DIVING INTO DEEP LEARNING

Artificial intelligence (AI) technology has made rapid progress over the last few years. Breakthroughs in deep learning set new performance levels for various AI applications including speech recognition, language translation, recommendation and computer vision. It has enabled a first wave of successful consumer applications driven by companies like Google, Uber, Facebook and Netflix. Now a second AI wave is on the horizon, driven by industrial applications. This article gives an overview of the latest AI developments, the main techniques and their application to industrial and control engineering problems.

ALBERT VAN BREEMEN

Introduction

Applying AI to industrial and control engineering problems is not new. During the 70's and 80's expert systems became a popular AI technology to create industrial planning and diagnostics systems [1]. These systems take human expert knowledge and turn it into a database of if-then rules. The control software consists of a database with expert rules and some logic to process these rules ('inference logic'). This approach is particularly successful in domains where the problem cannot be modelled well using first principles and one has to rely on the intuition and knowledge of human experts.

During the early 90's a variation of expert systems using fuzzy logic became popular (see the box). The foundation of fuzzy set theory was laid in 1965 by Lofti Zadeh [1], but only during the late 80's it started to be applied to control engineering problems.

While rule-based AI systems provide a way to translate human expert knowledge into a (control) program, it is at the same time also limited by the knowledge of the human expert. During the mid-90's another AI technology became popular that tackled this problem. This AI technology is named artificial neural networks or just neural networks in short. A neural network is a machine learning AI algorithm that learns nonlinear relationships from data (Figure 1). It can be used as a generic building block to build nonlinear adaptive control systems [3]. Applications of neural network-based control systems include nonlinear feedforward control in mechatronic positioning systems and optimised setpoint control [4].

The last few years a new AI technology is rapidly maturing. This technology is called deep learning and it builds upon the foundation of neural networks. The remainder of this article explains this technology and investigates its application to manufacturing and control problems.

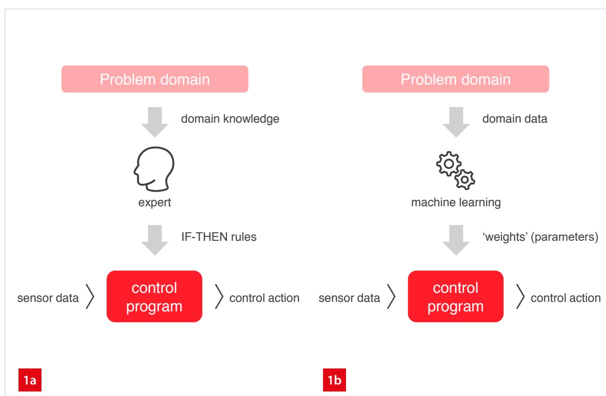
AUTHOR'S NOTE

Albert van Breemen is an independent AI & machine learning consultant with over 20 years of experience in high-tech technology and innovation. He worked for companies including Philips and ASML.

info@aiblog.nl
www.aiblog.nl

Fuzzy logic

While expert systems are based on binary logic (yes, no), fuzzy logic uses truth values between 0 and 1. For example, a human expert might state a knowledge rule like IF temperate-is-high THEN set-heater-low. With binary logic the predicate temperate-is-high is defined by thresholding the temperature sensor data at a specific value. Fuzzy logic, however, defines a range of temperature values, each value being 'high', but with a different truth value. Fuzzy logic provides an intuitive method to design control systems for complex nonlinear processes that is robust to sensor data uncertainty. The technology was popular in Asia and several fuzzy logic-based controllers where used in consumer and industrial applications, such as rice cookers, washing machines and cement mills [2].

