Scene Classification using Approximate Nearest Neighbors

Zhiyuan Wang

Advisor: Prof. Ying Wu

February 1, 2015

1 Introduction

With the goal to make machine vision have the ability to understand the real world as good as human, scene understanding is one of the most fundamental and challenging problems of high level computer vision. It is a crucial part or potentially useful for many applications, to name just a few, content-based image and video retrieval, human computer interaction, autonomous driving, vision-based robot navigation, visual surveillance. (Apple Inc. is searching for scientist/engineer with knowledge on scene understanding right now, but I don't know what they are doing ...) Scene understanding is also a complex task. Image segmentation, object detection, object recognition, 3D vision, semantic analysis and even ontology-driven natural language processing, all of them can be a subtask of scene understanding. In literature, the research on scene understanding begin from some high level computer vision tasks like image segmentation, scene classification [19, 30], to a combination of computer vision, cognitive science [29, 20], context and concept modeling, semantic analysis and summarization [33, 11] which involved with some natural language processing methods, on a higher level. L. Li et al. proposed a hierarchical generative model within a probabilistic framework using tags and context to do automatic classification, annotation and segmentation. [11] Recently since the widely usage of depth camera, 3D scene understanding also become an emerging research topic, especially for indoor scene. [12, 5, 15] 3D scene understnding is slightly different than 2D scene, with more focus on geometric understanding, object localization and the reasoning behind them. [5] used YouTube videos to learn scene semantics from long-term observation of people. They associate human pose with scene objects, and using object recognition to imporove prediction of human pose in video. In [12], D. Lin jointly solved scene classification and 3D object recognition by a conditional random field integrating both geometric and semantic context with global features (scene appearance, ranking potential, etc.) of generated candidate cuboids. N. Silberman et al. proposed a semantic segmentation models can better identify individual instances of the same classes by introducing a new higher-order loss function that directly minimizes the coverage metric and evaluate a variety of region features, including those from a convolutional network. [15] (see more at ECCV 2014: Tutorial on 3D Scene Understanding)

Human is capable to recognize a scene by a glance, and gain an insight in a short time. The rapid improvement on multi-class object recognition also make real time scene understanding become possible. [9] proposed a markov decision method doing online feature selection, cost-sensitive dynamic feature selection in their paper, to optimize any-time object and scene recognition.

Scene classification is a step stone toward total scene understanding. It gives us a categrization of scene or tags of an image like what we get from a glance, thus can be used as a global feature or prior in some higher-level and coarse-to-fine problems. [19] presented GIST descriptor computes a wavelet image decomposition, which represents an image by the output of Gabor-like filters with tuned to different orientation and scales. [10] improved bag-of-features method to predict scene categarization by using spatial pyramid which considers the object layout of an image. The most related work to our project is [30]. In [30], Xiao et al. investigated several different features' performance for scene classificaiton, and eventually combine them together to train a kernel SVM classifier achieved the best accuracy on their proposed large scale datasets.

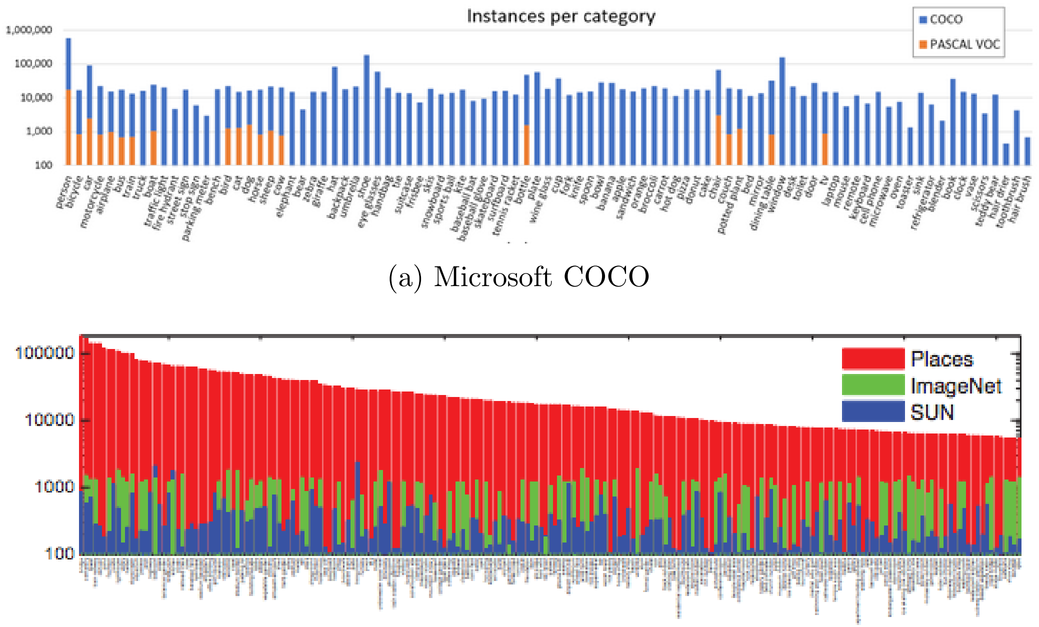
2 Large Scale Scene Understanding Datasets

