**CNN Features off-the-shelf: an Astounding Baseline for Recognition**

**CNN Features off-the-shelf: an Astounding Baseline for Recognition 论文笔记**

https://blog.csdn.net/weixin\_37904412/article/details/80558850

**Abstract**

Recent results indicate that the **generic descriptors extracted from the convolutional neural networks are very powerful**. This paper adds to the mounting evidence that this is indeed the case.

We report on a series of experiments conducted for different recognition tasks using the publicly available code and model of the **OverFeat network** which was trained to perform object classiﬁcation on ILSVRC13.

**We use features extracted from the OverFeat network as a generic image representation to tackle the diverse range of recognition tasks of object image classiﬁcation, scene recognition, ﬁne grained recognition, attribute detection and image retrieval applied to a diverse set of datasets.**

We selected these tasks and datasets as they gradually move further away from the original task and data the OverFeat network was trained to solve. Astonishingly, we report consistent superior results compared to the highly tuned state-of-the-art systems in all the visual classiﬁcation tasks on various datasets.

For instance retrieval it consistently outperforms low memory footprint methods except for sculptures dataset. The results are achieved using a linear SVM classiﬁer (or L2 distance in case of retrieval) applied to a feature representation of size 4096 extracted from a layer in the net.

The representations are further modiﬁed using simple augmentation techniques e.g. jittering. **The results strongly suggest that features obtained from deep learning with convolutional nets should be the primary candidate in most visual recognition tasks**.

Q: what is jittering?

https://demonstrations.wolfram.com/ImageJitterFilter/

An artistic effect can be applied to an image by replacing each pixel with a random pixel from a neighborhood of the specified radius. This is called a jitter filter in some image processing contexts.  


**1. Introduction**

*“Deep learning. How well do you think it would work for your computer vision problem?” Most likely this question has been posed in your group’s coffee room. And in response someone has quoted recent success stories [29, 15, 10] and someone else professed skepticism. You may have left the coffee room slightly dejected thinking “Pity I have neither the time, GPU programming skills nor large amount of labelled data to train my own network to quickly ﬁnd out the answer”.*

But when the convolutional neural network **OverFeat** [38] was recently made publicly available it allowed for some experimentation.

In particular **we wondered now**, **not whether one could train a deep network speciﬁcally for a given task, but if the features extracted by a deep network** - one carefully trained on the diverse ImageNet database to perform the speciﬁc task of image classiﬁcation - **could be exploited for a wide variety of vision tasks**.

We now relate our discussions and general ﬁndings because as a computer vision researcher you’ve probably had the same questions:

Prof: First off has anybody else investigated this issue?

Student: Well it turns out Donahue et al. [10], Zeiler and Fergus [48] and Oquab et al. [29] have suggested that **generic features can be extracted from large CNNs** and provided some initial evidence to support this claim. **But they have only considered a small number of visual recognition tasks**. It would be fun to more thoroughly investigate how powerful these CNN features are. How should we start?

Prof: The simplest thing we could try is to **extract an image feature vector from the OverFeat network** and **combine this with a simple linear classiﬁer**. The feature vector could just be the responses, with the image as input, from one of the network’s ﬁnal layers. For which vision tasks do you think this approach would be effective?

Student: Deﬁnitely image classiﬁcation. Several vision groups have already produced a big jump in performance from the previous sate-of-the-art methods on Pascal VOC. But maybe ﬁne-tuning the network was necessary for the jump? I’m going to try it on Pascal VOC and just to make it a little bit trickier the MIT scene dataset.

Answer: **OverFeat does a very good job even without ﬁne-tuning (section 3.2** for details).

Prof: Okay so that result conﬁrmed previous ﬁndings and is perhaps not so surprising. We asked the OverFeat features to solve a problem that they were trained to solve. And ImageNet is more-or-less a superset of Pascal VOC. Though I’m quite impressed by the indoor scene dataset result. What about a less amenable problem?(不太容易解决的问题呢)

Student: I know **ﬁne-grained classiﬁcation**. Here we want to **distinguish between sub-categories of a category such as the different species of ﬂowers**. Do you think the more generic OverFeat features have sufﬁcient representational power to pick up the potentially subtle differences between very similar classes?

Answer: It worked great on a standard bird and ﬂower database. In its most simplistic form it didn’t beat the latest best performing methods but it is a much cleaner solution with ample scope for improvement. Actually, **adopting a set of simple data augmentation techniques (still with linear SVM) beats the best performing methods.** Impressive! (Section 3.4 for details.)

Prof: Next challenge attribute detection? Let’s see **if the OverFeat features have encoded something about the semantic properties of people and objects**.

Student: Do you think the global CNN features extracted from the person’s bounding box can cope with the articulations(衔接) and occlusions present in the H3D dataset. All the best methods do some sort of **part alignment before classiﬁcation and during training**. ？？

Answer: Surprisingly the CNN features on average beat poselets and a deformable part model for the person attributes labelled in the H3D dataset. Wow, how did they do that?! They also work extremely well on the object attribute dataset. **Maybe these OverFeat features do indeed encode attribute information? (Details in section 3.5.)**

Prof: Can we push things even further? Is there a task OverFeat features should struggle with compared to more established computer vision systems? Maybe instance retrieval. This task drove the development of the SIFT and VLAD descriptors and the bag-of-visual-words approach followed swiftly afterwards. Surely these highly optimized engineered vectors and mid-level features should win hands down(轻易获胜) over the generic features?

Student: I don’t think CNN features have a chance if we start comparing to methods that also incorporate 3D geometric constraints. Let’s focus on descriptor performance. Do new school descriptors beat old school descriptors in the old school descriptors’ backyard?

Answer: Very convincing. **Ignoring systems that impose 3D geometry constraints the CNN features are very competitive on building and holiday datasets (section 4).** Furthermore, doing standard instance retrieval feature processing (i.e. PCA, whitening, renormalization) it shows superior performance compared to low memory footprint methods on all retrieval benchmarks except for the sculptures dataset.

Student: The take home message from all these results?

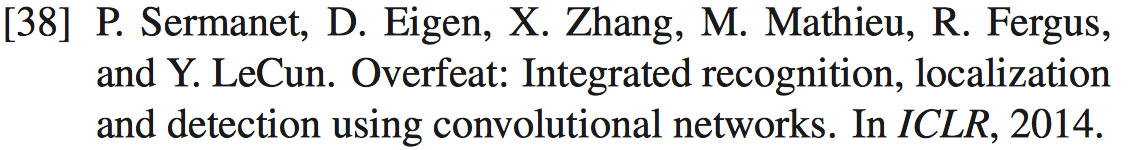
Prof: It’s all about the features! SIFT and HOG descriptors produced big performance gains a decade ago and now deep convolutional features are providing a similar breakthrough for recognition. Thus, applying the well-established computer vision procedures on CNN representations should potentially push the reported results even further. In any case, if you develop any new algorithm for a recognition task then it must be compared against the strong baseline of generic deep features + simple classiﬁer.

