





Convolutional Neural Networks Implementations for Computer Vision

Paweł Michalski¹ , Bogdan Ruszczak² , and Michał Tomaszewski¹  

¹ Faculty of Electrical Engineering, Automatic Control and Informatics,
Institute of Computer Science, Opole University of Technology,
Prószkowska 76, 45-758 Opole, Poland

{p.michalski,m.tomaszewski}@po.opole.pl

² Faculty of Economy and Management, Luboszycka 7, 45-036 Opole, Poland
b.ruszczak@po.opole.pl

Abstract. The paper covers the current state of the art regarding the use of machine learning mechanisms, and in particular the deep convolutional neural networks used in the field of computer vision. In the article there has been presented the current definition of deep learning and specific dependencies between related fields such as machine learning and artificial intelligence. The practical part of the work consists of three components: the features of the structure of the convolutional neural network, the distinction of its key elements, the description of their actions, the compilation of information about available learning sets used in network testing and verification processes, and the review of the implementation of convolutional neural networks, which had a significant impact on development of discipline. To illustrate the great potential of the presented tools for solving computer vision tasks, the study highlights examples of their applications. The possibility of using convolutional neural networks for identification of technical objects in digital images is indicated.

Keywords: Deep learning · Convolutional neural networks
Computer vision · Artificial intelligence · Machine learning

1 Introduction

Processing the large data sets (big data) is currently a rapidly growing discipline due to increasing demand in a number of areas, such as medical data processing [1], supporting the movement of autonomous cars [2], or the processing of the digital images [3]. Machines, as a tool to replace people's work, are now equipped with sets of sensors that respond to human senses. In the case of people, the sense of sight is one of the most important senses, and provides a great deal of information about the surroundings. Despite the fact that for a long time there

Paweł Michalski, PhD. Eng., Assistant Professor; Bogdan Ruszczak, PhD. Eng., Assistant Professor; Michał Tomaszewski, PhD. Eng., Associate Professor.

has been a wide range of vision sensors available to record images in various electromagnetic fields, it is still troublesome to process the vision information in the shortest time possible, especially since many of these tasks require real-time analysis. The autonomous machine that performs the tasks required should very often have the ability to determine its position relative to the surroundings, such as furniture, traffic congestion in rooms or other road users [4, 5]. Machines, like people, must make a specific decision based on acquired images. Another issue is to identify elements of the environment, which is widely analyzed by researchers around the world [6].

Another task, pattern recognition, which for most people is not difficult, in the case of computers turns out to be a much more complicated process. This is most often caused by the necessity to map 3D objects with a single image or a set of two-dimensional images. Additionally, the recording of images can be done in different lighting conditions, which also affects the degree of difficulty of the process. Another difficulty may be partial occlusion of the subject or image projection from another perspective as compared to the master image. Despite the differences in lighting and in spite of the various points of view, the task of classifying a facility by man in most cases is still successful. In the case of automated processes performed by the machines it is required to use one of the common patterns matching algorithms, such as Scale Invariant Feature Transform (SIFT), Speed Up Robust Feature (SURF) or Robust Independent Elementary Features (BRIEF). Some variants of these algorithms, for example Oriented FAST or Rotated BRIEF (ORB) can accurately detect a pattern with deformations such as the change of scaling or rotation [7]. The key to a good object recognition is the set of features which describe it. Such features are then sought for in test images, and on this basis a set of features, describing the object being tested, is created.

Classic Computer Vision (CV) algorithms unfortunately have their limitations and usually get the correct results with small deformations of objects, and under similar lighting conditions, which were associated with the building of a model set of characteristics describing an object. Widespread use of Convolutional Neural Networks (CNN) as a tool to aid the machines to make decisions based on a set of images of the surrounding world was possible through two processes. The first was the appearance of CPUs with high computing abilities, especially high-performance graphics cards processors which do very well in the network learning process. The second was the development of Machine Learning (ML) techniques, in particular Deep Learning (DL) algorithms.

