Deep Learning in Object Recognition, Detection, and Segmentation

Xiaogang Wang
The Chinese University of Hong Kong
xgwang@ee.cuhk.edu.hk

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

X. Wang. *Deep Learning in Object Recognition, Detection, and Segmentation.* Foundations and Trends® in Signal Processing, vol. 8, no. 4, pp. 217–382, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-117-7

Foundations and Trends® in Signal Processing
Volume 8, Issue 4, 2014

Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing

Information for Librarians

Foundations and Trends® in Signal Processing, 2014, Volume 8, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in Signal Processing
Vol. 8, No. 4 (2014) 217–382
© 2016 X. Wang
DOI: 10.1561/2000000071

Deep Learning in Object Recognition, Detection, and Segmentation

Xiaogang Wang
The Chinese University of Hong Kong
xgwang@ee.cuhk.edu.hk

Contents

Full text available at: http://dx.doi.org/10.1561/2000000071

Abstract

As a major breakthrough in artificial intelligence, deep learning has achieved very impressive success in solving grand challenges in many fields including speech recognition, natural language processing, computer vision, image and video processing, and multimedia. This article provides a historical overview of deep learning and focus on its applications in object recognition, detection, and segmentation, which are key challenges of computer vision and have numerous applications to images and videos.

The discussed research topics on object recognition include image classification on ImageNet, face recognition, and video classification. The detection part covers general object detection on ImageNet, pedestrian detection, face landmark detection (face alignment), and human landmark detection (pose estimation). On the segmentation side, the article discusses the most recent progress on scene labeling, semantic segmentation, face parsing, human parsing and saliency detection. Object recognition is considered as whole-image classification, while detection and segmentation are pixelwise classification tasks. Their fundamental differences will be discussed in this article. Fully convolutional neural networks and highly efficient forward and backward propagation algorithms specially designed for pixelwise classification task will be introduced.

The covered application domains are also much diversified. Human and face images have regular structures, while general object and scene images have much more complex variations in geometric structures and layout. Videos include the temporal dimension. Therefore, they need to be processed with different deep models. All the selected domain applications have received tremendous attentions in the computer vision and multimedia communities.

Through concrete examples of these applications, we explain the key points which make deep learning outperform conventional computer vision systems. (1) Different than traditional pattern recognition systems, which heavily rely on manually designed features, deep learning automatically learns hierarchical feature representations from massive training data and disentangles hidden factors of input data

through multi-level nonlinear mappings. (2) Different than existing pattern recognition systems which sequentially design or train their key components, deep learning is able to jointly optimize all the components and crate synergy through close interactions among them. (3) While most machine learning models can be approximated with neural networks with shallow structures, for some tasks, the expressive power of deep models increases exponentially as their architectures go deep. Deep models are especially good at learning global contextual feature representation with their deep structures. (4) Benefitting from the large learning capacity of deep models, some classical computer vision challenges can be recast as high-dimensional data transform problems and can be solved from new perspectives.

Finally, some open questions and future works regarding to deep learning in object recognition, detection, and segmentation will be discussed.

X. Wang. *Deep Learning in Object Recognition, Detection, and Segmentation.* Foundations and Trends® in Signal Processing, vol. 8, no. 4, pp. 217–382, 2014.
DOI: 10.1561/2000000071.

1

Historical overview of deep learning

This chapter will give an overview of the development of deep learning back to neural networks in 1940s, some high-impact results it has achieved since 2006, and the major differences between deep models and other machine learning models. It will also explain why neural networks were once given up by many researchers and why they became popular again since 2006.

1.1 Machine learning

Since deep learning is a subarea of machine learning, we first give a very brief introduction on what machine learning is about. Given input data $\mathbf{x}$, the goal of machine learning is to predict the output $\mathbf{y}$ through a mapping function $\mathbf{y} = f(\mathbf{x})$. If $\mathbf{y}$ is a discrete value (i.e. class label), it is a classification problem. $\mathbf{y}$ can also be a high-dimensional real-valued vector, and then it is a regression problem. Machine learning is to find the mapping function f through a set of training samples. f is assumed to be characterized with a set of parameters θ. Deep learning keeps the same goal.

At the training stage, θ is estimated from a set of training samples $\{\mathbf{x}_i\}$ with their annotated target outputs $\{\mathbf{y}_i\}$. The prediction accuracy of the learned f on test data is largely affected by the learning capacity of f as well as the scale of the training data. In the past decades, the scale of training data was small and machine learning research focused on solving the overfitting problem, i.e. the learned f has high prediction accuracy on the training data, while it performs poorly on the test data. Overfitting is caused by the mismatch between the learning capacity and the scale of training data. A well known phenomenon is the curse of dimensionality. As the dimensionality of input data $\mathbf{x}$ increases, the number of parameters as well as the learning capacity of f increases, which makes the overfitting problem even worse. In order to solve the overfitting problem, much research has been done on how to reduce model capacity by reducing the number of parameters and adding various types of regularity.