Student: The take home message from all these results?

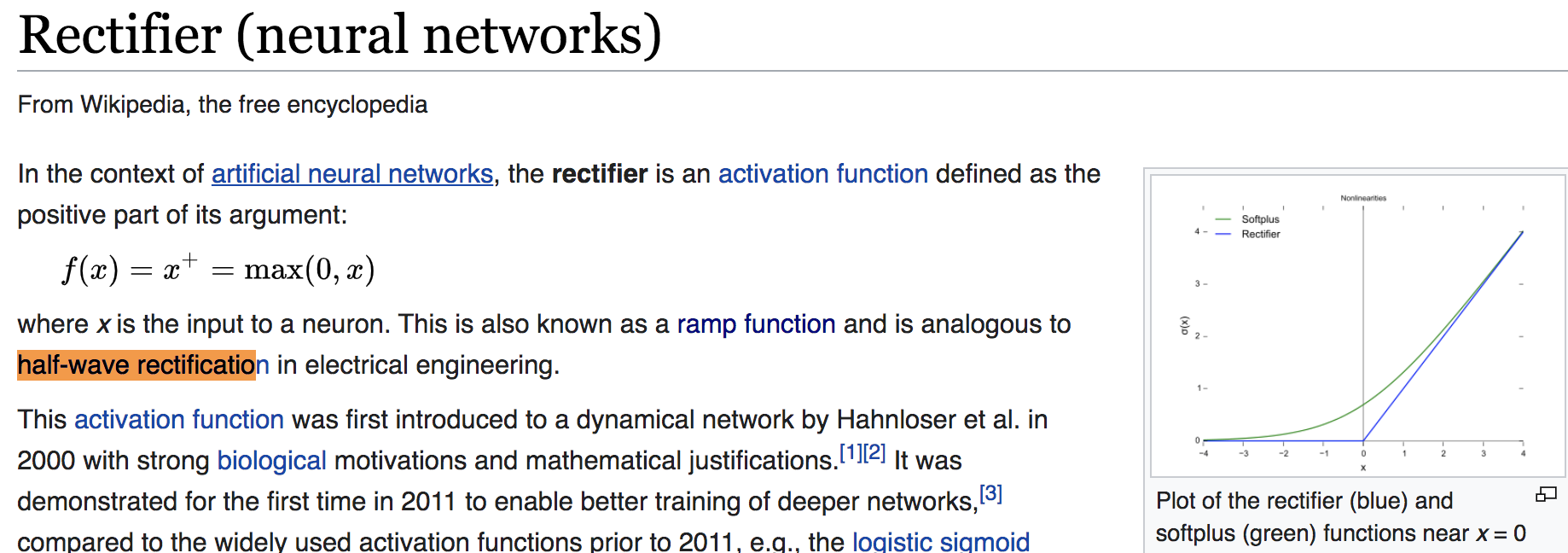
Prof: It’s all about the features! SIFT and HOG descriptors produced big performance gains a decade ago and now deep convolutional features are providing a similar breakthrough for recognition. Thus, applying the well-established computer vision procedures on CNN representations should potentially push the reported results even further. **In any case, if you develop any new algorithm for a recognition task then it must be compared against the strong baseline of generic deep features + simple classiﬁer.**

**2. Background and Outline**

In this work we use the **publicly available trained CNN called OverFeat [38]**. The structure of this network follows that of Krizhevsky et al. [22].



The convolutional layers each contain 96 to 1024 kernels of size 3×3 to 7×7. Half-wave rectiﬁcation is used as the nonlinear activation function. **Max pooling** kernels of size 3×3 and 5×5 are used at different layers to **build robustness to intra-class deformations**.



We used the “large” version of the OverFeat network. It takes as input color images of size 221×221. Please consult [38] and [22] for further details.

OverFeat was trained for the image classiﬁcation task of ImageNet ILSVRC 2013 [1] and obtained very competitive 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 [12].**

We report results on a series of experiments we conducted on different recognition tasks. **The tasks and datasets were selected such that they gradually move further away from the task the OverFeat network was trained to perform**. We have two sections for visual classiﬁcation (Sec. 3) and visual instance retrieval (Sec. 4) where we review different tasks and datasets and report the ﬁnal results.

The crucial thing to remember is that the CNN features used are trained only using ImageNet data though the simple classiﬁers are trained using images speciﬁc to the task’s dataset.

大意： CNN features used are trained only using ImageNet data；Simple classiﬁers are trained using images speciﬁc to the task’s dataset.

Finally, we have to point out that, given enough computational resources, optimizing the CNN features for speciﬁc tasks/datasets would probably boost the performance of the simplistic system even further [29, 15, 51, 43, 41].

