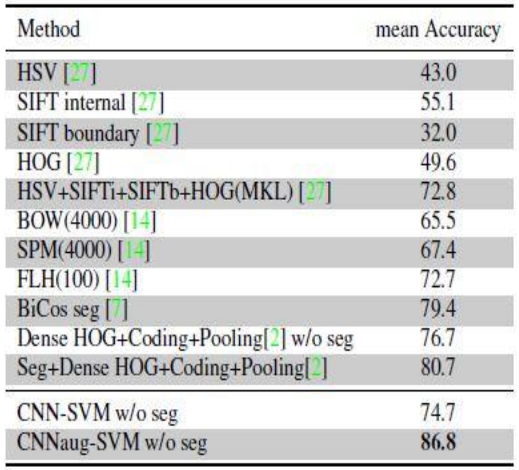
FINE GRAINED RECOGNITION

- Table 3 shows the results of the CNN-SVM compared to the top performing baselines on the Caltech-UCSD 200-2011 dataset.
- Table 4 shows the performance of CNN-SVM and other baselines on the Oxford flower's dataset.
- The CNN-SVM outperforms all basic representations and their multiple kernel combination even without using segmentation.



Method	Part info	mean Accuracy
Sift+Color+SVM[45]	✗	17.3
Pose pooling kernel[49]	✓	28.2
RF[47]	✓	19.2
DPD[50]	✓	51.0
Poof[5]	✓	56.8
CNN-SVM	✗	53.3
CNNaug-SVM	✗	**61.8**
DPD+CNN(DeCaf)+LogReg[10]	✓	**65.0**

Table 3: **Results on CUB 200-2011 Bird dataset.** The table distinguishes between methods which use part annotations for training and sometimes for evaluation as well and those that do not. [10] generates a pose-normalized CNN representation using DPD [50] detectors which significantly boosts the results to **64.96**.



Method	mean Accuracy
HSV [27]	43.0
SIFT internal [27]	55.1
SIFT boundary [27]	32.0
HOG [27]	49.6
HSV+SIFTi+SIFTb+HOG(MKL) [27]	72.8
BOW(4000) [14]	65.5
SPM(4000) [14]	67.4
FLH(100) [14]	72.7
BiCos seg [7]	79.4
Dense HOG+Coding+Pooling[2] w/o seg	76.7
Seg+Dense HOG+Coding+Pooling[2]	80.7
CNN-SVM w/o seg	74.7
CNNaug-SVM w/o seg	**86.8**

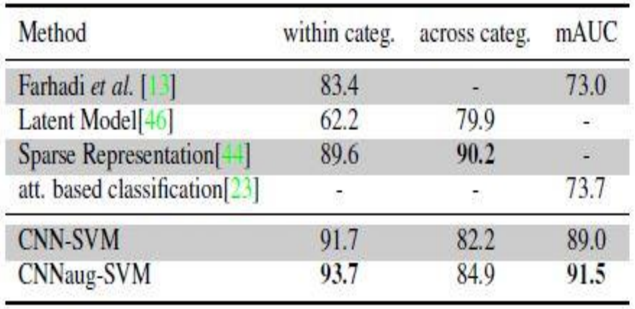
Table 4: **Results on the Oxford 102 Flowers dataset.** All the methods use segmentation to subtract the flowers from background unless stated otherwise.

ATTRIBUTE DETECTION

- Within the context of computer vision, an attribute is defined as some semantic or abstract quality which different instances/categories share.
- Two sets of datasets are used for attribute detection. They are below:
  - **UIUC 64 object attributes dataset:** There are 3 categories of attributes in this dataset:
    - shape (e.g. is 2D boxy)
    - part (e.g. has head)
    - material (e.g. is furry).
  - **The H3D dataset:** which defines 9 attributes for a subset of the person images from Pascal VOC 2007. The attributes range from “has glasses” to “is male”.

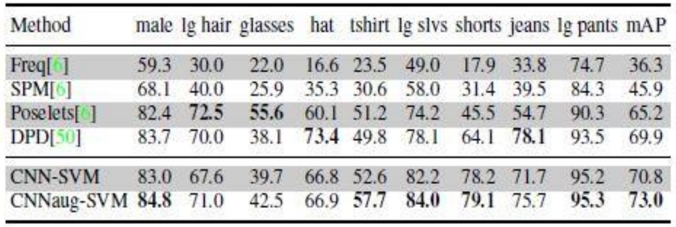
ATTRIBUTE DETECTION

- Table 5 compares CNN features performance to state-of the-art for UIUC 64 object attributes dataset.
- Table 6 reports the results of the detection of 9 human attributes on the H3D dataset



Method	within categ.	across categ.	mAUC
Farhadi *et al.* [13]	83.4	-	73.0
Latent Model[46]	62.2	79.9	-
Sparse Representation[44]	89.6	**90.2**	-
att. based classification[23]	-	-	73.7
CNN-SVM	91.7	82.2	89.0
CNNaug-SVM	**93.7**	84.9	**91.5**

Table 5: **UIUC 64 object attribute dataset results.** Compared to other existing methods the CNN features perform very favorably.