Both ML and DL are concepts in the field of research of Artificial Intelligence (AI) or Computational Intelligence (CI). The machine learning process can be divided into two main groups: unsupervised learning and supervised learning, where the machine learns based on test data. The test data usually comes in the form of input/output pairs, written to give you basic information to make learning the future decisions of the system. Such a learning process involves the participation of man in the learning process. During supervised learning process for each input information must first be assigned a correct response at the output

of the circuit. It's not always possible to use this automatic learning method, then the use of unattended teaching remains an alternative. DL is implemented by large-scale neural networks, which the best example are the deepest convolutional neural networks described in this paper.

The rapid development of AI techniques is measured with some troubles, resulting from the activity of the proposed algorithms. In the discussion initiated by the authors in [8], the main limitations of AI were identified. During the processing large collections of learning, it's an difficult task to automatically identify costly and often repetitive activities in the process of searching for a solution. With limited information, the algorithm has no basis for automatically eliminating certain sets of solutions, even if it might seem pointless to humans. One of the limitations is also the insufficient skill of AI algorithms to adapt to changing conditions. Their work runs smoothly if they are powered with data for which they have been prepared, or more broadly, to solve even complex problems but with the right structure.

2 Deep Learning

DL is based on in-depth AI research and is part of ML, and was created during the explosion of popularity of AI-related IT tools. The illustration below (Fig. 1) shows technologies sometimes called “narrow AI” – they perform precisely defined tasks with similar efficiency as humans, and sometimes even better. Examples of such tools can be automatic classification of photos in Pinterest or face recognition in photos posted on Facebook.

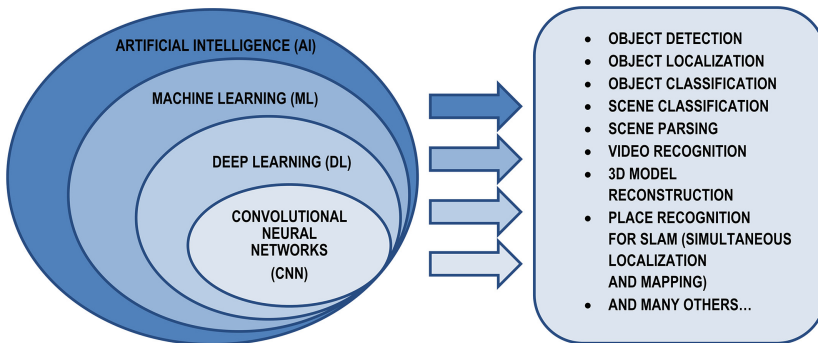


Fig. 1. Relationships between artificial intelligence, machine learning, deep learning and convolutional neural networks and examples of applications in computer vision.

In the simplest sense, ML is a process that comprises of parsing data, automatic learning based on them, and subsequent results determination or prediction estimation. ML is rather training to perform specific tasks using a large set of data than manually algorithms coding to solve problems.

ML is an attempt to tackle the challenges of AI at the beginning of its development, which have been tried before with, for instance: decision tree learning, clustering, reinforcing learning, Bayesian networks, inductive logic programming, etc. Those methods allowed to perform a narrow range AI only. Nowadays researchers try to solve these issues using ML. One of the first major projects in ML was Google Brain, which created the TensorFlow library, which is one of the popular tools for creating and teaching neural networks, published under the open Apache license. Another interesting use of deep learning in practice is the voice recognition implemented in applications like Google Now and Apple Inc.'s Siri.

DL has enabled the practical solution of many tasks where ML was not enough. Issues such as the navigation of autonomous cars, recommendations for consumers in sales systems or support for preventive health care are currently being implemented by the DL. One of the areas where using such algorithms is nowadays broadly implemented is CV.

3 Convolutional Neural Networks as a Technique for Deep Learning

According to [9], Deep Learning is a section of Machine Learning, centered around algorithms modeling high-level abstraction in data sets, using multiple layers of nonlinear transformations. The most well-known and used group of deep learning algorithms are Convolutional Neural Networks because of their wide application possibilities in recognizing different patterns, particularly in detecting objects in digital images.

The idea of CNN has been around for a long time, but there were many problems that inhibited its development. For example, with the enlarging number of network's layers, the number of model parameters increases, which, in conjunction with the simultaneous rise in the size of the input training base, significantly expands the demand for computing power. The vanishing gradient problem, associated with the use of the reverse propagation algorithm, also required a solution.