In recent years, as the emergence of large scale training data, people observed that the performance of f on test data got improved when the dimensionality of input data increased, which was called "blessing of dimensionality" [27], because larger training data required larger learning capacity. As illustrated in Figure 1.1, the performance of machine learning models with shallow structures (e.g. SVM and Boosting) gets saturated when training data becomes very large because of their limited learning capacity. They face the underfitting problem, i.e. their prediction accuracy on large-scale training data is not satisfactory.

Differently, deep neural networks could have much larger learning capacity, because of their very large numbers of parameters and deep architectures. Therefore, when training data is small, deep learning does not show major advantage compared with other machine learning methods and could even perform worse because of the overfitting problem. Under the setting of machine learning with large scale training data, deep learning makes a big difference. In order to solve the underfitting problem, it requires effectively increasing the learning capacity of models, better optimization techniques (so that the training process will not get stuck at a bad local minimum), and enough computation resources (so that the training process can be completed within a lim-

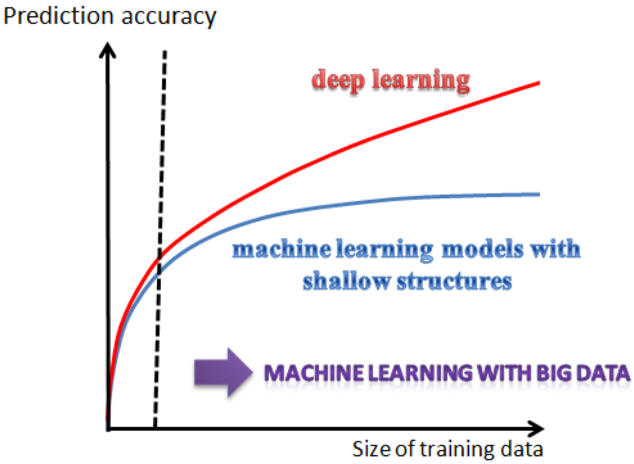


Figure 1.1: The performance of machine learning changes with the scale of training data. As the training data becomes very large, the performance of machine learning models with shallow structures gets saturated because their limited learning capacity, while the performance of deep learning keeps increasing. In the past decades, machine learning research focused on solving the overfitting problem because only small training data was available. With large-scale training data, people need to solve the underfitting problem, which is the focus of deep learning.

ited period). The research focus of deep learning has been shifted from solving the overfitting problem to these aspects, which have not been well explored in the past decades.

1.2 Neural networks

Deep models are neural networks with deep structures. The history of neural networks can be traced back to the 1940s [115]. It was inspired by simulating the human brain system and the goal was to find a principled way to solve general learning problems. It was popular in 1980s and 1990s. In 1986, Rumelhart, Hinton, and Williams published backpropagation in Nature [120], and it has been widely used to train neural networks until now. In the following subsections, we will introduce the structure of multilayer neural networks, feedforward operation used to predict output from input, and backward propagation. However, neural networks were eventually given up by most researchers because of multiple reasons which will be explained in Section 1.2.4.

1.2.1 Multilayer neural networks

The computation units of neural networks are called neurons and are organized into multiple layers. Neurons in adjacent layers are connected with weights. However, neurons in the same layer are not connected. In feedforward operation, neurons in a lower layer pass signals to neurons in its upper layer. A neuron is activated if its received signals are strong enough. Similar to the brain, some connections between neurons are stronger, while some are weaker, indicated by different weights. Figure 1.2 shows an example of a three layer neural network with an input layer, a hidden layer, and an output layer. $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_d)$ is a d dimensional input data vector. $\mathbf{h} = (h_1, \ldots, h_j, \ldots, h_k)$ are responses at n_H hidden neurons. $\mathbf{z} = (z_1, \ldots, z_k, \ldots, z_c)$ are the predicted outputs at c output neurons of the neural network. In the training set, each sample $\mathbf{x}$ is associated with a target vector $\mathbf{t}$. It is expected that output $\mathbf{y}$ predicted by the learned neural network is close to the target $\mathbf{t}$ as possible.

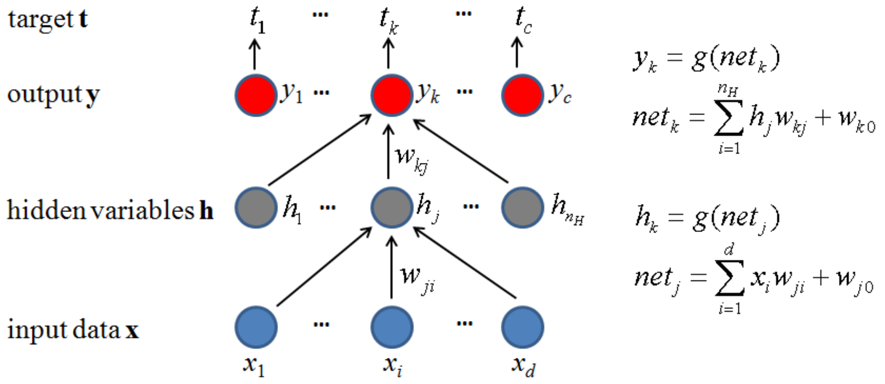

target $\mathbf{t}$
output $\mathbf{y}$
hidden variables $\mathbf{h}$
input data $\mathbf{x}$
t_1 ... t_k ... t_c
y_1 ... y_k ... y_c
w_{kj}
h_1 ... h_j ... h_{n_H}
w_{ji}
x_1 ... x_i ... x_d
$y_k = g(net_k)$
$net_k = \sum_{i=1}^{n_H} h_j w_{kj} + w_{k0}$
$h_k = g(net_j)$
$net_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0}$

Figure 1.2: Architecture of a three-layer neural network.