Large scale image datasets play a crucial role on the development of mult-class object recognition toward a generalized computer vision system. Back to year 2004 tackling object recognition on databases such Caltech 101 is still a challenging task. (wow, it is almost 10 years ago) But today collecting data is much easier, due to the growth of online photo community like Flickr. It is not surprising, ImageNet now has 14,197,122 images in tens of thousand classes, and the number will still be increasing. Thanks to crowsourcing, making annotation is also easier and feasible for large scale image datasets. [21] Some researchers even proposed some attribute-based approaches [22, 7], which heavily depend on crowsourcing with human in the loop of object recognition. (I personally do not quite into them, simply don't) Another approach think post-processing is not necessary, no annotation is needed. Tiny image dataset [27], a scene recognition dataset, contains 80M 32×32 images in 75k classes.

In [30], J. Xiao et al. performed experiments on 15 classes scene database [10] and introduced SUN (Scene UNderstanding) dataset with 397 categories containing more than 100 images per class. Recently, B. Zhou et al. introduced a new scene-centric image dataset Place with a standard CNN architecture, which is 60 times larger than SUN database. T. Lin et al. create Microsfot COCO database with a focus on common object in context. [28] Similar as Places, they managed to provide a database with each class distributed more evenly.

To study semantic scene meaning's relation with visual information, C. Zitnick et al. proposed a way to synthesis Abstract Images with objects and background. [33]

NYUv2 [16] is a RGBD image dataset initially created for 3D indoor



(a) Microsoft COCO

(b) Places

scene segmentation. It contains 1449 densely labeled pairs of aligned RGB and depth images, 464 scenes of 3 cities, and 407,024 unlabeled frames.

We use the same dataset as [30].

3 Method

3.1 Approximate Nearest Neighbor Search

As many experiments in computer vision and machine learning suggest using large scale database is the key to obtain a good performance on real life problem. [6] Due to the growing size of large scale database, scalability should be considered carefully. Particularly in computer vision, most features are high-dimensional, e.g. a color 32×32 tiny image has 3072 dimensions, that becomes more important. Scalability is simply a limitation of traditional nearest neighbor searching algorithm. Since we already get a boost due to the size of the training set, an approximate algorithm with a little loss of accuarcy but still get a close result and a remarkable speed-up is an appropriate choice.

M. Mujia et al. built FLANN (Fast Library for Approximate Nearest Neighbors) which integrates many approximate nearest neighbor search algorithms. [13] The following subsections are summarized from [14]. (but I will concentrate all the relevant contents(description, empirical result, etc.) in each section.)

3.1.1 Partitioning Trees

The kd-tree is one of the best known nearest neighbor algorithm. It is very effective in low dimensionality spaces, but its performance decreases quickly for high dimensional data. Thus some variations of kd trees is proposed to deal with this problem. They can be roughly classified into four classes.

"Error bound" approach. Like other common approximate algorithms, "error bound" approach search by considering $(1 + \epsilon)$-approximate nearest neighbors. A priority queue is used to speed up the search.

"Time bound" approach. These approach approximating the nearest neighbor by limiting the time spent during the search, where the k-d tree search is stopped early after examining a fixed number of leaf nodes, and in practice it has been found to give better results than the error-constrained algorithm. (strange, a little bit, but it's empirical result)

Multiple randomized k-d trees. In a wide range of comparisons, multiple randomized trees ar among one of the most effective methods for matching high dimensional data.