In recent years there has been a rapid development of CNN, and the impact of many difficulties has been significantly reduced as a result of research into new concepts for the broader neural network, including [9–16]. At the same time, there has been an increase in the ability to create very large datasets [17] based on a big data revolution.

4 Components of the Convolutional Neural Networks

What are Convolutional Neural Networks? In the construction of these networks, there are four main types of layers that perform the basic tasks of such networks: convolution, pooling, normalization and connection. As a result of the Convolution Layer, the input image is processed by a variety of convolutional filters

to extract the characteristics contained in those parts. Pooling Layer is being use for reduce the size of the information being analyzed, thereby decreasing the sensitivity of the network to the distortion of the analyzed scene. The basic methods used in this layer are max pooling, when the largest value is selected in the parsed window and averaging, when its value is averaged. The ReLED layer (Rectified Linear Units Layer) by data normalisation increases the network's ability to solve nonlinear problems. CNN consist of multiple layers on successive levels, but the last link in such a network is the submission of results to the final layer – Fully Connected Layer. This layer results in the final rating, allowing the various training categories assignment.

The distinguishing feature of CNN over classical neural networks is that the number of layers is much higher. The Fig. 2 shows the structure of CNN. The depth of neural network architecture is defined as the length of the longest path between the input and output neurons. There is no precise threshold of the layers number, allowing one to call the network “deep”, but it has been assumed that it refers to the network with more than two hidden layers.

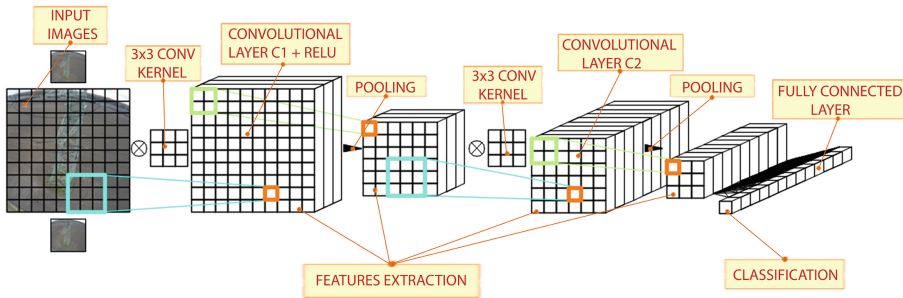


Fig. 2. The general structure of the convolutional neural network with distinction of the most essential components.

The general structure of the CNN described above can be modified. An alternative to the ReLU layer may be to use the Softmax or Parametric Rectified Linear Units (PReLU) function [16], which could improve the network structure. Another interesting approach to CNN modification using the parallel processing of several convolutional layers (Inception Module) is presented in [18].

The process of designing a CNN requires matching its structure - determining the number of layers and their respective layout, as well as the parameters of their work. The preparation of each layer consists in defining the so-called hyperparameters, such as depth defining the number of neurons (different type and size features or filters on each of the Convolutional Layers), stride deciding how dense sampling of layers will be performed, zero-padding specifying how to supplement incomplete information (eg. on the edges of the processed images), and the number of neurons in the Fully Connected Layer.

The CNN learning involves processing input datasets (e.g. image collection) with assigned categories. This is done in unattended mode. The mechanism used to train the network that distinguishes the convolutional neural networks is the error backpropagation. This method allows to improve the grading scales of the individual layers based on the observation of the adjustment evaluation error function. The purpose of the operation is to modify the weight of the classifier so that the observed adjustment error is lower. The phenomenon of the overfitting CNN is also worth mentioning. Especially Fully Connected Layer, which stores most of the information for the network, is susceptible to this treat. A special technique, called dropout, allows to stop sub-elements learning before overtraining the entire network.

5 ImageNet as a Source of Data for Convolutional Neural Network

Since the widespread use of digital documents and the availability of global data interchange (internet development), digital imaging specialists have worked to design more sophisticated algorithms for indexing, downloading, organizing, and commenting on multimedia data. CNNs learning to recognize specific objects in digital images requires a large number of digital images to be stored in a database which must be catalogued according to a specific hierarchy. This led to the idea of creating a large, indexed database of digital images - datasets. The most commonly used database for this purpose is the ImageNet [19] project, which is designed to conduct research to identify a variety of objects. ImageNet is a collection of digital images organized according to the hierarchy used in the WordNet dictionary. In this Princeton University dictionary, every significant term which is possible to be described by many words or phrases is called a “set of synonyms” or “synset”. Today there are more than 100,000 synonyms in the WordNet glossary, most of them nouns (80,000+).