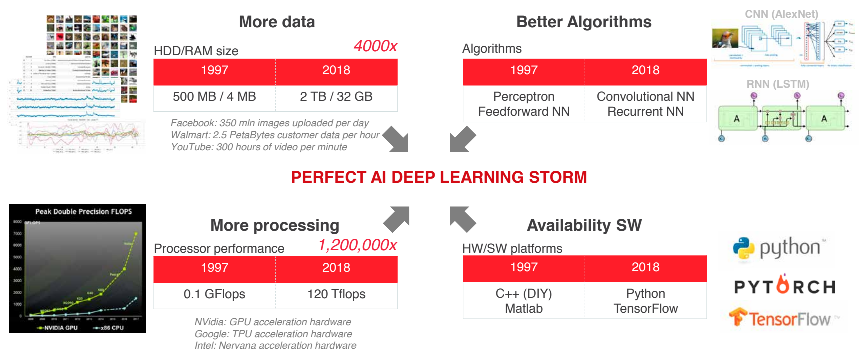


Expert-based design. *Machine learning-based design.*

Deep learning

Deep learning became popular after the 2012 edition of the yearly held 'ImageNet Large Scale Visual Recognition Challenge' (ILSVRC) [5]. During this challenge teams compete to develop the best program to detect objects in images. Up to 2012 the mainstream approach to build these programs was to use human expert knowledge and tuning. Alex Krizhevsky and his team, however, used a neural network with massively more parameters (62 million) than normally used.

As training on a CPU processor would take too long, they used GPUs instead. A GPU has massively more computational cores than a CPU, with each core being able to perform basic multiply-and-addition operations of floating point numbers. As this is the core calculation of a neural network (see further), Krizhevsky et al. could speed up their training significantly.



2

Perfect AI deep learning storm due to four technological developments.

In the end, they trained a winning neural network on two Nvidia GTX 580 GPUs (1,500 Gflops per GPU) in five to six days. Their solution achieved a 15.4% error rate, which was 10.8 percentage points better than the best system the previous year, and thus won the 2012 ILSVRC. Since then, other teams adopted this approach of using large neural networks, large datasets and GPU training. This approach is now called deep learning and has achieved impressive results in domains such as playing Go [7] and autonomous driving [8].

A perfect AI storm

Currently the field of AI is experiencing a perfect storm. Four technological developments accelerate the development of deep learning, since its conception in 2012 (Figure 2). These include:

1. Data storage & generation:
   Storage capacity in computer systems increased a factor of ~4,000 over the last 20 years. Simultaneously, cheap sensors and Internet connectivity fuel the creation of massive amounts of data; big data. Currently, big datasets of any domain, such as medical, retail, engineering or manufacturing, are freely available on websites such as *Kaggle.com* and *openml.org*, or are present within companies. Deep learning requires large datasets to train models and there is no technology bottleneck anymore for gathering and storing them.
2. Computational power:
   Processor power increased a factor of 1,200,000 over the last 20 years. Companies such as Nvidia and Google have been developing AI-optimised processors such as GPUs (graphical processing unit) and TPUs (tensor processing unit) [9]. These AI optimised processors give a performance boost of 10 to 50 times compared to CPUs.
3. AI algorithms:
   New types of neural networks and machine learning methods have been developed that more efficiently handle large amounts of data. The next section discusses the details of some of those algorithms.
4. Open-source software:
   The availability of open-source software has dramatically shortened the time to develop AI applications. Nowadays anybody has access to major AI software platforms, such as Python/Scikit-Learn [10], TensorFlow [11] or PyTorch [12]. Experimental results can be more easily shared and compared, which speeds up developments.

Deep learning

Deep learning can be classified into four different types of algorithms, each optimised to handle specific data types and problems.

Artificial neural networks

The main technique behind deep learning is artificial neural networks (ANNs). An ANN is a model that is able to learn nonlinear relations between inputs and outputs. The basic building block of an ANN is a node called perceptron, which is a mathematical model of a biological neuron (see the box on the next page) and was formulated in 1943 by McCulloch & Pitts (referenced in [13]).