1.2.2 Feedforward operation

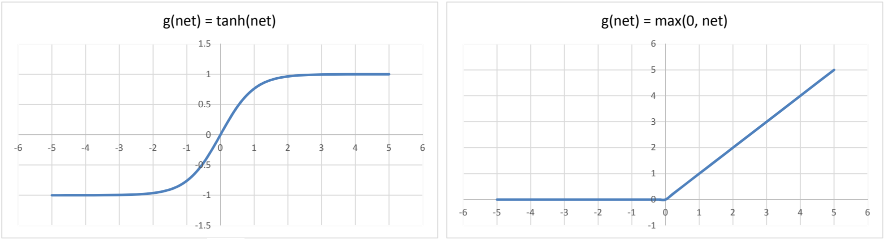
At each hidden neuron j, the weighted sum of input neurons is first computed as

$$net_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0}. \tag{1.1}$$

net_j is considered as the net activation of the hidden neuron. $\{w_{ji}\}$ are the weights of connections between the input layer and the hidden layer, and $\{w_{j0}\}$ are the bias terms. The hidden neuron emits an output y_j through a nonlinear activation function, i.e.

$$y_j = g(net_j). \tag{1.2}$$

The tanh function as shown in Figure 1.3 was widely used as the nonlinear activation function in the past. In recent years, it was found that Rectified Linear Unit (ReLU) leads to sparse neural responses and is more effective in many cases. There are also other choices, such as Parameterized Rectified Linear Unit (PReLU) [63]. Taking ReLU as an example, the hidden neuron emits no response unless the activation is larger than a threshold.



(a) (b)

Figure 1.3: Examples of nonlinear activation functions. (a) is the tanh fuction, i.e. $g(net) = \frac{e^{net} - e^{-net}}{e^{net} + e^{-net}}$. (b) is the Rectified Linear Unit (ReLU), i.e. $g(net) = max(0, net)$.

In the output layer, each output neuron k also first compute its net activation from the signals sent by hidden neurons,

$$net_k = \sum_{j=1}^{n_H} y_j w_{kj} + w_{k0}. \tag{1.3}$$

$\{w_{kj}\}$ are the weights and $\{w_{k0}\}$ are the bias terms. The output neuron k emits z_k through the nonlinear activation function of its net activation, i.e.

$$z_k = g(net_k). \tag{1.4}$$

Summarizing Eq. (1.1) - (1.4), the output of the neural network is equivalent to a set of discriminant functions

$$f_k(\mathbf{x}) \equiv z_k = g\left(\sum_{j=1}^{n_H} w_{kj} g\left(\sum_{i=1}^{d} w_{ji}x_i + w_{j0}\right) + w_{k0}\right). \tag{1.5}$$

It is achieved by a series of linear and nonlinear transforms computed at multiple layers.

1.2.3 Backpropagation

Training a neural network is to find an optimal set of weights (including bias terms) $\mathbf{W}$ to minimize an objective function $J(\mathbf{W})$, such that the predicted outputs $\mathbf{z}$ of training samples are close to the targets $\mathbf{t}$ as possible. Backpropagation (BP) [120] proposed in 1980s is still the most widely used method for supervised training of neural networks. It is a gradient descent algorithm. Weights are randomly initialized and them updated iteratively. At each iteration, weights are changed in a direction to reduce the objective function,

$$\mathbf{W} \longleftarrow \mathbf{W} - \eta \nabla J(\mathbf{W}), \tag{1.6}$$

where η is a hyperparameter of learning rate and $\nabla J(\mathbf{W})$ is gradient of the objective function w.r.t. weights $\mathbf{W}$. As shown in Figure 1.4 (a), training samples are fed in the input layer of the neural network. With feedforward operation, outputs are predicted in the output layer. Prediction errors are computed by comparing with the target values. With BP, errors are propagated back to each layer and used to compute the gradients of weights in each layer. A detailed description of the BP algorithm can be found in [46].