Table 1. ImageNet database (April 30, 2010) [19].

Total number of non-empty synsets	21,841
Total number of images	14,197,122
Number of images with bounding box annotations	1,034,908
Number of synsets with SIFT features	1000
Number of images with SIFT features	1,200,000

Based on dependencies defined in WordNet, nearly 15 million URLs of digital images have been added to ImageNet by 2016, which were manually described to name the objects. Additionally, over a million digital images have bordered named objects. There has been created a giant set of digital images used by

scientists from around the world. Table 1 shows selected statistics of ImageNet database.

Currently ImageNet is widely used by scientists around the world in developing new methods in the field of CV, particularly as a training material for CNNs. Since 2010, ImageNet has been organizing the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where teams of researchers create competitive software in the field of valid classification and object and scene detection, based on ImageNet. The Contest is a natural successor to the Caltech 101 and PASCAL VOC projects, but on a much larger scale - for comparison, the PASCAL VOC image database in 2012 contained only about 21,738 images grouped in 20 classes.

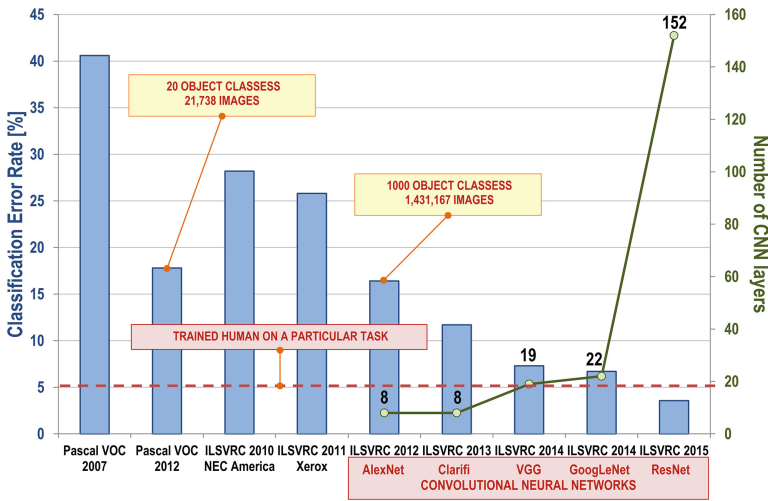


Fig. 3. Achievements in image classification. Based on: [17, 19–23].

With each new contest and the growth of the ImageNet database, there has been a sharp increase in the processing of digital images year after year (Fig. 3). In 2011, the ILSVRC rate of error was 25%. Compared to the winner of the Pascal VOC the average error rate was 40.65% in 2007 [22] and 17.8% in 2012 [23] respectively. In the following years, the development of digital image processing techniques for analysis and the use of CNNs have reduced error rates in subsequent years to a few percent. It was announced in 2015 [17] that the software outperformed human capabilities in narrow image analysis tasks from ImageNet. ResNet, the winner of the ILSVCR in 2015, achieved 4.8% during, and 3.57% after the competition.

Recognizing the positive results of the research that has been made with the ImageNet database, other similarly developed data sets have recently been launched. For example, DigitalGlobe has released a free of charge SpaceNet [24]

repository of high resolution satellite imagery to support the development of new remote sensing applications (automatic recognition and extraction of satellite imagery). ImageNet is currently the largest categorized database of digital images, but it contains images in low resolution. Data provided in the SpaceNet repository will primarily help in developing computer vision algorithms for automatic building detection, but the authors believe that they can also be used to identify technical objects (power lines, transport infrastructure, industrial networks). In the near future, SpaceNet's repository will offer far greater possibilities, with DigitalGlobe announcing it will include as many as 60 million satellite scenes.

As building large (massive) datasets is time consuming and costly, a number of work is being done to optimize these activities. Recent interest in Generative Adversarial Networks (GAN) has been raised in [15]. These are methods that could be used to create data sets that are similar to input data. In order to approximate the concept of the GAN, the distinction between generative models and discriminatory models must first be indicated.