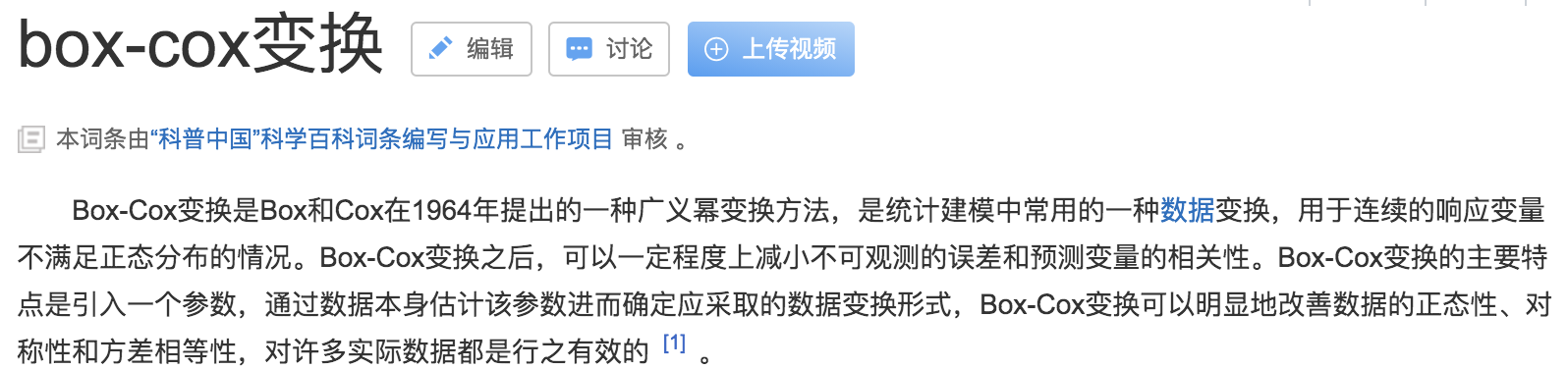
**3. Visual Classiﬁcation**

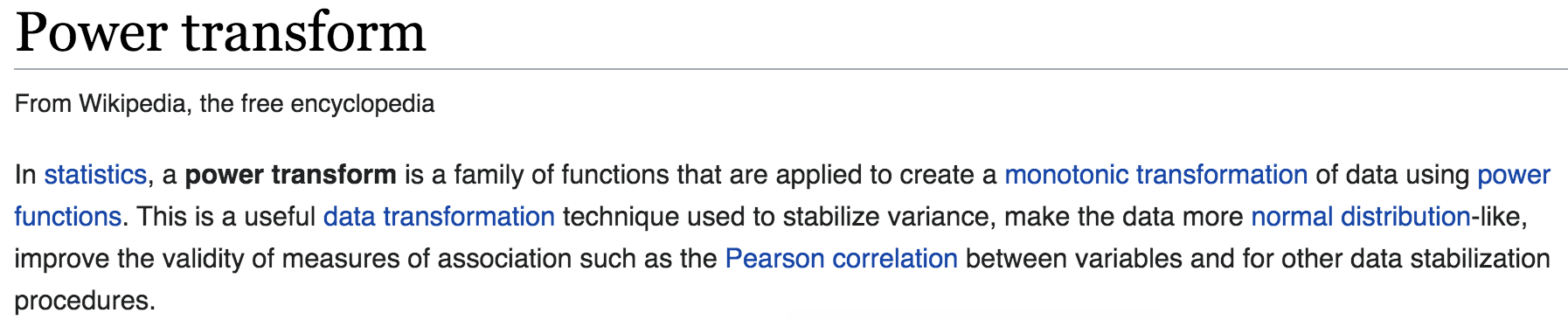
Here we go through different tasks related to visual classiﬁcation in the following subsections.

**3.1. Method**

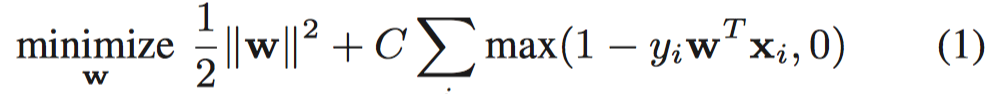
For all the experiments, unless stated otherwise, we **use the ﬁrst fully connected layer (layer 22) of the network as our feature vector**. Note the max-pooling and rectiﬁcation operations are each considered as a separate layer in OverFeat which differs from Alex Krizhevsky’s ConvNet numbering. For all the experiments we resize the whole image (or cropped sub-window) to 221×221. This **gives a vector of 4096 dimensions**. We have two settings:

* The feature vector is further **L2 normalized** to unit length for all the experiments. We use the 4096 dimensional feature vector in combination with a Support Vector Machine (SVM) to solve different classiﬁcation tasks (CNN-SVM).
* We further augment the training set by adding cropped and rotated samples and doing componentwise power transform and report separate results (CNNaug+SVM).

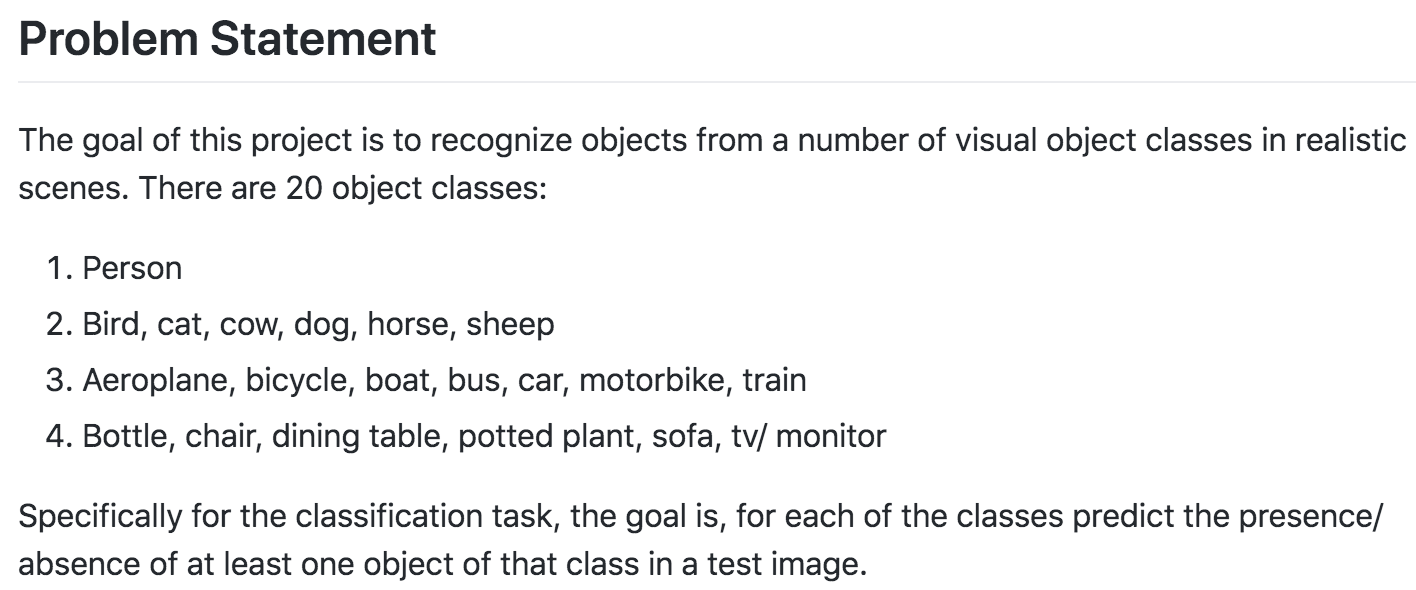




For the classiﬁcation scenarios where the labels are not mutually exclusive (e.g. **VOC Object Classiﬁcation or UIUC Object attributes**) we **use a one-against-all strategy**, in the rest of experiments we use **one-against-one linear SVMs with voting.** For all the experiments we use a linear SVM found from eq.1, where we have training data {(x i , y i )}.



**VOC**



Further information can be found in the implementation details at section 3.6.

**3.2. Image Classiﬁcation**

To begin, we adopt the CNN representation to tackle the problem of image classiﬁcation of objects and scenes. The system should assign (potentially multiple) semantic labels to an image. **Remember in contrast to object detection, object image classiﬁcation requires no localization of the objects.** The CNN representation has been optimized for the object image classiﬁcation task of ILSVRC. Therefore, in this experiment the representation is more aligned with the ﬁnal task than the rest of experiments. **However, we have chosen two different image classiﬁcation datasets, objects and indoor scenes, whose image distributions differ from that of ILSVRC dataset**.

**3.2.1 Datasets**

We use two challenging recognition datasets, Namely, Pascal VOC 2007 for object image classiﬁcation [12] and the MIT-67 indoor scenes [36] for scene recognition.