The surface of the objective function of a neural network is typically highly complex with many local minima as shown in Figure 1.4 (b). There is no theoretical guarantee that the global minimum can be

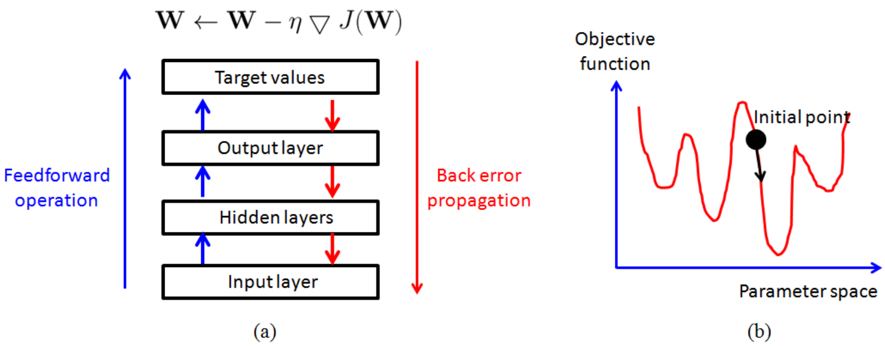


Figure 1.4: Backpropagation. (a) Illustration the BP process of training neural networks. (b) The performance of the trained neural networks with BP depends on the initialization point.

achieved by BP on general neural networks. The local minimum reach by gradient descent depends on the initialization of network weights. Some works [66] have been done to pretrain neural networks such that they can start with a good initialization point and reach a better local minimum after the convergence of BP.

Given n training samples, in batch gradient descent, the objective function can be expressed as

$$J(\mathbf{W}) = \sum_{p=1}^{n} J_p(\mathbf{W}), \tag{1.7}$$

where $J_p(\mathbf{W})$ is the prediction cost on the pth training sample, and the weights are updated as

$$\mathbf{W} \longleftarrow \mathbf{W} - \eta \sum_{p=1}^{n} \nabla J_p(\mathbf{W}). \tag{1.8}$$

However, when the training set is large, evaluating the sum-gradient is computationally expensive. Stochastic gradient descent samples a subset of summand functions at every iteration. This is very effective in the case of large-scale machine learning problems. In stochastic training, the training set is divided into mini-batches, and the true gradient of $J(\mathbf{W})$ is approximated at a mini-batch of samples. Estimate of the gradient is noisy, and the weights may not move precisely down the gradient at each iteration, but is much faster than batch learning. On the

Full text available at: http://dx.doi.org/10.1561/2000000071

other hand, noise may result in better solutions. The weights fluctuate, which makes it possible to jump out of bad local minima.

1.2.4 Difficulties of employing neural networks

People encountered several major problems when employing neural networks in various applications in 1980s and 1990s. Neural networks typically have a large number of parameters and it was difficult to train them. It was easy for neural networks to overfit on training sets, while they performed poorly on test sets. It lacked large scale training data, which made the overfitting problem even more severe. Even a relatively large training set only had a few hundred training samples. Moreover, with very limited computational power available in 1980s and 1990s, it took a long time to train a small neural network. In general, the performance of neural networks was not significantly better than other machine learning tools and it was much more difficult to train neural networks. Therefore, many researchers gave up neural networks in early 2000s and turned to other machine learning tools such as SVM, Boosting, decision tree, and K-Nearest Neighbor.

1.3 Other machine learning models

Other machine learning models can be approximated with neural networks with only one or two hidden layers. Therefore, they are called models with shallow structures. An example of SVM is shown in Figure 1.5. The prediction function of SVM can be written as

$$f(\mathbf{x}) = b + \sum_{i=1}^{M} K(\mathbf{x}_i, \mathbf{x}). \tag{1.9}$$

$\mathbf{x}$ is a test sample. $\mathbf{x}_i$ is a support vector. There are totally M support vectors. K is the kernel function to measure the similarity between $\mathbf{x}$ and $\mathbf{x}_i$. As shown in Figure 1.5, SVM can be implemented with a three-layer neural network with $M+1$ hidden neurons. $K(\mathbf{x}_i, \mathbf{x})$ is output at each hidden neuron i.

These models have loose ties with biological systems. Instead of solving general learning problems, people designed specific systems

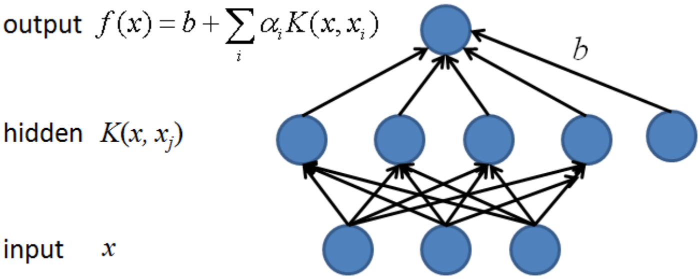


Figure 1.5: SVM can be approximated with a three layer neural network.

(models) for specific tasks and used different handcrafted features. For example, HMM-GMM was used in speech recognition, SIFT was used in object recognition, LBP was used in face recognition, and HOG was used in human detection.

1.4 Deep learning

Deep learning has become popular since 2006 [67, 66]. A major breakthrough in deep learning was first achieved in speech recognition [65]. It outperformed HMM-GMM, which dominated the field for many years, by a large margin. There are a few reasons making neural networks successful again. First of all, a key reason is the emergence of large scale training data with annotations. For example, ImageNet [36] has millions of images with annotated class labels. With large-scale training data, deep neural networks show significant advantages compared with shallow models because of their very large learning capacity. With the fast development of high performance parallel computing systems, such as GPU clusters, it has become much easier to train large-scale deep neural networks with millions of parameters.