Method	male	lg hair	glasses	hat	tshirt	lg slvs	shorts	jeans	lg pants	mAP
Freq[6]	59.3	30.0	22.0	16.6	23.5	49.0	17.9	33.8	74.7	36.3
SPM[6]	68.1	40.0	25.9	35.3	30.6	58.0	31.4	39.5	84.3	45.9
Poselets[6]	82.4	**72.5**	**55.6**	60.1	51.2	74.2	45.5	54.7	90.3	65.2
DPD[50]	83.7	70.0	38.1	**73.4**	49.8	78.1	64.1	**78.1**	93.5	69.9
CNN-SVM	83.0	67.6	39.7	66.8	52.6	82.2	78.2	71.7	95.2	70.8
CNNaug-SVM	**84.8**	71.0	42.5	66.9	**57.7**	**84.0**	**79.1**	75.7	**95.3**	**73.0**

Table 6: **H3D Human Attributes dataset results.** A CNN representation is extracted from the bounding box surrounding the person. All the other methods require the part annotations during training. The first row shows the performance of a random classifier. The work of Zhang *et al.* [51] has adapted the CNN architecture specifically for the task of attribute detection and achieved the impressive performance of **78.98** in mAP. This further highlights the importance of adapting the CNN architecture for different tasks given enough computational resources.

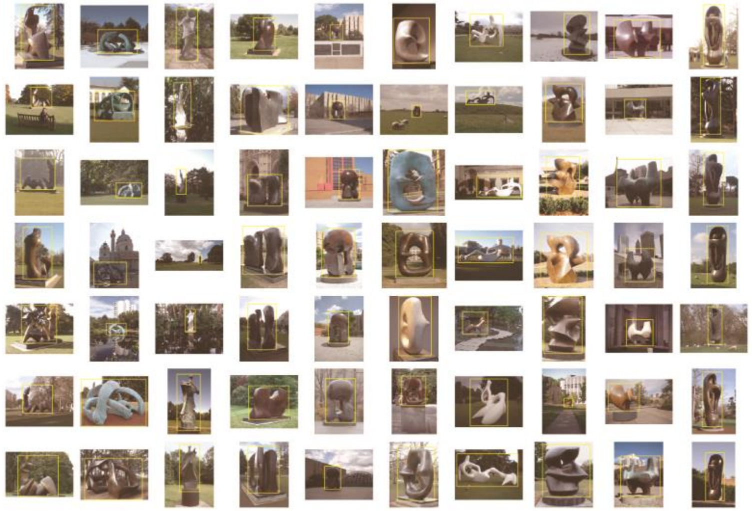
VISUAL INSTANCE RETRIEVAL

- The CNN representation is compared to the current state-of-the-art retrieval pipelines including VLAD, BoW, IFV, Hamming Embedding, and BoB.
- Unlike the CNN representation, all the above methods use dictionaries trained on similar or same dataset as they are tested on.
- For a fair comparison between the methods, results posted with relevant order of dimensions and excluded post-processing.
- Datasets are used for visual instance retrieval are below:
  - **Oxford 5k building:** 5063 reference photos gathered from flickr, and 55 queries of different buildings.
  - **Paris 6k buildings:** 55 queries images of buildings and monuments from Paris and 6412 reference photos.
  - **Sculptures 6k:** 70 query images and contains 6340 reference images.
  - **Holidays dataset:** contains 1491 images of which 500 are queries. It contains images of different scenes, items and monuments.
  - **Uk bench:** 2250 items each from four different viewpoints.