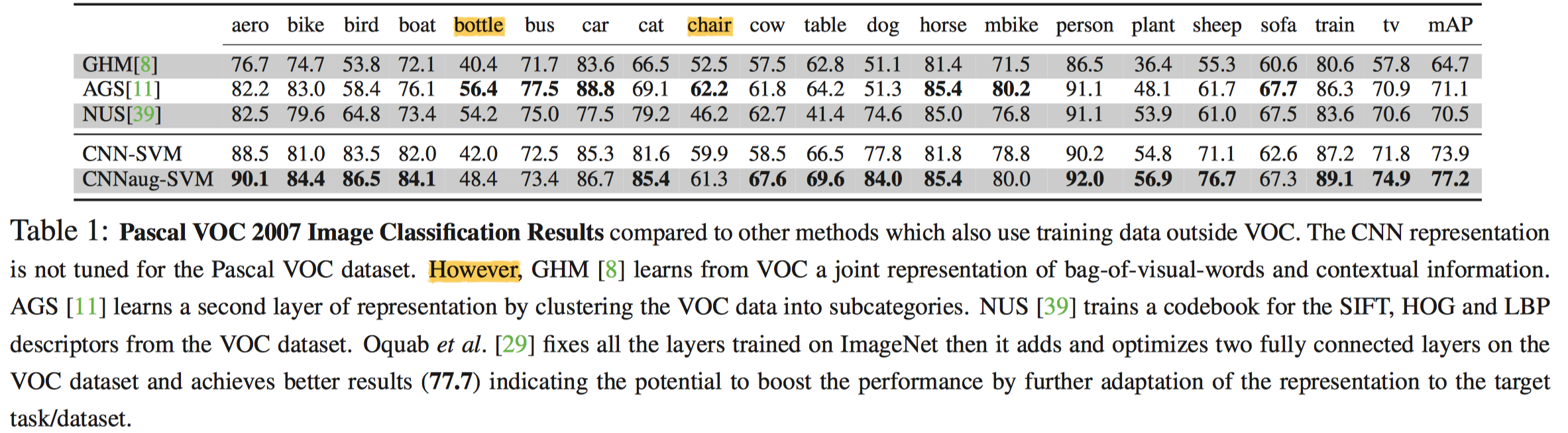
**Pascal VOC**. Pascal VOC 2007 [12] contains ∼10000 images of 20 classes including animals, handmade and natural objects. The **objects are not centered** and in general **the appearance of objects in VOC is perceived to be more challenging than ILSVRC.** Pascal VOC images come with bounding box annotation which are not used in our experiments.

**MIT-67 indoor scenes**. The MIT scenes dataset has 15620 images of 67 indoor scene classes. The dataset consists of different types of stores (e.g. bakery, grocery) residential rooms (e.g. nursery room, bedroom), public spaces (e.g. inside bus, library, prison cell), leisure places (e.g. buffet, fastfood, bar, movie theater) and working places (e.g. ofﬁce, operating room, tv studio). **The similarity of the objects present in different indoor scenes makes MIT indoor an especially difﬁcult dataset compared to outdoor scene datasets**.

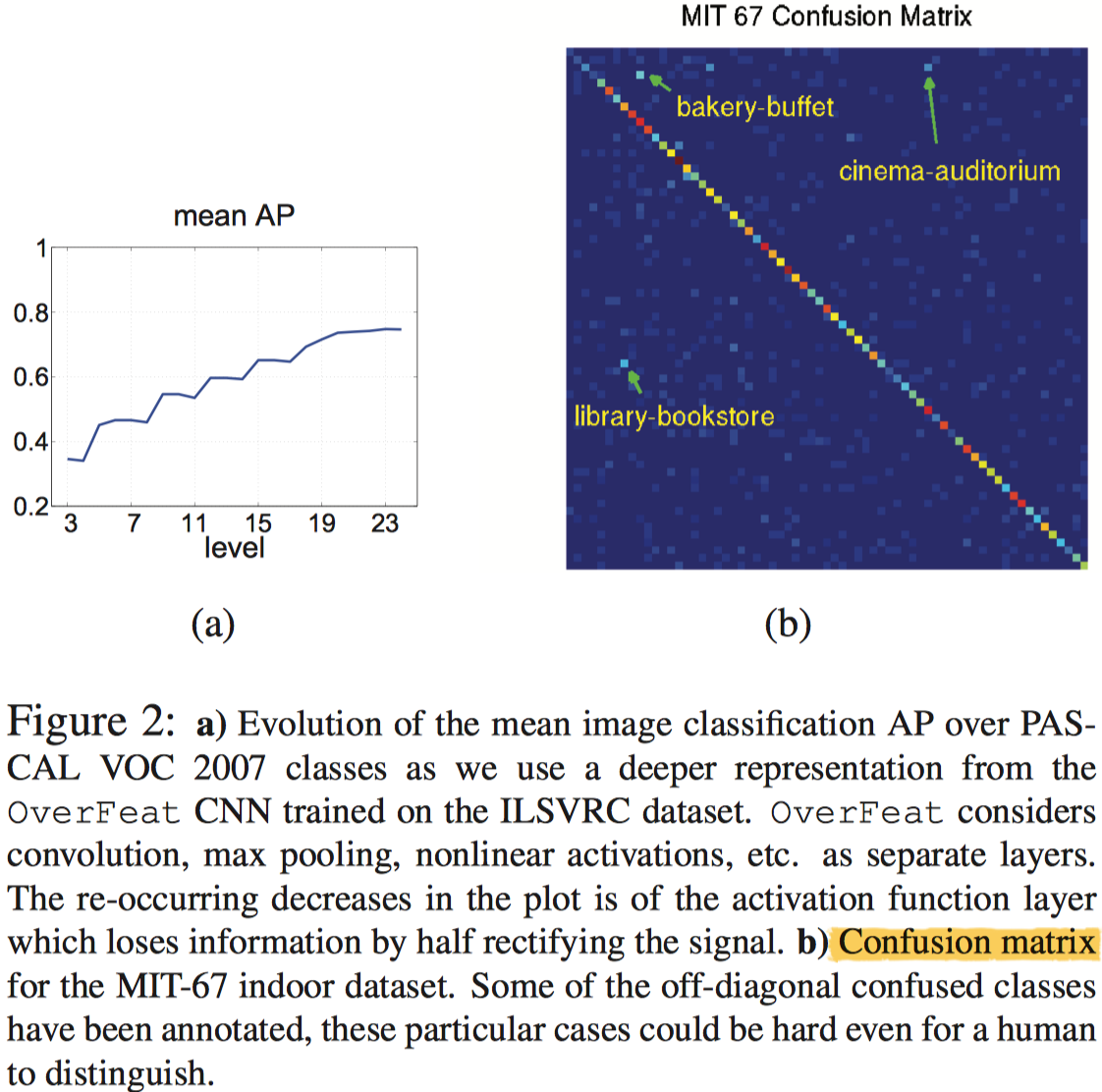
**3.2.2 Results of PASCAL VOC Object Classiﬁcation**

Table 1 shows the results of the OverFeat CNN representation for object image classiﬁcation. The performance is measured using average precision (AP) criterion of VOC 2007 [12].