Moreover, there has been significant advances in the design of network structures, models, and training strategies. For example, unsupervised and layerwise pre-training has been proposed. It makes a neural network reach a good initialization point. Based on that, fine-tuning with BP can find a better local minimum. It helps to solve the underfitting problem in large-scale training sets to some extent. Dropout and data augmentation [80] have been proposed to solve the overfitting problem in training. Batch normalization [96] has been proposed to train very deep neural networks efficiently. Various network structures such as AlexNet [80], Clarifai [173], Overfeat [125], GoogLeNet [138], and VGG [128] have been extensively studied to optimize the performance of deep learning.

1.5 Deep learning achievements in computer vision

1.5.1 Object recognition and detection

Deep learning started to have a huge impact on computer vision in 2012, when Hinton's group won the ImageNet Large Scale Visaul Recognition Challenge (ILSVRC) with deep learning [80]. Before that, there were attempts to apply deep learning to relatively small datasets and the obtained improvement was marginal compared with other computer vision methods. The computer vision community was not fully convinced that deep learning would bring revolutionary breakthrough without strong evidence on grand challenges until 2012.

ILSVRC is one of the most important grand challenges in computer vision, and has drawn the a lot of attention recently especially after the great success of deep learning in 2012. It was originally proposed in 2009 [36]. The challenge was to classify images collected from the web into $1,000$ categories. Its training data includes more than one million images, much large than other datasets previously used to evaluate deep learning, such as MNIST [1]. This competition has been running for several years and many top computer vision groups participated in the competition. However, different computer vision systems for object recognition tended to converge and there was no real breakthrough until 2012. This section reviews the ILSVRC results from 2012 to 2014, so that readers can understand how fast deep learning has been developing in computer vision.

Hinton's group participated in this challenge at ILSVRC 2012. As shown in Table 1.1, the teams ranking from No. 2 to No. 4 all used conventional computer vision technologies and handcrafted features.

[1] http://yann.lecun.com/exdb/mnist/

Rank	Group	Top-5 error rate (%)	Description
1	U. Toronto	15.132	Deep learning
2	U. Tokyo	26.172	Handcrafted features
3	U. Oxford	26.979	Handcrafted features
4	Xerox/INRIA	27.058	Handcrafted features

Table 1.1: Performance of top ranked groups on the image classification task in ILSVRC 2012. Since each image from ImageNet may contain multiple objects, top-5 error rate was commonly used for evaluation. Deep learning outperformed other computer vision methods based on handcrafted features by more than 10%.

The differences between their classification accuracies were less than 1%. Since each image from ImageNet may contain multiple objects, top-5 error rate was commonly used for evaluation. The classification of an image is considered as correct if its labeled ground truth is among the top five classes predicted by the model. However, Hinton's group outperformed them by more than 10%, reaching the top-5 error rate of 15.3%. They employed the convolutional neural network (CNN) [83] implemented with two GPUs.

The computer vision community was shocked by this result. Many people believed that a revolutionary breakthrough was brought by deep learning to this field. Shortly thereafter, people found that the visual feature representation learned from ImageNet could be well generalized to other datasets and computer vision tasks, such as object detection [56], image segmentation [97], image retrieval [154] and object tracking [69]. For example, another well known object recognition and detection challenge is PASCAL VOC. However, its training set is too small to train deep models. Girshick *et al.* [56] applied the features learned from ImageNet with the image classification task and deep CNN to object detection on PSACAL VOC. The detection rate was improved by 20%. This conclusion has significant impact. It indicates that once better features are learned by deep learning on ImageNet, many other computer vision problems can be improved accordingly. Therefore, deep learning on ImageNet has become the engine driving the computer vision field. That is one of the reasons that it has drawn most attention recently.

Rank	Group	Top-5 error rate (%)	Description
1	NYU	11.197	Deep learning
2	NUS	12.535	Deep learning
3	Oxford	13.555	Deep learning

Table 1.2: Performance of top ranked groups on the image classification task in ILSVRC 2013.

Rank	Group	mAP (%)	Description
1	UvA-Euvision	22.581	Handcrafed features
2	NEC-MU	20.895	Handcrafed features
3	NYU	19.400	Deep learning

Table 1.3: Performance of top ranked groups on the object detection task in ILSVRC 2013.

In ILSVRC 2013, the teams ranking top 20 all used deep learning. As shown in Table 1.2, the winner deep model was called Clarifai from NYU. The error rate was reduced to 11.19%. In that year, an object detection challenge was added. It required detecting objects of 200 categories from $40,000$ test images. It is much more challenging than image classification, since each image may contain multiple objects of different categories. The highest mean Average Precision (mAP) was only 22.58%. The top two winners still used handcrafted features instead of deep learning.

In ILSVRC 2014, much deeper CNNs were employed. As shown in

Rank	Group	Top-5 error rate (%)	Description
1	Google	6.656	Deep learning
2	Oxford	7.325	Deep learning
3	MSRA	8.062	Deep learning

Table 1.4: Performance of top ranked groups on the image classification task in ILSVRC 2014.