Mathematically, a perceptron is defined by the following equation:

$$= f(\sum_{i=1}^{n} w_i \cdot x_i + b)$$

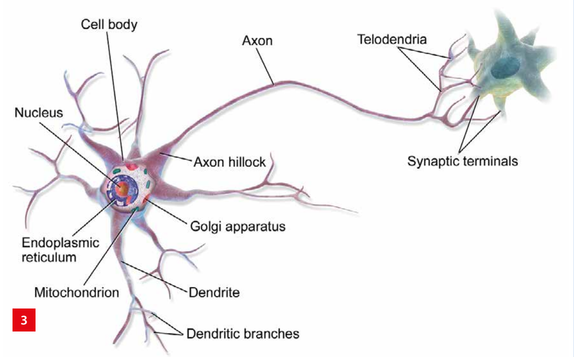
with:

$$f(z) = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}$$

Its output y is a function f of the summation of weighted inputs x_i and a bias. The parameters w_i and b (called 'weights' and 'bias') are determined by using a dataset and

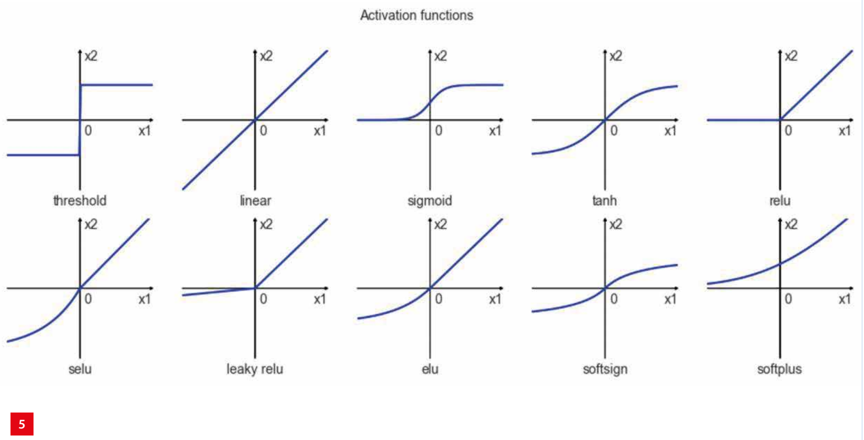
Biological neuron

A biological neuron (Figure 3) consists of a cell body, dendrites and axon. Via the dendrites a neuron receives signals from its neighbouring neurons. These signals are added up by the cell body of the neuron. If the total of received signals is higher than some threshold the neuron will fire a signal via its axon to other neighbouring neurons. Not every signal received is as important as other signals. Some signals even inhibit the firing of a neuron. McCulloch & Pitts [13] modelled this by using weights.



3

Animation of a (multipolar) neuron [14].



5

Different types of activation functions of a perceptron.

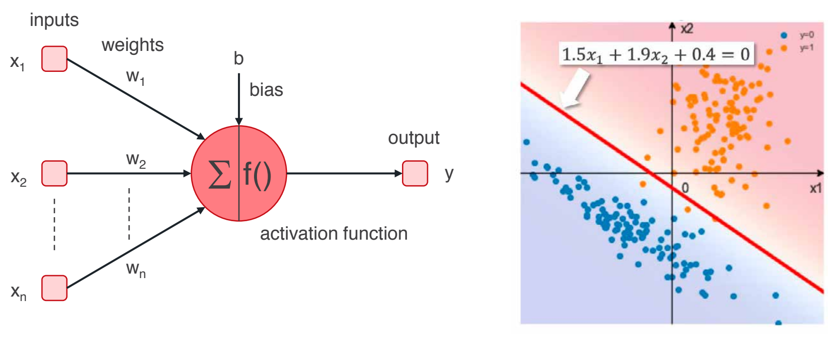
an optimisation method that minimises the prediction error of the perceptron. This is called 'training a neural network'.

One way to interpret the perceptron equation is that it divides a data space linearly into two classes (see Figure 4). The function *f* is called the activation function of the perceptron. Over the years many activation functions have been proposed (see Figure 5) and choosing the best one for a specific problem is one of the ANN design challenges.

A perceptron in itself has limited computational strength. Real-world datasets often are nonlinear and data points from classes are spread over several clusters in the data space (Figure 6). By combining many perceptrons in a layer-like topology (see Figure 7), a model is created that is able to learn more complex functions for problems such as regression and classification.