**Since the original representation has been trained for the same task (on ILSVRC) we expect the results to be relatively high**. We **compare the results only with those methods which have used training data outside the standard Pascal VOC 2007 dataset**. We can see that the method outperforms all the previous efforts by a signiﬁcant margin in mean average precision (mAP). Furthermore, it has superior average precision on 10 out of 20 classes. **It is worth mentioning the baselines in Table 1 use sophisticated matching systems**. The same observation has been recently made in another work [29].



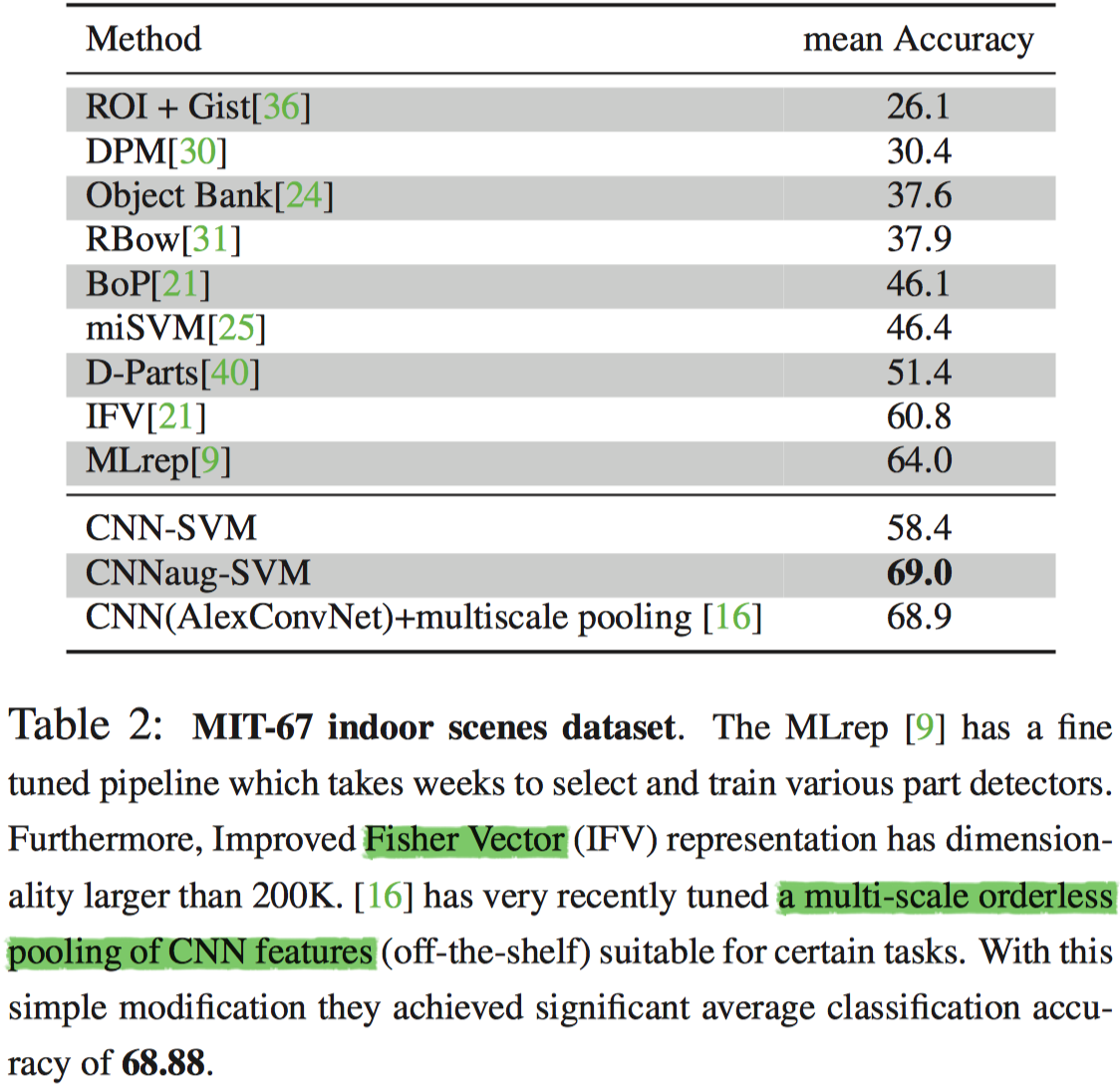
**Different layers.** **Intuitively** one **could reason** that the **learnt weights for the deeper layers could become more speciﬁc to the images** of the training dataset **and the task** it is trained for. Thus, **one could imagine the optimal representation for each problem lies at an intermediate level** of the network. To further study this, we trained a linear SVM for all classes using the output of each network layer. The result is shown in Figure 2a. Except for the fully connected last 2 layers the performance increases. We observed the same trend in the individual class plots. **The subtle drops in the mid layers (e.g. 4, 8, etc.) is due to the “ReLU” layer which half-rectiﬁes the signals. Although this will help the non-linearity of the trained model in the CNN, it does not help if immediately used for classiﬁcation.**



**3.2.3 Results of MIT 67 Scene 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 representation with linear SVMs training signiﬁcantly outperforms a majority of the baselines. The non-CNN baselines beneﬁt from a broad range of sophisticated designs. confusion matrix of the CNN-SVM classiﬁer on the 67 MIT classes. It has a strong diagonal. The few relatively bright off-diagonal points are annotated with their ground truth and estimated labels. One can see that in these examples the two labels could be challenging even for a human to distinguish between, especially for close-up views of the scenes.



**3.3. Object Detection**

Unfortunately, we have not conducted any experiments for using CNN off-the-shelf features for the task of object detection. But it is worth mentioning that Girshick et al. [15] have reported remarkable numbers on PASCAL VOC 2007 using off-the-shelf features from Caffe code. We repeat their relevant results here. 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 our evidences of how powerful the CNN features off-the-shelf are for visual recognition tasks.

Finally, by further ﬁne-tuning the representation for PASCAL VOC 2007 dataset (not off-the-shelf anymore) they achieve impressive results of 53.1.

**3.4. Fine grained Recognition**

Fine grained recognition has recently become popular due to its huge potential for both commercial and cataloging applications. **Fine grained recognition is specially interesting because it involves recognizing subclasses of the same object class such as different bird species, dog breeds, ﬂower types, etc.** The advent of many new datasets with ﬁne-grained annotations such as Oxford ﬂowers [27], Caltech bird species [45], dog breeds [1], cooking activities [37], cats and dogs [32] has helped the ﬁeld develop quickly. **The subtlety of differences across different subordinate classes (as opposed to different categories) requires a ﬁne-detailed representation. This characteristic makes ﬁne-grained recognition a good test of whether a generic representation can capture these subtle details.**

大意就是 通用表示的方法好不好使，得用Fine grained Recognition来判定

**3.4.1 Datasets**

We evaluate CNN features on two ﬁne-grained recognition datasets CUB 200-2011 and 102 Flowers.

**Caltech-UCSD Birds (CUB) 200-2011** dataset [45] is chosen since many recent methods have reported performance on it. 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 which are sometimes even hard for humans to distinguish.**

Multiple levels of annotation are available for this dataset - bird bounding boxes, 15 part landmarks, 312 binary attributes and boundary segmentation.