The discriminant model in the learning process fits a function that assigns the input data (x) to the desired category (y). Under probabilistic circumstances, the conditional distribution $P(y|x)$ is directly taught. In this way, extended hierarchical models [13] are created that represent the probability distributions of data processed by deep CNN. The generative model tries to learn simultaneously the probability of a set of input and category data simultaneously, e.g. $P(x, y)$. This probability can be converted to $P(y|x)$, respectively, for assigning an object to a set, using the Bayes' theorem, but additionally generative capabilities can also be used to create new sets of data (x, y).

GAN implementation is based on two models of neural networks competing with each other. One model gets unclassified input and generates samples (generator). The second model (discriminator) receives samples from both the generator and the input learning data, and must be able to distinguish between the two data sources. These two networks play a continuous game of fixed sum, where one player's profit is the loss of the other. During the game the generator learns to produce more realistic samples, and the discriminator learns better and better distinguishes the generated data from the real data. The most important benefit of GAN is that we can provide information about the backpropagation gradient from the generator to discriminator network. Both types of models are useful, but generative models have one interesting advantage over discriminatory models - they can understand and explain the basic structure of input data even when they are not assigned to any category. This is particularly important when working on real-world data modeling, where uncategorized data is very extensive, and the acquisition of the described data is expensive and often impractical.

6 Selected Examples of Convolutional Neural Networks

LeNet

CNNs are a type of network that try imitate the human perception of the environment. It is the first stage in the processing of information that is received through the senses. The originator of the concept of processing information in a similar manner by machines was a scholar, Frank Rosenblatt, a psychologist who studied the processes of animal learning. He was also the creator of the first image recognition system which he called the perceptron. Minsky and Seymour Papert some years later showed limitations on perceptrons that temporarily inhibited work in this field [25]. Another important discovery was the development of neocognitron, which was developed as a handwriting recognition system [26]. The first significant CNN was the LeNet-5 network created by Yann LeCun. LeNet-5 had a multilayered structure and allowed the use of backward error propagation algorithm in the network learning process [10]. The network was designed to recognize handwriting and print and was able to correctly classify characters after tampering (inter alia translation, scale, rotation, squeezing, stroke). As input, a 32×32 pixel image was provided. The network structure consisted of 6 layers (3 convolutional layers, 2 subsampling layers, 1 fully connected layers). The network was learned using the MNIST dataset.

AlexNet

The AlexNet is the implementation that achieved the highest score in the ILSVRC test in 2012, obtaining a Top5 test error of 16.4% (Top5 error is an indicator of whether the search object was in the one out of 5 best-fit categories). The author of the network is Alex Krizhevsky, his network is based on 5 convolutional layers, max pooling, dropout, and 3 fully connected layers [14]. The network structure uses also the Rectified Linear Unit (ReLU) activation function. Input to the first convolutional layer is given in the form of a matrix of $224 \times 224 \times 3$. Filters used in the first layer have a dimension of 11×11 . The network has been pre-trained using the ImageNet database and allows classifying objects for 1000 possible classes. Network learning was done using the stochastic gradient descent method. The network has 650 K neurons and using 630 M connections generates 60 million parameters, so its authors have devised a dropout mechanism to remove some neurons and their connections during the learning process. Implementation of ReLU activation with dropout during the network learning process let one to skip the pre-training phase when very large amount of labeled data is available [27]. The AlexNet learning lasted 5 days with 2 GTX 580 graphics cards.

ZF NET

Another network to focus on ZF Net was developed by two researchers, Matthew Zeiler and Rob Fergus in 2013 and it get a Top5 error rating of 11.2% in ILSVRC.

The network has been developed on the basis of the AlexNet with several modifications. The main advantage is the reduction of the filter window to 7×7 , resulting in much more information concerning the original image and the use of the activation functions: cross-entropy and the loss for the error function. The network was taught using the GTX 580 for twelve days. The most interesting component of the network was the deconvolutional layer that could reverse the convolution process, resulting i.e. in the ability to visualize activation of neurons in intermediate layers. The deconvolutional network, using grouping and regularization, in a reversed way, allow to obtain the corresponding pixels to the input image [28]. The visualization mechanism, developed in ZF Net, could be used to preview how each layer works, and allows to introduce network architecture improvements.