Hyperlane k-d trees using non-axis-aligned partitioning, including PCA-tree, RP-tree, etc. As the overhead of evaluating multiple dimensions, these approaches are not more efficient than a randomized k-d tree decomposition.

K-d trees on decomposed subspace. These approaches decompose the space into several small subsapce by various clustering algorithms instead of using hyperplanes, including hierarchical k-means, vp-tree, etc. (Again, we found the overhead of building such decomposition still can be expensive on some database.)

3.1.2 Hashing Based Nearest Neighbor Techniques

One of the trick to make algorithm scalable is using hashing-based method, and recently hash has also been introduced to speed up many computer vision methods. [4, 31] Probably the most common known one is locality sensitive ashing (LSH). The performance of hashing methods is highly dependent on the quality of the hashing functions they use and large body of research has been targeted at improving hashing methods by using data-dependent hashing functions computed using various learning techniques, such as spectral hashing, parameter sensitive hashing, kernelized LSH, etc. Different LSH algorithms provide theoretical guarantees on the search quality, and have been sucessfully used in many projects, however in M. Mujia's expereiments they are often outperformed by randomized k-d trees and the priority search k-means trees.

3.1.3 Nearest Neighbor Graph Techniques

The graph methods build a graph structure in which points are vertices and edges connect each point to its nearest neighbors. The query points are used to explore this graph using various strategies, such as start from multiple well separated seeds, best-first search, in order to get closer to their nearest neighbors. The construction time of K-NN graph stucture should also be considered. (Some peopel argue that KGrpah library is faster than FLANN. I haven't got time to test)

3.1.4 Automatic Configuration of NN Algorithms

We can formulate the automatic configuration as a procedure to choose an NN algorithm A with parameter θ in a certain parameter space Θ, which minimizes a cost function $c(\theta)$

$$c(\theta) = \frac{s(\theta) + w_b b(\theta)}{\min_{\theta \in \Theta}(s(\theta) + w_b b(\theta))} + w_m m(\theta), \tag{1}$$

where $s(\theta)$, $b(\theta)$ and $m(\theta)$ represent the search time, tree build time and memory overhead for the tree(s) constructed and queried with parameters θ.

The weights w_b and w_m are used to control the relative importance of the build time and memory overhead.

The above optimization is performed in two steps: a global exploration of the parameter space using grid seach followed by a local optimization. In the second step implement further locally explore the parameter space and fine-tune the best solution using Nelder-Mead downhill simplex method.

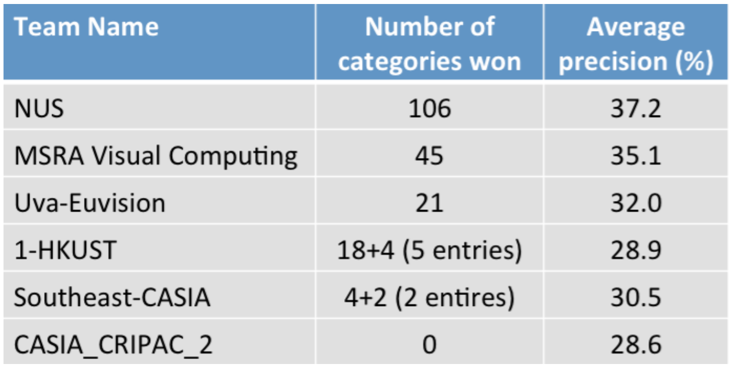
FLANN uses random sub-sampling cross-validation to generate the data and query points when run the optimization.

3.1.5 Scalable Nearest Neighbor Search

M. Mujia et al. also point out that many papers show good performance using simple non-parametric methods in conjunction with large scale databases, but one problem is that it is hard to load the data into memory, so they proposed a distributed nearest neighbors algorithm [14], but this is out of my report's scope.