**The majority of the methods applied use the bounding box and part landmarks for training. In this work we only use the bounding box annotation during training and testing.**

**Oxford 102 ﬂowers** dataset [27] contains 102 categories. Each category contains 40 to 258 of images. **The ﬂowers appear at different scales, pose and lighting conditions.** **Furthermore, the dataset provides segmentation for all the images.**

**3.4.2 Results**

Table 3 reports the results of the CNN-SVM compared to the top performing baselines on the CUB 200-2011 dataset. **The ﬁrst two entries of the table represent the methods which only use bounding box annotations.** The rest of baselines use part annotations for training and sometimes for evaluation as well.

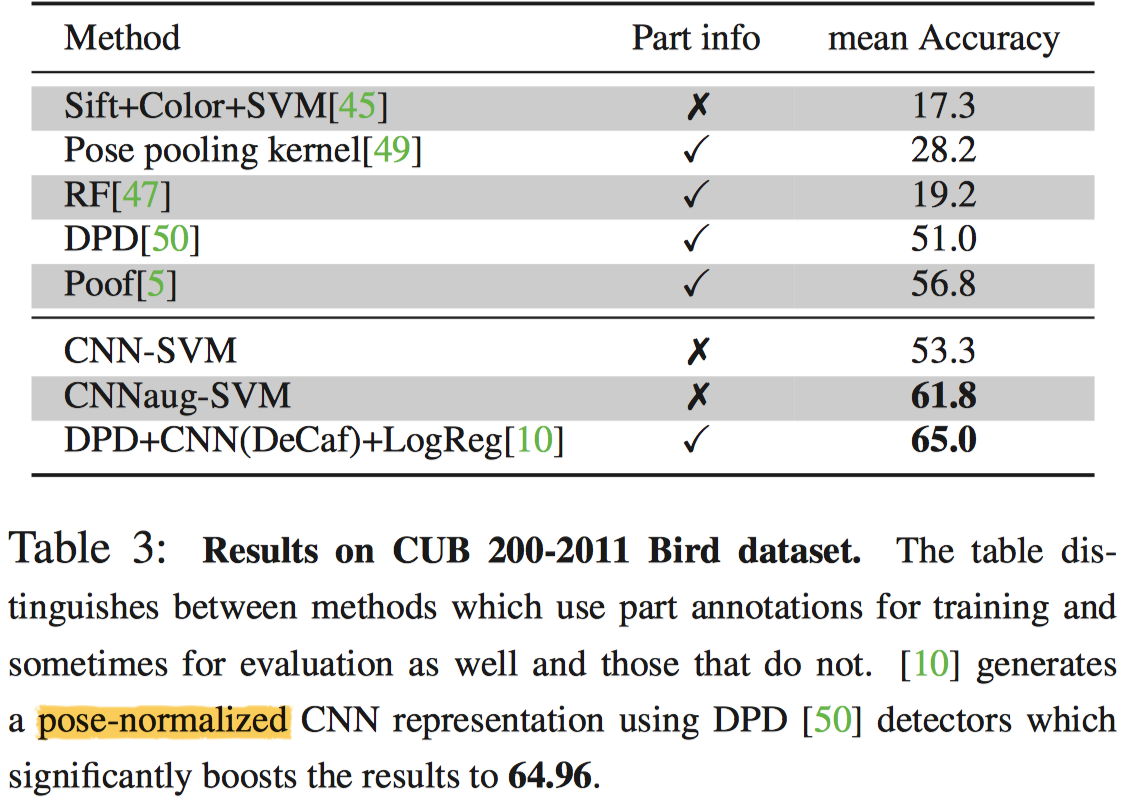
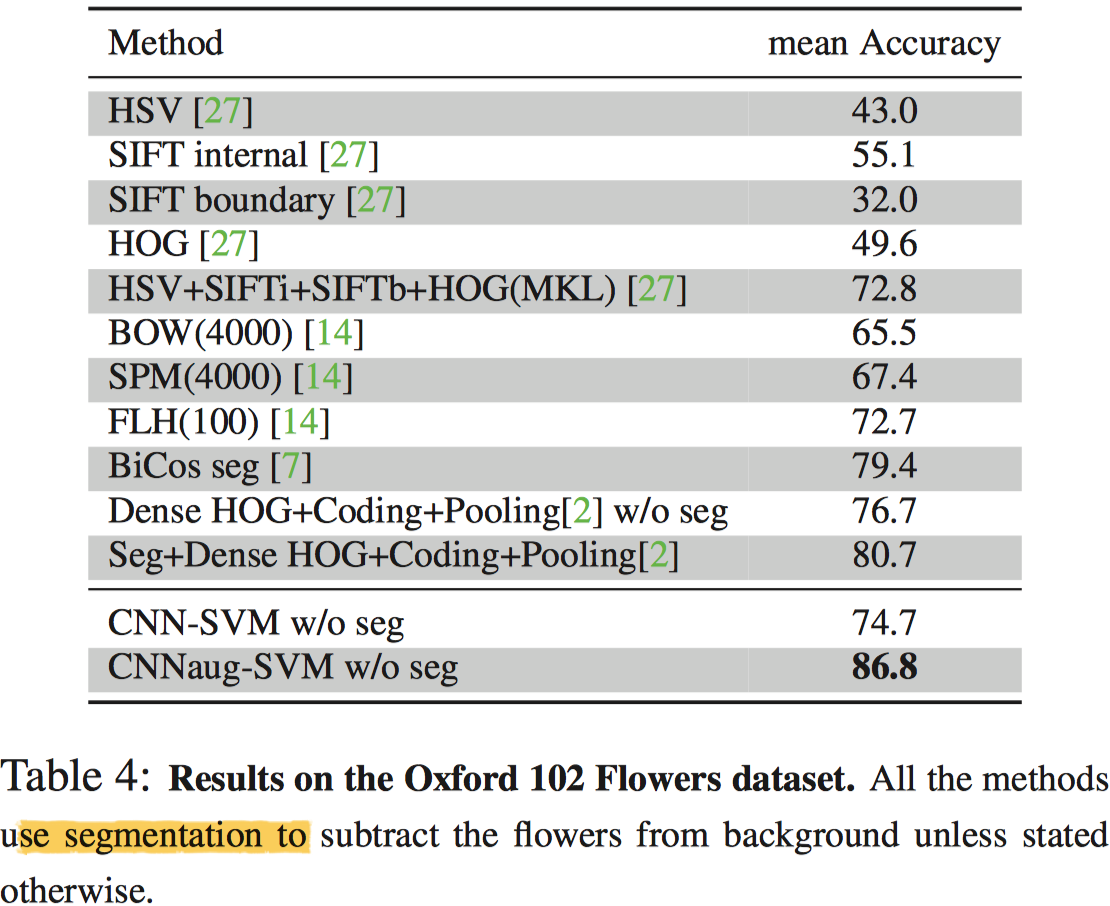


Table 4 shows the performance of CNN-SVM and other baselines on the ﬂowers dataset. All methods, bar the CNNSVM, use the segmentation of the ﬂower from the background. It can be seen that CNN-SVM outperforms all basic representations and their multiple kernel combination even without using segmentation.