VGG

The VGG Net was created in 2014 by the Visual Geometry Group team, which, unlike the previously mentioned networks, distinguishes the increased number of layers used. The network structure was originally developed in several variants ranging from 11 up to 19 layers. The aim of this approach, developed by Karen Simonyan and Andrew Zisserman, was the idea of to investigate how the depth of the network affects the results [29]. The main difference with respect to earlier networks is the use of 3×3 filters, in comparison AlexNet uses 11×11 and ZF Net 7×7 . Consecutive convolution of two 3×3 filter layers yields an effective 5×5 receptive field and a combination of three 3×3 convolution layers results in a effective 7×7 receptive field. This approach allows to maintain benefits of a large filter with the use of several smaller ones and to reduce the number of parameters to learn. The network was learned using the mini-batch gradient descent method.

GoogLeNet

GoogLeNet is the network that won the ILSVRC competition in 2014 and achieved 6.7% error rate in the Top5 category. The network developed by Google has a different structure than the previously discussed. It consists of 22 layers so the trend of deepening the network structure (shown in Fig. 3) is preserved. The increasing number of layers causes the raise in the number of parameters and could lead to overfitting and the computational cost growth. The GoogLeNet network generates up to 12 times less parameters than AlexNet, and while being significantly more accurate. Such optimization has been achieved by introducing a mechanism called the Inception Module, what is an important enhancement over the sequential approach of the remaining networks. In the previous implementations of the CNNs it was necessary to choose whether to perform a pooling or a convolution operation, and to determine the filter size. The Inception Module allows the parallel use of different size filters [18]. For example, it distinguishes general (5×5) and local (1×1) filters at the same time. Then the

concatenation of all results into a single vector, used as the next layer input, has to be performed. Overall, the GoogLeNet architecture comprises of a total of over 100 layers, 9 of which were the inception modules. The network structure is currently being developed under the codename Inception, with the most up-to-date version called Inception V4 [30].

ResNet

ResNet was developed by a Microsoft Research Asia team in 2015. It comprises of 152 layers and in the ILSVRC 2015 competition it got an error rate of 3.57%. This result is spectacular, it was the first time when AI achieve better results than human who in the image classification task, depending on their knowledge and experience, oscillate between 5 and 10%. The researchers while defining ResNet found out that the shallow structures are less error-prone during the learning process due to use of residual transition between layers [31]. The concept of residual transition assumes that after passing through several layers, the result is summed up with the input of the CNN. This results in faster learning process of the network. Another change from previous approaches is the abandonment of the fully connected layer to the global average pooling in order to simplify the relation between feature maps and categories, what makes the network results more meaningful and interpretable.

7 Summary and Conclusions

The progress in the classification of objects in digital images by CNN is significant, exceeding even human capabilities. However, it is important to remember that deep learning algorithms can classify only images belonging to strictly defined set of categories. People can recognize a much larger number of classes, create a new ones, while simultaneously being able to analyze the context of the scene [17]. Despite the significant development of CNN there are still some limitations to their common application. In order to achieve higher universality of current solutions the learning mechanisms has to be developed. Many ongoing CNN work is currently centered around the implementation of this tool for solving further Computer Vision tasks. Current editions of ILSVRC competition are focused on object detection, scene classification and scene parsing.

The authors now recognize the wide potential for the development of specialized applications of automated inspection methods, mainly for objects that have not yet become a part of training datasets. The CNNs seems to be suitable to use with technical facilities and equipment, such as power infrastructure. The monitoring systems in this sector already collect the vast amount of data and this brings the need of automated tools introduction to process them. Preparing datasets for new domains and advancing analysis mechanisms on its basis may be necessary. Deep learning mechanisms require very large sets of data, but their preparation requires a considerable effort, it is time and cost consuming, since the assignment of categories to specific objects is still performed manually.

Currently, the work is being done on optimizing these activities, such as the minimizing required learning sets size, while the maintaining the same effectiveness of the trained networks. In order to achieve this the current learning algorithms improvement is essential, for instance using GAN.