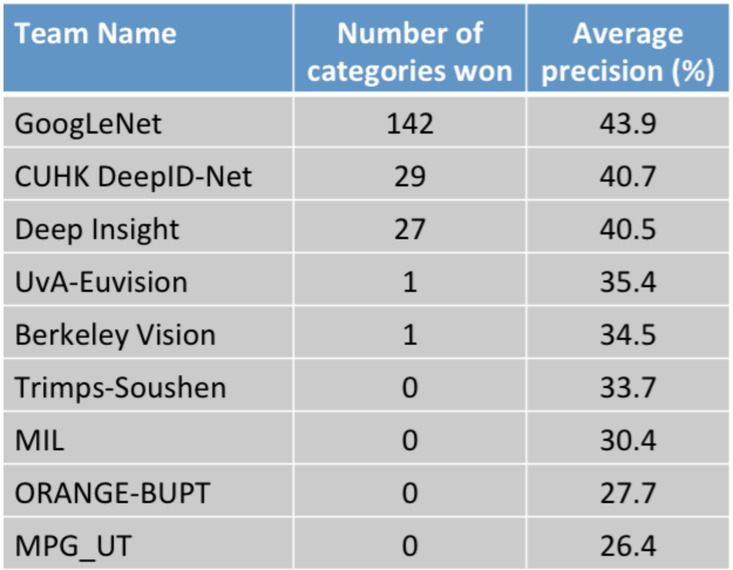
3.2 Convolutional Neural Network

In recent years, convolutional neural network model has made an impressive breakthrough in various vision tasks. Some researcher called the year 2012 a turning point [17] (not because 21 December). As fig. 2 shows, most of the top entries use CNN-based model. Google's team propose a multibox network approach which futher improve their model's mAP to 55.7. [2] (they claim that they expect the highest quality proposal generation method to be



Team Name	Number of categories won	Average precision (%)
NUS	106	37.2
MSRA Visual Computing	45	35.1
Uva-Euvision	21	32.0
1-HKUST	18+4 (5 entries)	28.9
Southeast-CASIA	4+2 (2 entires)	30.5
CASIA_CRIPAC_2	0	28.6

(a) Result with provided training data only.



Team Name	Number of categories won	Average precision (%)
GoogLeNet	142	43.9
CUHK DeepID-Net	29	40.7
Deep Insight	27	40.5
UvA-Euvision	1	35.4
Berkeley Vision	1	34.5
Trimps-Soushen	0	33.7
MIL	0	30.4
ORANGE-BUPT	0	27.7
MPG_UT	0	26.4

(b) Result of entries with additional data

Figure 2: ILSVRC Object Detection Result [18]

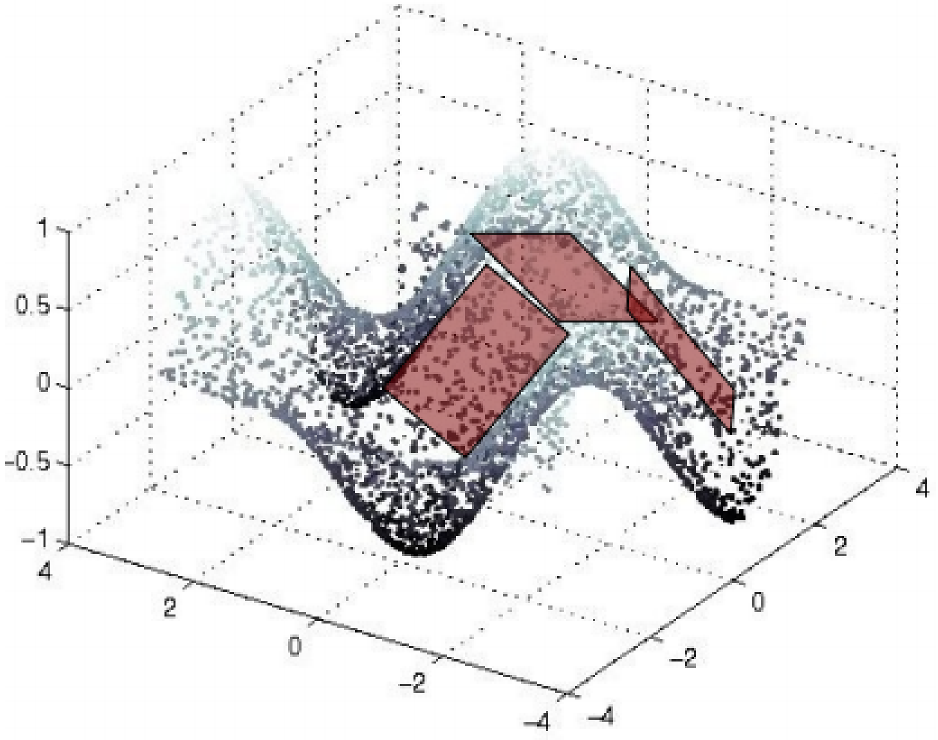


Figure 3: ReLu layers is a complex combination of piece-wisth linear tilings [23]

learned from scratch as well, though they also admit some recent success of novel sophisticated proposal generation methods). Convolution, sharing parameter and recently abondon fully-connected layer make CNN fast, though training a CNN model on large datasets (like 2 million images) still need days or weeks [32].