****

**3.5. Attribute Detection**

An **attribute** within the context of computer vision is **deﬁned as some semantic or abstract quality which different instances/categories share**.

计算机视觉中的属性被定义为不同的实例/类别共享的一些语义或抽象的特征。

**3.5.1 Datasets**

We use two datasets for attribute detection. The ﬁrst dataset is the **UIUC 64** object attributes dataset [13]. There are **3 categories of attributes in this dataset**: **shape** (e.g. is **2D boxy**), **part** (e.g. **has head**) or **material** (e.g. is **furry**).

有三类属性：形状（例如是二维的），部位（如有头部），材质（例如毛茸茸的）

The second dataset is the **H3D dataset** [6] which deﬁnes **9 attributes** for a subset of the person images from Pascal VOC 2007. The attributes range from “**has glasses**” to **“is male”**.

**3.5.2 Results**

Table 5 compares CNN features performance to state-of-the-art. Results are reported for both across and within categories attribute detection (refer to [13] for details).

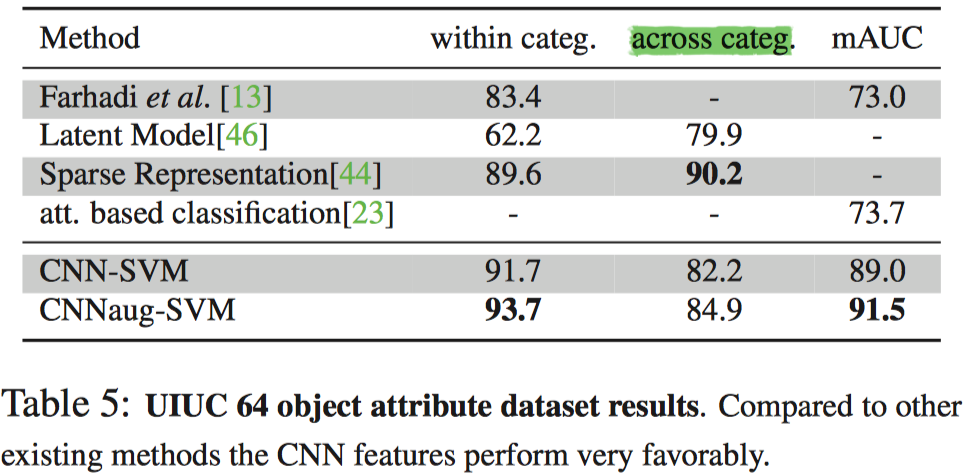
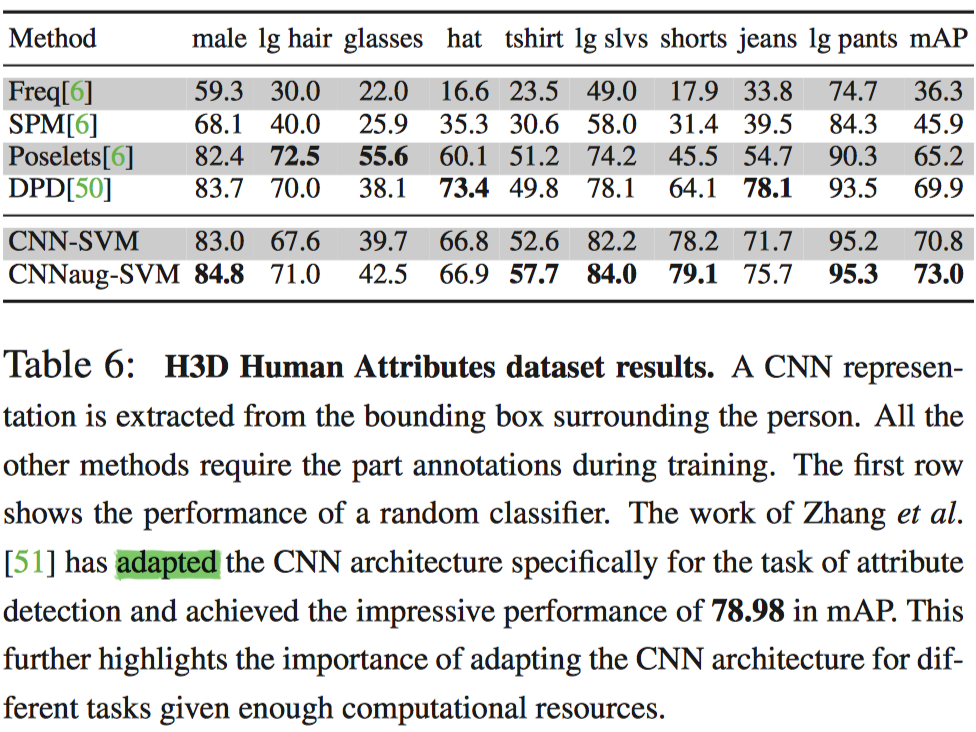


Table 6 reports the results of the detection of 9 human attributes on the H3D dataset including poselets and DPD [50]. **Both poselets and DPD use part-level annotations during training** while **for the CNN we only extract one feature from the bounding box around the person**. The CNN representation performs as well as DPD and signiﬁcantly outperforms poselets.



**3.6. Implementation Details**

We have used precomputed linear kernels with libsvm for the CNN-SVM experiments and liblinear for the CNNaugSVM 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). The cropping is done such that the subwindow contains 4/9 of the original image area from the 4 corners and the center. We noted the following phenomenon for all datasets. At the test time, when we have multiple representations for a test image, taking the sum over all the responses works outperforms taking the max response. In CNNaug-SVM we use signed component-wise power transform by raising each dimension to the power of 2. For the datasets which with bounding box (i.e. birds, H3D) we enlarged the bounding box by 150% to include some context. In the early stages of our experiments we noticed that using one-vs-one approach works better than structured SVM for multi-class learning. Finally, we noticed that using the imagemagick library for image resizing has slight adverse effects compared to matlab imresize function. The crossvalidated SVM parameter (C) used for different datasets are as follows. VOC2007:0.2, MIT67:2 , Birds:2, Flowers:2, H3D:0.2 UIUCatt:0.2. 2

**4. Visual Instance Retrieval**

In this section we **compare the CNN representation to the current state-of-the-art retrieval pipelines** including VLAD[4, 52], BoW, IFV[33], Hamming Embedding[17] and BoB[3]. **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, **we only report results on representations with relevant order of dimensions** and **exclude post-processing methods like spatial re-ranking and query expansion**.