The present challenge for the CNN development is the proposal of their new structures (of higher number of layers or enhanced architectures). Network hyper-parameters such as: depth, the number of filters in each convolution, and their size, are correlated, and their fitting requires later verification. A lot of work is currently underway to optimize this process (inter alia [9, 16, 18, 29, 32]) and the authors anticipate the appearance new deep learning solutions.

References

1. Batra, S., Sachdeva, S.: Suitability of data models for electronic health records database. In: Srinivasa, S., Mehta, S. (eds.) BDA 2014. LNCS, vol. 8883, pp. 14–32. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13820-6_2
2. Bagloee, S.A., Tavana, M., Asadi, M., et al.: Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. *J. Mod. Transport.* **24**(4), 284–303 (2016). <https://doi.org/10.1007/s40534-016-0117-3>
3. Pal, S.K., Meher, S.K., Skowron, A.: Data science, big data and granular mining. *Pattern Recogn. Lett.* **67**(2), 109–112 (2015). <https://doi.org/10.1016/j.patrec.2015.08.001>
4. Häne, C., Sattler, T., Pollefeys, M.: Obstacle detection for self-driving cars using only monocular cameras and wheel odometry. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS. Hamburg (2015). <https://doi.org/10.1109/IROS.2015.7354095>
5. Salman, Y.D., Ku-Mahamud, K.R., Kamioka, E.: Distance measurement for self-driving cars using stereo camera. In: Proceedings of the 6th International Conference on Computing and Informatics, ICOCI 2017, Kuala Lumpur (2017)
6. Hohm, A., Lotz, F., Fochler, O., Lueke, S., Winner, H.: Automated Driving in Real Traffic: from Current Technical Approaches towards Architectural Perspectives. SAE Technical Paper (2014)
7. Karami, E., Prasad, S., Shehata, M.: Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images. In: Newfoundland Electrical and Computer Engineering Conference, IEEE, Newfoundland and Labrador Section At St. John's, NL (2015). <https://doi.org/10.13140/RG.2.1.1558.3762>
8. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Man, D.: Concrete Problems in AI Safety (2016). arxiv.org/abs/1606.06565
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *IEEE* **86**(11), 2278–2324 (1998)
11. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer, S.C., Kolen, J.F. (eds.) A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press, Hoboken (2001)
12. Hinton, G.E.: To recognize shapes, first learn to generate images. *Prog. Brain Res.* **165**, 535–547 (2007)

13. Bengio, Y.: Learning Deep Architectures for AI. Now Publishers, Boston (2009)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems (2012)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks (2014). arxiv.org/abs/1406.2661
16. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (2015). arxiv.org/abs/1502.01852
17. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015). <https://doi.org/10.1007/s11263-015-0816-y>
18. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
19. ImageNet Project. <http://image-net.org>
20. Cao, J., et al.: A parallel Adaboost-Backpropagation neural network for massive image dataset classification, *Sci. Rep.* **6**(38201) (2016). <https://doi.org/10.1038/srep38201>
21. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH (2014). <https://doi.org/10.1109/CVPR.2014.222>
22. Marszalek, M., Schmid, C., Harzallah, H., Weijer, J.: Learning object representations for visual object class recognition. In: Visual Recognition Challenge workshop, ICCV (2007)
23. Yan, S., Dong, J., Chen, Q., Song, Z., Pan, Y., Xia, W., Huang, Z., Hua, Y., Shen, S.: Generalized hierarchical matching for sub-category aware object classification. In: Visual Recognition Challenge workshop, ECCV (2012)
24. SpaceNet. <http://explore.digitalglobe.com/spacenet>
25. Papert, S., Minsky, M.: Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge (1988)
26. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980). <https://doi.org/10.1007/BF00344251>
27. Srivastava, N., et al.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
28. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
29. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-scale Image Recognition (2014). arxiv.org/abs/1409.1556
30. Szegedy, C., et al.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning (2016). arxiv.org/abs/1602.07261
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas (2016). <https://doi.org/10.1109/CVPR.2016.90>
32. Yong-Deok, K., Eunhyeok, P., Sungjoo, Y., Taelim, C., Lu, Y., Dongjun, S.: Compression of Deep Convolutional Neural Networks for Fast and Low Power Mobile Applications (2016). arxiv.org/abs/1511.06530