Rank	Group	mAP (%)	Description
1	Google	43.933	Deep learning
2	CUHK	40.656	Deep learning
3	DeepInsight	40.452	Deep learning
4	UvA-Euvision	35.421	Deep learning
5	Berkley	34.521	Deep learning

Table 1.5: Performance of top ranked groups on the object detection task in ILSVRC 2014.

	RCNN	Berkley Vision	DeepInsight	GoogLeNet (Google)	DeepID-Net (CUHK)
Avg	n/a	n/a	40.5	43.9	50.3
Single	31.4	34.5	40.2	38.0	47.9

Table 1.6: Summary of mAP on ImageNet with different deep learning based object detection methods. "Single" represents the results achieved with single models. "Avg" represents the results achieved with model averaging. It has been well known that model averaging generally leads to improvement on image classification and object detection.

Table 1.4, GoogLeNet [138] had more than 20 layers, and won both the image classification and object detection challenges. VGG [128] from Oxford won the localization challenge also with a very deep network. The image classification top-5 error rate was reduced to 6.66% and the mAP for object detection was largely improved to 43.93% as shown in Table 1.5. Table 1.6 summaries the progress of deep learning based object detection on ImageNet. RCNN [56] was the first widely used deep learning pipeline for general object detection and was proposed in 2013. The most recent work DeepID-Net [109] has significantly advanced the state-of-the-art to mAP of 50.3.

1.5.2 Face recognition

Another major challenge in computer vision is face recognition. Labeled Faces in the Wild (LFW) [73] is the most well known benchmark in face recognition. Most of the groups or companies working on face recognition reported their results on LFW. Many face recognition datasets were collected in lab environments under controlled condition. In 2007, Huang *et al.* created the LFW dataset, which included face images of celebrities from the web, in order to evaluate face recognition performance in unconstrained conditions. Its test set includes $6,000$ pairs of images and computation algorithms need to tell whether an image pair comes from the same person or not. The chance of random guess is 50%. According to the study [82], when only the central face regions (excluding hair) were cropped and shown to humans, the face verification accuracy by human eyes was 97.53%. When the whole images including hairs were shown to humans, the face verification accuracy by human eyes was 99.20%. A classical face recognition method, i.e. Eigenface [148], only has 60% accuracy on LFW. It shows that the dataset is quite challenging. The best performing non-deep-learning technology [27] obtained 96.33% face verification on LFW. With deep learning, it was the first time for DeepID2 [135] to achieve face verification accuracy of 99.15% on LFW, comparable with human performance on this benchmark. Now the new state-of-the-art DeepID2+ [137] and FaceNet [123] have achieved face verification accuracy of 99.45% and 99.63% on LFW respectively, surpassing human performance.

1.5.3 Impact on industry

Deep learning brings big impact on the computer vision community as well industry. Six months after Hinton's group won ILSVRC 2012, both Google and Baidu released their new visual search engines by applying the same deep model used by Hinton's group in ILSVRC 2012 to their own data. It was observed that the average precision was doubled. The following paragraph is from the news released by Google:

"On our test set we saw double the average precision when compared to other approaches we had tried. We acquired the rights to the technology and went full speed ahead adapting it to run at large scale on Google's computers. We took cutting edge research straight out

of an academic research lab and launched it, in just a little over six months"

Hinton joined Google a few months after he won ILSVRC 2012. Baidu established the Institute of Deep Learning in 2012, and recruited Andrew Ng, a well known professor from Stanford working on deep learning, as the director of their new lab in the silicon valley in May 2014. In December 2014, Facebook established a new AI lab in the NewYork City, dedicated to deep learning, and recruited Yann LeCun as the director, who is a well known pioneer on deep learning. In January 2014, Google spent 400 million US dollars to acquire DeepMind, a startup company working on deep learning. Nowadays, many startup companies emerge and work on computer vision applications with deep learning technologies. MIT technology review listed deep learning as one of the top ten breakthrough technologies in 2013.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2009.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(11):2271–2282, 2012.

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[4] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2007.

[5] B. Amberg and T. Vetter. Optimal landmark detection using shape models and branch and bound. In *Proc. IEEE Int'l Conf. Computer Vision*, 2011.

[6] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Proc. IEEE Int'l Conf. Computer Vision*, 2011.

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *Proc. Int'l Conf. Human Behavior Understanding*, 2011.

[8] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transforms. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1997.

[10] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[11] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[12] Y. Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 7, 2009.

[13] Y. Bengio, A. Courville, and P. Vincent. Representation learning: a review and new perspectives. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(8):1798–1827, 2013.

[14] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Proc. Neural Information Processing Systems*, 2007.

[15] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai. Better mixing via deep representations. In *Proc. Int'l Conf. Machine Learning*, 2013.

[16] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: identifying a person of interest from a media collection. *TR MSU-CSE-14-1*, 2014.

[17] L. Bourdev and J. Brandt. Robust object detection via soft cascade. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.

[18] L. Bourdev and J. Malik. Poselets: body part detectors trained using 3d human pose annotations. In *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

[19] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *Proc. European Conf. Computer Vision*, 2008.

[20] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(8):1872 – 1886, 2013.