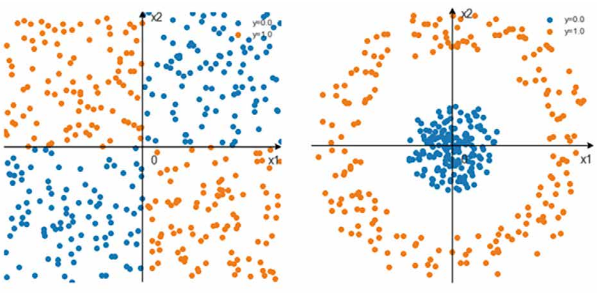
Convolutional neural networks

Processing images is useful for many applications, including quality control, security, autonomous driving, and augmented reality. While images can be fed into an ANN directly, it would result in neural networks with many inputs and parameters, as every pixel becomes one input of the ANN. A more efficient way is to use neural networks with so-called convolutional layers. A neural network layer consists of a bank of image filters, each having a limited size of 3 x 3 or 5 x 5 parameters. Neural networks where the first layers are convolutional layers, are called convolutional neural networks (CNNs); see Figure 8. Compared to an ANN, a CNN has far less parameters to train.



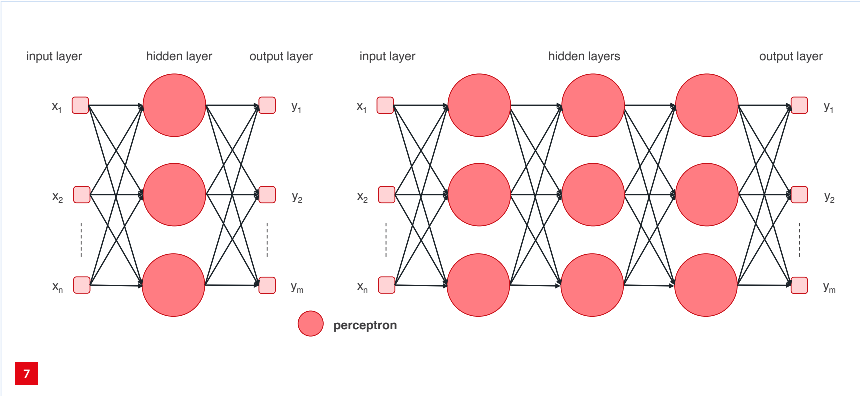
4

Perceptron and an example of a linear separation of a 2D data space.

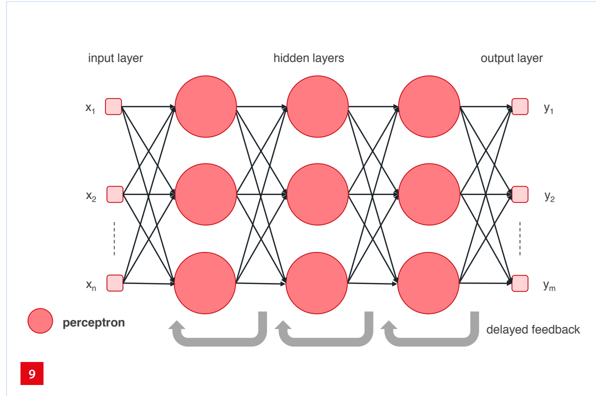


6

Two examples of artificial datasets that cannot be separated by a single line and cannot be learned by a single perceptron.