CNN is just a model, and ”deep” (non linear mapping between layers) make it have more express power. While structure of deep neural network is still under intense research, some state-of-the-art model simply use linear

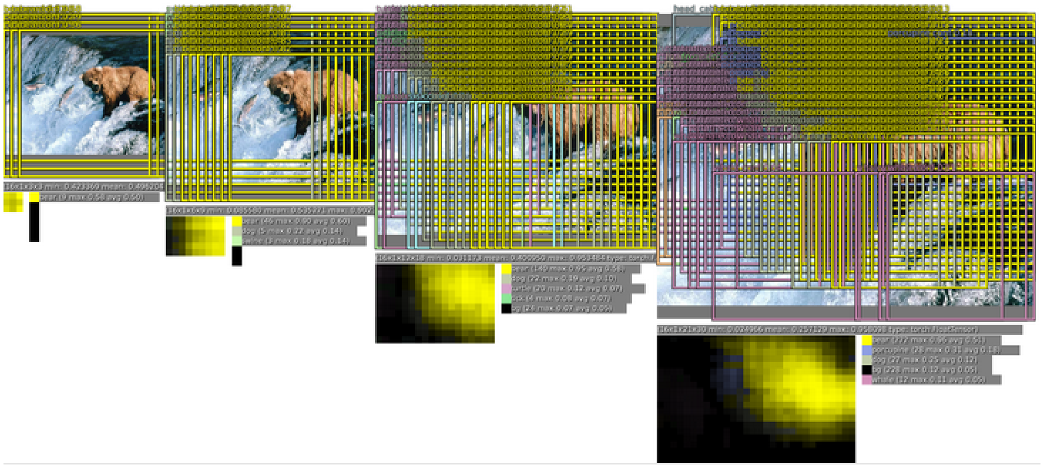

bear (9 max 0.58 avg 0.50)
bear (46 max 0.90 avg 0.60)
dog (5 max 0.22 avg 0.14)
swine (3 max 0.18 avg 0.14)
bear (149 max 0.95 avg 0.54)
dog (22 max 0.19 avg 0.10)
turtle (20 max 0.12 avg 0.07)
bg (24 max 0.07 avg 0.05)
bear (172 max 0.96 avg 0.51)
porcupine (26 max 0.31 avg 0.18)
dog (27 max 0.25 avg 0.12)
bg (228 max 0.12 avg 0.05)
whale (12 max 0.11 avg 0.05)

Figure 4: overFeat, Dense Detection [26]

structure. [3] Pooling, softmax, hinge-loss, these techniques are also used on other vision systems. The hierachical structure give CNN the ability to fulfill multiscale and context-denpendent task (like image pyramid, which be used in SIFT, etc.), and each layer is a concatenation of filters (like filter bank, adaboost). Convolutional features are local(like gabor filter) and invariant (with concatenation and pooling). (I still need to see more in literatrue to justify my conclusion. I remember there are some other advantages but I forget the source.)

[1] use overfeat-produced featrue with a linear (L2) SVM on various image recognition tasks and make a comparison with traditional enigneered features. ([1] is an interesting paper to read, even for people don't like CNN.)