[21] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou. Deep nonlinear metric learning with independent subspace analysis for face verification. In *Proc. ACM Multimedia*, 2012.

[22] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[23] X. Cao, D. Wipf, F. Wen, and G. Duan. A practical transfer learning algorithm for face verification. In *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[24] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[25] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. *arXiv:1507.06550*, 2015.

[26] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. European Conf. Computer Vision*, 2012.

[27] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[28] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Proc. Neural Information Processing Systems*, 2014.

[29] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[30] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[31] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.

[32] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[33] G. Cybenko. Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

[34] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.

[35] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[36] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2009.

[37] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *Proc. European Conf. Computer Vision*, 2012.

[38] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

[39] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[40] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2009.

[41] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36:1532–1545, 2014.

[42] P. Dollar, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *Proc. European Conf. Computer Vision*, 2012.

[43] P. Dollar, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *Proc. British Machine Vision Conference*, 2009.

[44] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2009.

[45] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[46] R. O. Duda, P. E. Hart, and D. G. Stork, editors. *Pattern Classification*, chapter multilayer neural networks, pages 282–335. John Wiley & Sons, Inc., 2000.

[47] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object clclass (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[48] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation. *arXiv:1504.07159*, 2015.

[49] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35:1915–1929, 2013.

[50] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[51] P. F. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[52] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:55–79, 2005.

[53] W. A. Freiwald and D. Y. Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010.

[54] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[55] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[57] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. *arXiv:1403.1840*, 2014.

[58] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

[59] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *Pooc. Int'l Conf. Automatic Face and Gesture Recognition*, 2008.

[60] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Proc. Neural Information Processing Systems*, 2006.

[61] J. Hastad. Almost optimal lower bounds for small depth circuits. In *Proc. ACM Symposium on Theory of Computing*, 1986.

[62] J. Hastad and M. Goldmann. On the power of small-depth threshold circuits. *Computational Complexity*, 1:113–129, 1991.

[63] K. He, Z. Zhang, S. Ren, and Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv:1502.01852*, 2015.

[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proc. European Conf. Computer Vision*. 2014.

[65] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition — the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012.

[66] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1544, 2006.

[67] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

[68] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[69] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *Proc. Int'l Conf. Machine Learning*, 2015.

[70] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[71] X. Hou, J. Harel, and C. Koch. Image signature: highlighting sparse salient regions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34:194 – 201, 2012.

[72] G. B Huang, H. Lee, and E. Learned-Miler. Learning hierarchical representation for face verification with convolutional deep belief networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[73] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miler. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[74] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

[75] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[76] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zeng, and S. Li. Salient object detection: a discriminative regional feature integration approach. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[77] K. Kang and X. Wang. Fully convolutional neural networks for crowd segmentation. *arXiv:1411.4464*, 2014.

[78] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[79] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierachies for visual recognition. In *Proc. Neural Information Processing Systems*, 2010.

[80] A. Krizhevsky, L. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Neural Information Processing Systems*, 2012.

[81] N. Kumar, P. N. Belhumeur, and S. Nayar. Facetracer: a search engine for large collections of images with faces. In *Proc. European Conf. Computer Vision*, 2008.

[82] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

[83] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

[84] Y. LeCun, L. Bottou, G. B. Orr, and K. Muller. Efficient backprop. Technical report, 1998.

[85] H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area v2. In *Proc. Neural Information Processing Systems*, 2007.

[86] H. Li, R. Zhao, and X. Wang. Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification. *arXiv:1412.4526*, 2014.

[87] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[88] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *Proc. IEEE Int'l Conf. Computer Vision*, 2012.

[89] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[90] C. Liang, X. amd Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *Proc. IEEE Int'l Conf. Computer Vision*, 2015.

[91] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *Proc. European Conf. Computer Vision*, 2008.

[92] M. Lin, Q.. Chen, and S. Yan. Network in network. *arXiv:1312.4400v3*, 2013.

[93] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets knn: Quasi-parametric human parsing. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[94] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proc. IEEE Int'l Conf. Computer Vision*, 2015.

[95] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *Proc. IEEE Int'l Conf. Computer Vision*, 2014.

[96] S. Loffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.

[97] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[98] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[99] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[100] P. Luo, X.. Wang, and X.. Tang. Pedestrian parsing via deep decompositional neural network. In *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[101] R. Mairon and O. Ben-Shahar. A closer look at context: From coxels to the contextual emergence of object saliency. In *Proc. European Conf. Computer Vision*, 2014.

[102] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[103] G. A. Miller, R. Beck, C. D. Fellbaum, D. Gross, and K. Miller. Wordnet: An online lexical database. *International Journal of Lexicograph*, 3:235–244, 1990.

[104] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33:1771–1782, 2000.

[105] V. Nair and G. Hinton. Implicit mixtures of restricted boltzmann machines. In *Proc. Neural Information Processing Systems*, 2008.

[106] S. Ohayon, W. A. Freiwald, and D. Y. Tsao. What makes a cell face selective? the importance of contrast. *Neuron*, 74:567–581, 2013.