Multi-layer perceptron. *Deep neural network.*

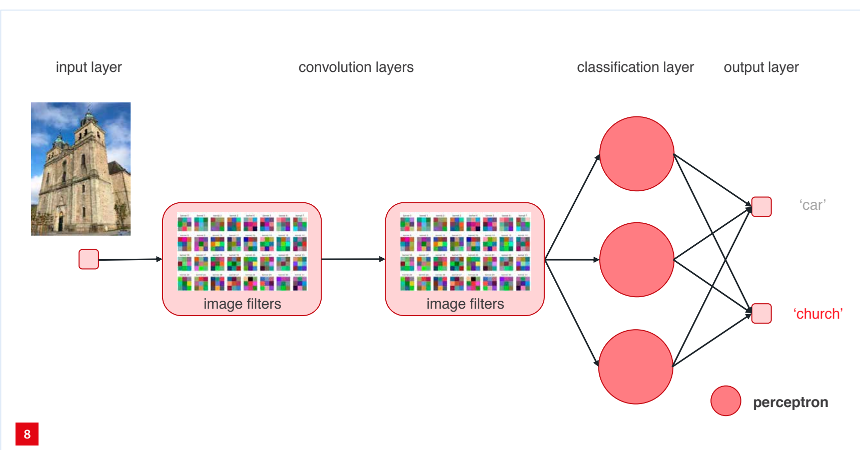


Recurrent neural network.

The more convolutional layers are added to a CNN, the more complex image processing the network can carry out. Over the last few years many different CNN architectures have been developed with names such as AlexNet, ResNet, Inception, and VCC [15]. These networks have been used to solve complex object detection and image segmentation problems. Training the parameters of these networks requires massive amount of data and processing capacity. Without hardware acceleration (i.e. by using AI-algorithm-optimised hardware), training the CNNs becomes impractical.

Recurrent neural networks

A third category of deep learning models are recurrent neural networks (RNNs); Figure 9. These networks differ from ANNs and CNNs in that they feedback a delayed output of nodes in layers as inputs to nodes in previous layers. By doing so, the network becomes 'dynamic' and gets interesting properties to handle time-series data. RNNs can be used for, e.g., speech recognition, video analysis, EEG signal processing, dynamic system identification and stock market analysis.



Convolutional neural network.

Although the concept of RNNs was already known in the 90's, no good training method for the weights was available. Because of the feedback loop the training could become unstable, resulting in either parameters vanishing to zero or exploding to infinitely large values. New RNN models called long short-term memory (LSTM) and gated recurrent unit (GRU) have been developed to overcome this problem.

Reinforcement learning

The last deep learning technique discussed here is reinforcement learning (RL) [17]. RL is a technique to determine a sequence of optimal actions for some problem. Where ANNs, CNNs and RNNs require a labelled dataset to do the training, RL uses a single reward signal at the end of some sequence of actions. This type of problem, where a performance signal is only received after some time (control) actions have been carried out, occurs frequently. Think of dynamic machine control, marketing campaigns, process optimisation and game play.

By combining RL with ANN, CNN or RNN powerful learning models can be developed. This has been illustrated by, a.o., the company DeepMind [18], which developed an AI system that beat the ruling world Go champion.

Applications

Industrial AI

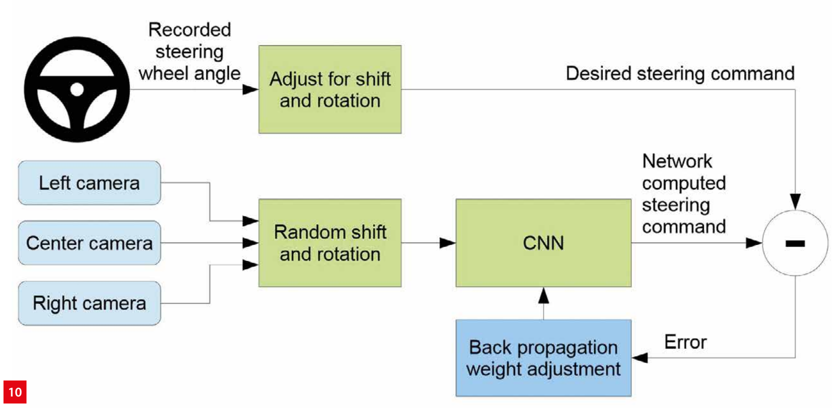
Industrial equipment and manufacturing processes generate huge amounts of data from sensors for monitoring, control and optimisation. Using deep learning AI algorithms could lead to new performance levels and applications to improve product quality, maximise uptime, optimise yield and lower cost of labour, energy and material usage. This has been recognised by some major industrial companies like GE, Bosch, Fanuc, Kuka and Siemens, which are all investing significantly in AI technology riding the wave of 'Industry 4.0', 'smart manufacturing', 'Internet of Things' and more.