CNN model could be used on smaller database by using a pre-trained

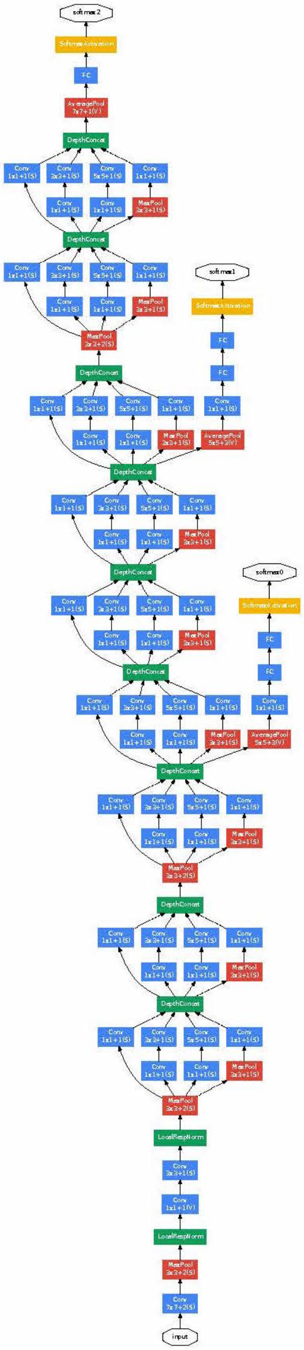


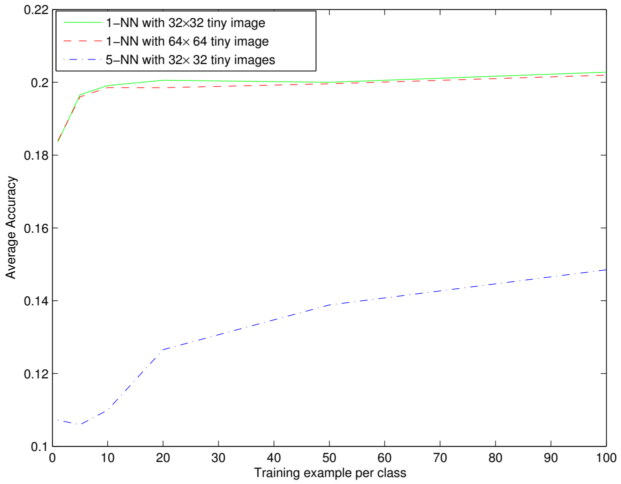
Figure 5: Stucture of GoogLeNet

model.

4 Experiments and Analysis

We randomly split samples of each class into training set and test set 10 times. For each split we use 1, 5, 10, 20, 50 and 100 images of each class for training and run test on a test set containing about 300 images for each class . We finally compute the average top-1 accuracy of the 10 splits. 1-NN with manahattan distance classifier get a better accuracy than 5-NN (we also test 10-NN is the same as 5-NN), but top-5 accuracy may differ which we haven't test. Using a larger image doesn't show improvemnt that justifies the conclusion in [27]. This naive KNN approach is still far behind state-of-the-art method which achieved 95% accuracy, but we may use scalable nearest neighbor search to handle problem with large dataset. Some futher improvement may be achieved by using space projection or more sophisticate feature like hog2-2, etc, using saliency based approach to first extract some objects first, extracting some semantic context via segmentation. The scene classification or categrization (to some extend) is difficult, because the variety within the same class, so some approaches are proposed to extract some subclass, and train model to handle such variety. An old fashion way is to trian separate models then ensemble them, but now with a better model we may handle it naturally during training.

I also have tried to train imagenet cnn model via caffe (a deep learning



framework) [8]. (But the installation is a bit nightmare, like other open source library. I spend days to fix link error and resolve compatibility issues.) Finally I got a memory not engouh exception when I trained the model with CUDA(the loss is around 7 at that time), and I don't think I could finish it in time (I'd rather to use a ec2 with tk40).

5 Discussion

5.1 Learned Features vs. Hand-made features

According to Y. Lecun[1], the drawback of hand-made features is that you have to make assumptions about the nature of the signal (image signal to computer vision). (That the only reason he said on this issue in class) In other words, hypotheses alao mean limits. Learned features have better

[1] http://techtalks.tv/talks/deep-learning/58122/, 21:00-25:00

generality and less ad-hoc design (while non-parameter approaches have the same advantage), e.g. histogram of sparse codes [24] has much simpler design options than HoG to tune. The discussion between learning and hand-made features reflects two important aspects of artificial intelligence, learning and reasoning. However, even deep learning approaches still need to be tuned [25] and reasoning in post-processing. In fact, CNN model integrates many best-practice from previous model, such as soft-max pooling in HoG (also used in other histogram-base approaches), hinge-loss in SVM, latent variable(latent SVM, HMM, etc.). Finally, many algorithm could become scalable, such Bradley-Fayyad-Reina(BFR) algorithm to K-Means, progress and sucess on other models are possible.

Generality of learned features is important for approaching ”real” AI (somewhat real), but it is fun to make features by hand. Keep calm, even fisher vector based method is still alive [17].