[107] W. Ouyang, X. Chu, and X.. Wang. Multi-source deep learning for human pose estimation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[108] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[109] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yan, Z. Wang, C. C. Loy, and X. Tang. Deepid-net: Deformable deep convolutional neural networks for object detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[110] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Modeling mutual visibility relationship in pedestrian detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[111] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *Proc. European Conf. Computer Vision*, 2010.

[112] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[113] P. O. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *Proc. Int'l Conf. Machine Learning*, 2014.

[114] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[115] W. Pitts and W. S. McCulloch. How we know universals: The preception of auditory and visual forms. *Bulletin of Mathematical Biophysics*, 9:127–147, 1947.

[116] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *Proc. Neural Information Processing Systems*, 2006.

[117] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *Proc. Neural Information Processing Systems*, 2007.

[118] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv:1403.6382*, 2014.

[119] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proc. Int'l Conf. Machine Learning*, 2011.

[120] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. *Nature*, 323(99):533–536, 1986.

[121] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[122] K. Sande, C. G. M. Snoek, and A. W. M. Smeulders. Fisher and vlad with flair. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[123] F. Schroff, D. Kalenichenko, and J. Phibin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[124] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *arXiv:1503.02351*, 2015.

[125] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. Int'l Conf. Learning Representations*, 2014.

[126] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[127] J. Shao, K. Kang, C. C. Loy, and X. Wang. Deeply learned attributes for crowded scene understanding. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[128] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[129] K. Sohn, G. Zhou, C. Lee, and H. Lee. Learning and selecting features jointly with point-wise gated boltzmann machines. In *Proc. Int'l Conf. Machine Learning*, 2013.

[130] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[131] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[132] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv:1502.00873*, 2015.

[133] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[134] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for computing face similarities. In *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[135] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proc. Neural Information Processing Systems*, 2014.

[136] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[137] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective and robust. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[138] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014.

[139] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[140] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale trainnig for face identification. *arXiv: 1406.5266*, 2014.

[141] Siyu Tang, Mykhaylo Andriluka, Anton Milan, Konrad Schindler, Stefan Roth, and Bernt Schiele. Learning people detectors for tracking in crowded scenes. *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[142] Y. Tian, P. Luo, X.. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *Proc. IEEE Int'l Conf. Computer Vision*, 2015.

[143] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[144] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *Proc. European Conf. Computer Vision*, 2012.

[145] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *Proc. European Conf. Computer Vision*, 2010.

[146] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Proc. Neural Information Processing Systems*, 2014.

[147] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[148] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[149] J. R. R. Uijlings, K. E. A. Van de Sande, T. Gevers, and W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171, 2013.

[150] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[151] P. Vincent, H. Larochelle, H. Bengio, and P. A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proc. Int'l Conf. Machine Learning*, 2008.

[152] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. In 11, editor, *Journal of Machine Learning Research*, pages 3371–3408, 2010.

[153] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[154] J. Wan, D. Wong, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proc. ACM Multimedia*, 2014.

[155] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[156] X. Wang and X. Tang. Unified subspace analysis for face recognition. In *Proc. IEEE Int'l Conf. Computer Vision*, 2003.

[157] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26:1222–1228, 2004.

[158] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1:270–280, 1989.

[159] L. Wiskott, J. Fellous, N. Kruger, and C. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[160] L. Wolf, T. Hassner, and Y. Taigman. Effective face recognition by combining multiple descriptors and learned background statistics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33:1978–1990, 2011.

[161] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.

[162] F. Xia, J. Zhu, P. Wang, and A. Yuille. Pose-guided human parsing with deep learned features. *arXiv:1508.03881*, 2015.

[163] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multi-pedestrian detection in crowded scenes: A global view. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[164] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[165] C. Yang, L. Zahng, H. Lu, X. Ruan, and M. Yang. Saliency detection via graph-based manifold ranking. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2013.

[166] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[167] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[168] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[169] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.

[170] L. Yann, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[171] A. Yao. Separating the polynomial-time hierarchy by oracles. In *Proc. IEEE Symposium on Foundations of Computer Science*, 1985.

[172] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[173] M. Zeiler. Clarifai. www.clarifai.com.

[174] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional neural networks. *arXiv preprint arXiv:1311.2901*, 2013.

[175] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[176] X. Zeng, W. Ouyang, and X. Wang. Deep learning of scene-specific classifier for pedestrian detection. In *Proc. European Conf. Computer Vision*, 2014.

[177] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2014.

[178] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Proc. IEEE Int'l Conf. Computer Vision*, 2005.

[179] Z. Zhang, P. Luo, C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proc. European Conf. Computer Vision*, 2014.

[180] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2015.

[181] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. *arXiv:1502.03240*, 2015.

[182] L. Zhu, Y. Chen, and A. Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(6):1029–1043, 2010.

[183] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2010.

[184] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2012.

[185] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity preserving face space. In *Proc. IEEE Int'l Conf. Computer Vision*, 2013.

[186] Z. Zhu, P. Luo, X. Wang, and Tang. X. Multi-view perceptron: a deep model for learning face identity and view representations. In *Proc. Neural Information Processing Systems*, 2014.