**4.1. Datasets**

We report retrieval results on **ﬁve common datasets** in the area as follows:

**Oxford5k** buildings[34] This is a collection of 5063 reference photos gathered from ﬂickr, and 55 queries of different buildings. **From an architectural standpoint the buildings in Oxford5k are very similar. Therefore it is a challenging benchmark for generic features such as CNN.**

从建筑角度来看，Oxford5k中的建筑物非常相似。 因此，对于CNN等通用功能而言，这是具有挑战性的基准

**Paris6k** buildings[35] Similar to the Oxford5k, this collection has 55 queries images of buildings and monuments from Paris and 6412 reference photos. **The landmarks in Paris6k have more diversity than those in Oxford5k.**

**Sculptures6k**[3] This **dataset brings the challenge of smooth and texture-less item retrieval.** It has 70 query images and contains 6340 reference images which is halved to train/test subsets. **The results on this dataset highlights the extent to which CNN features are able to encode shape.**

该数据集上的结果突出显示了CNN特征能够编码形状的程度

**Holidays** dataset[19] This dataset contains 1491 images of which 500 are queries. **It contains images of different scenes, items and monuments**. Unlike the ﬁrst three datasets, it **exhibits a diverse set of images**. For the above datasets we reported mAP as the measurement metric.

**UKbench**[28] **A dataset of images of 2250 items each from four different viewpoints**. The UKbench provides a **good benchmark for viewpoint changes**. We reported recall at top four as the performance over UKBench.

**4.2. Method**

Similar to the previous tasks we use the L2 normalized output of the ﬁrst fully connected layer as representation.

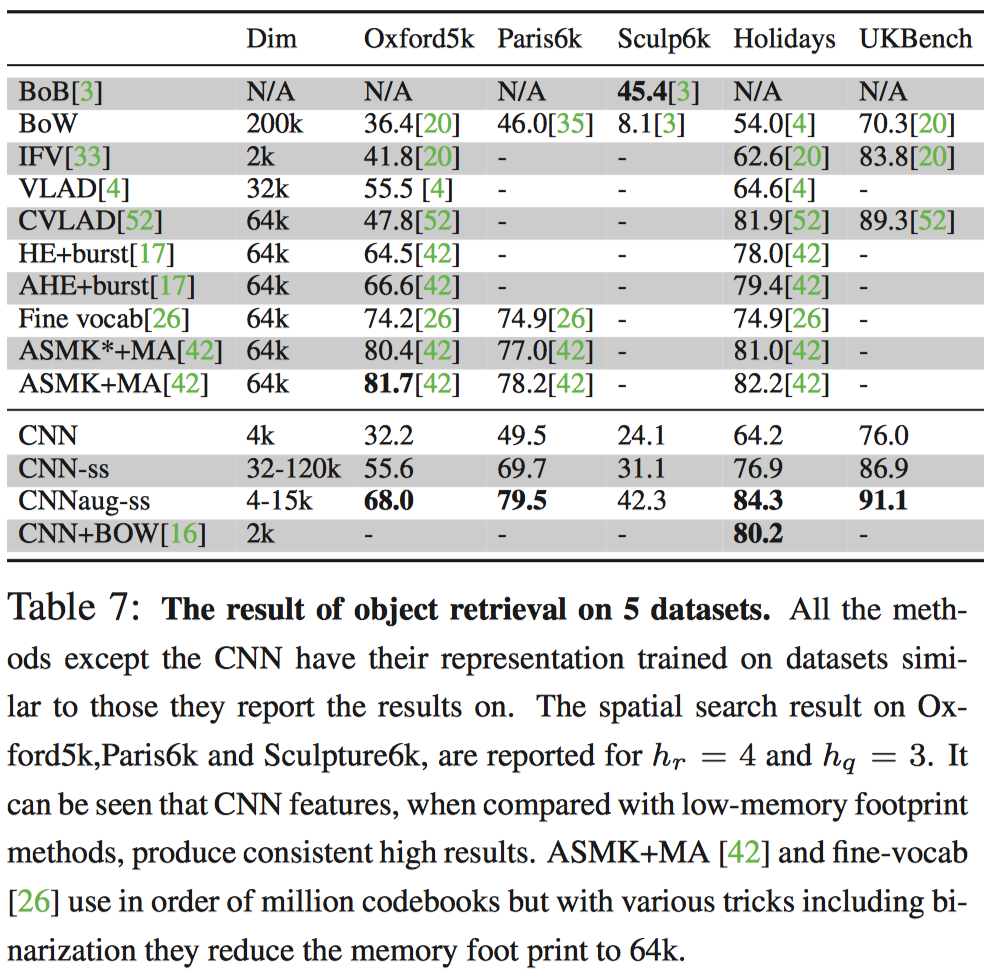
**Spatial search**. The items of interest can appear at different locations and scales in the test and reference images making some form of spatial search necessary. Our crude search has the following form. For each image we extract multiple sub-patches of different sizes at different locations. Let h (the number of levels) represent the number of different sized patches we extract. At level i, 1 ≤ i ≤ h, we extract i 2 overlapping sub-patches of the same size whose union covers the whole image. For each extracted sub-patch we compute its CNN representation. 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. Then, the distance between the reference and the query image is set to the average distance of each query sub-patch to the reference image. In contrast to visual classiﬁcation pipelines, we extract features from the smallest square containing the region of interest (as opposed to resizing). In the reset of the text, h r denotes to the number of levels for the reference image and similarly h q for the query image.

**Feature Augmentation**. Successful instance retrieval methods have many feature processing steps. Adopting the proposed pipeline of [18] and followed by others [16, 42] we process the extracted 4096 dim features in the following way: L2 normalize → PCA dimensionality reduction → whitening → L2 renormalization. Finally, we further use a signed component wise power transform and raise each dimension of the feature vector to the power of 2. For all datasets in the PCA step we reduce the dimensionality of the feature vector to 500. All the L2 normalizations are applied to achieve unit length.

**4.3. Results**

The result of different retrieval methods applied to 5 datasets are in table 7.

**Spatial search is only used for the ﬁrst three datasets which have samples in different scales and locations.** For the other two datasets we used the same jittering as explained in Sec. 3.1



It should be emphasized that we only reported the results on low memory footprint methods.

**5. Conclusion**

In this work, we used an off-the-shelf CNN representation, OverFeat, with simple classiﬁers to address different recognition tasks. The learned CNN model was originally optimized for the task of object classiﬁcation in ILSVRC 2013 dataset. Nevertheless, 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. The experiments conﬁrm and extend the results reported in [10]. We have also pointed to the results from works which speciﬁcally optimize the CNN representations for different tasks/datasets achieving even superior results. Thus, it can be concluded that from now on, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task.

在这项工作中，我们使用OverFeat CNN特征向量结合一个简单的分类器来解决不同的识别任务。 CNN模型最初是在ILSVRC 2013数据集中对图像分类任务进行了训练。 CNN提取的特征向量，表明它将成为更先进方法的强有力竞争者。 对于各种识别任务和不同的数据集观察到相同的趋势，这强调了学习的CN表示具备的有效性和一般性。 因此，CNN通过深度学习获得的特征应该是大多数视觉识别任务的主要选择。