References

[1] Josephine Sullivan Stefan Carlsson Ali Sharif Razavian, Hossein Azizpour. Cnn features off-the-shelf: an astounding baseline for recognition. (arXiv:1403.6382 [cs.CV]), 2014.

[2] Dumitru Erhan Dragomir Anguelov Christian Szegedy, Scott Reed. Scalable, high-quality object detection. (arXiv:1412.1441 [cs.CV]), 2014.

[3] Yangqing Jia Pierre Sermanet Scott Reed Dragomir Anguelov Dumitru Erhan Vincent Vanhoucke Andrew Rabinovich Christian Szegedy, Wei Liu. Going deeper with convolutions. (arXiv:1409.4842), 2014.

[4] Thomas Dean, Mark A Ruzon, Mark Segal, Jonathon Shlens, Sudheendra Vijayanarasimhan, and Jay Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1814–1821. IEEE, 2013.

[5] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. In *Proc. 12th European Conference on Computer Vision*, 2012.

[6] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[7] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Advances in Neural Information Processing Systems*, pages 3464–3472, 2014.

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[9] Sergey Karayev, Mario Fritz, and Trevor Darrell. Anytime recognition of objects and scenes. In *Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on*. IEEE, 2014.

[10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[11] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2036–2043. IEEE, 2009.

[12] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1417–1424. IEEE, 2013.

[13] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application VISSAPP'09)*, pages 331–340. INSTICC Press, 2009.

[14] Marius Muja and David G. Lowe. Scalable nearest neighbor algorithms

for high dimensional data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36, 2014.

[15] David Sontag Nathan Silberman and Rob Fergus. Instance segmentation of indoor scenes using a coverage loss. In *ECCV*, 2014.

[16] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[17] Hao Su Jonathan Krause Sanjeev Satheesh Sean Ma Zhiheng Huang Andrej Karpathy Aditya Khosla Michael Bernstein Alexander C. Berg Li Fei-Fei Olga Russakovsky, Jia Deng. Imagenet large scale visual recognition challenge. (arXiv:1409.0575v2 [cs.CV]), 2014.

[18] Jia Deng Jonathan Krause Alexander Berg Fei-Fei Li Olga Russakovsky, Sean Ma. Imagenet large scale visual recognition challenge (ilsvrc) 2014: Introduction. 2014.

[19] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision (IJCV)*, 42(3):145–175, 2001.

[20] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[21] P. Perona P. Welinder. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *CVPR*, 2010.

[22] Devi Parikh and Kristen Grauman. Implied feedback: Learning nuances of user behavior in image search. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 745–752. IEEE, 2013.

[23] Marc'Aurelio Ranzato. Supervised deep learning. Presented as Tutorial on Deep Learning for Vision, CVPR 2014., 2014.

[24] Xiaofeng Ren and Deva Ramanan. Histograms of sparse codes for object detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3246–3253. IEEE, 2013.

[25] Trevor Darrell Jitendra Malik Ross Girshick, Jeff Donahue. Rich feature hierarchies for accurate object detection and semantic segmentation. (arXiv:1311.2524v5 [cs.CV]), 2014.

[26] Pierre Sermanet. Object detection with deep learning. Presented as Tutorial on Deep Learning for Vision, CVPR 2014., 2014.

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

[28] Serge Belongie James Hays Pietro Perona-Deva Ramanan Piotr Dollr C.

Lawrence Zitnick Tsung-Yi Lin, Michael Maire. Microsoft coco: Common objects in context. (arXiv:1405.0312 [cs.CV]), 2014.

[29] Dirk B Walther, Barry Chai, Eamon Caddigan, Diane M Beck, and Li Fei-Fei. Simple line drawings suffice for functional mri decoding of natural scene categories. *Proceedings of the National Academy of Sciences (PNAS)*, 108(23):9661–9666, 2011.

[30] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[31] Wei Zhang, Hongzhi Li, Chong-Wah Ngo, and Shih-Fu Chang. Scalable visual instance mining with threads of features. In *Proceedings of the ACM International Conference on Multimedia*, pages 297–306. ACM, 2014.

[32] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[33] C Lawrence Zitnick, Ramakrishna Vedantam, and Devi Parikh. Adopting abstract images for semantic scene understanding. In *Special Issue on the best papers at the 2013 IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR) IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). IEEE, 2015.