VISUAL INSTANCE RETRIEVAL

Oxford5k and Paris 6k buildings

Holidays Dataset



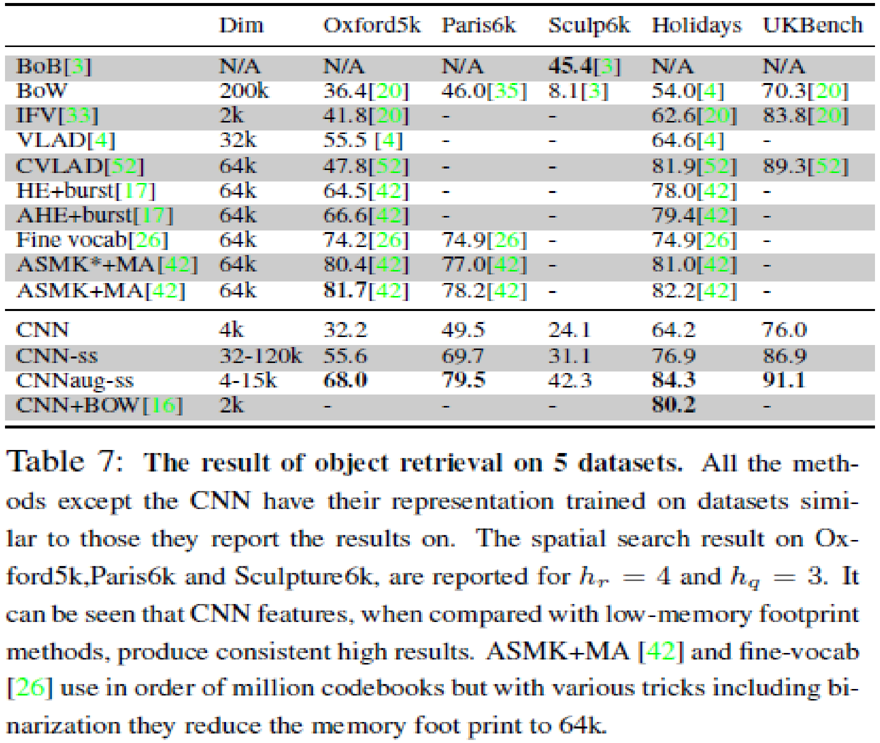
Sculpture 6k

VISUAL INSTANCE RETRIEVAL

- L2 normalized output of the first fully connected layer is used as representation.
- The items of interest can appear at different locations and scales in the test and reference images which made spatial search necessary.
- For each image, multiple sub-patches of different sizes at different locations are extracted.
- For each extracted sub-patch, its CNN representation is computed.
- The distance between a query sub-patch and a reference image is defined as the minimum L2 distance between the query sub-patch and respective reference sub-patches.
- The smallest square containing the region of interest is extracted.
- Feature Augmentation: The extracted 4096 dim features is processed in the following way:
  - L2 normalize → PCA dimensionality reduction → whitening → L2 renormalization → a signed component wise power transform and raise each dimension of the feature vector to the power of 2.

VISUAL INSTANCE RETRIEVAL

The result for different retrieval methods for 5 datasets:



	Dim	Oxford5k	Paris6k	Sculp6k	Holidays	UKBench
BoB[3]	N/A	N/A	N/A	**45.4**[3]	N/A	N/A
BoW	200k	36.4[20]	46.0[35]	8.1[3]	54.0[4]	70.3[20]
IFV[33]	2k	41.8[20]	-	-	62.6[20]	83.8[20]
VLAD[4]	32k	55.5 [4]	-	-	64.6[4]	-
CVLAD[52]	64k	47.8[52]	-	-	81.9[52]	89.3[52]
HE+burst[17]	64k	64.5[42]	-	-	78.0[42]	-
AHE+burst[17]	64k	66.6[42]	-	-	79.4[42]	-
Fine vocab[26]	64k	74.2[26]	74.9[26]	-	74.9[26]	-
ASMK*+MA[42]	64k	80.4[42]	77.0[42]	-	81.0[42]	-
ASMK+MA[42]	64k	**81.7**[42]	78.2[42]	-	82.2[42]	-
CNN	4k	32.2	49.5	24.1	64.2	76.0
CNN-ss	32-120k	55.6	69.7	31.1	76.9	86.9
CNNaug-ss	4-15k	**68.0**	**79.5**	42.3	**84.3**	**91.1**
CNN+BOW[16]	2k	-	-	-	**80.2**	-

Table 7: **The result of object retrieval on 5 datasets.** All the methods except the CNN have their representation trained on datasets similar to those they report the results on. The spatial search result on Oxford5k,Paris6k and Sculpture6k, are reported for $h_r = 4$ and $h_q = 3$. It can be seen that CNN features, when compared with low-memory footprint methods, produce consistent high results. ASMK+MA [42] and fine-vocab [26] use in order of million codebooks but with various tricks including binarization they reduce the memory foot print to 64k.

OVERVIEW

- Limitations:
  - Can not compete VLAD, SIFT for instance retrieval without 3D geometric constraints.
- Strength:
  - A good feature representation for classification and recognition problems.
  - Considering geometric constraints, works better that VLAD, SIFT for instance retrieval .

CONCLUSION

- Off-the-shelf CNN representation, OverFeat, with simple classifiers is used to address different recognition tasks.
- The learned CNN model was originally optimized for the task of object classification in ILSVRC 2013 dataset.
- It showed itself to be a strong competitor to the more sophisticated and highly tuned state-of-the-art methods.
- The same trend was observed for various recognition tasks and different datasets which highlights the effectiveness and generality of the learned representations.
- It can be concluded that, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task.

REFERENCE

1. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.
2. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, pp.1097-1105.