Table 1
Comparison of consumer vs industrial AI applications.

Consumer applications	Industrial applications
Meaning of data often clear (e.g. seeing a movie, liking a news item).	Data often noisy and their meaning less clear.
AI runs on the cloud without hard timing deadlines.	AI runs on the edge and real-time responses are expected.
False negatives/positives do not lead to disasters.	Prediction mistakes could lead to unsafe situations.
Predictions may only cost less then €0.001.	High predictions cost €10 - 1,000 as often much more is at stake.
Consumers do not ask why a recommendation was made.	Complex models must be interpretable.

Applying AI technology to industrial problems is more challenging than applying it to consumer applications (see Table 1). Noisy data, real-time performance and safety are among the challenges to deal with in an industrial application. Therefore, industry uses models that can be interpreted by a human expert. While this is possible with rule-based expert systems, deep learning models are more abstract and difficult to interpret. A mix of in-depth engineering domain knowledge and AI knowledge will help in gradually introducing deep learning technology in industrial applications.

Smart cameras
Developing smart cameras for industrial applications is currently one of the low-hanging deep learning fruits. Using CNNs for this purpose has been well researched lately and impressive results have been obtained. This technology is now ready to be used for developing applications in domains such as agriculture, wind mill inspection, medical image analysis and waste inspection.



10

An application of CNNs for autonomous driving as described by Nvidia [19].

Besides using CNNs to develop smart object detection systems, they can also be used to develop advanced end-to-end control systems. Nvidia describes an application of CNNs for autonomous driving and trained a CNN for driving a car autonomously (Figure 10). To train the CNN a dataset was created by recording the video footage from front cameras and the steering actions of a driver. With this end-to-end deep learning approach they succeeded in their application [19]. The same approach was also used to create an autonomous flying drone [20].

Maintenance
Another area where CNNs are used is equipment maintenance. By training CNNs to identify machine parts and their health condition augmented reality maintenance systems can be built. Such systems use augmented reality to present relevant information over an image of a machine for the maintenance engineer.

Machine uptime is an important factor determining the overall profit of a manufacturing process. Unexpected machine downtime due to part failure directly leads to additional costs. Predictive maintenance is a method to predict machine part failures before they occur. When a failure is expected within a period of time, the part can be replaced during regular maintenance service within that period.

Digital twin
The concept of a digital twin is often discussed in this context [21]. A digital twin is a digital copy of an asset, system or process. It is used for asset monitoring, predictive maintenance and planning by simulation. A digital twin consists of sensors, data gathering, data storage, data analysis and visualisation. Because very large numbers of digital twins are often managed by a manufacturer, they are implemented using scalable cloud services, such as provided by Amazon Web Services.

The quality of a digital twin depends among others on how well the future behaviour of an asset can be predicted from the data. Deep learning recurrent neural networks are a promising deep learning technique to use, because of the time-series nature of the data to deal with. While traditional approaches depend on manual feature engineering and domain expertise, RNNs automatically extract the right features from time-series data.

Siemens applied the digital twin concept to the application of predictive maintenance of gas turbines. For every turbine data from over 500 sensors are collected. Using AI and virtual reality, engineers can remotely monitor the asset and prevent problems early on [22].

Optimisation

Reducing material and energy usage in manufacturing processes is always an important challenge to tackle, in particular with climate change on the agenda of many companies nowadays. Deep learning models can learn the complex interactions between (parts of) machines and use this knowledge for optimisation.

Google applied deep learning to reduce energy consumption in its data centres. This is a challenging control problem because interaction between the various components is so complex that an intuitive understanding cannot be built. Furthermore, internal and external (like the weather) conditions change so quickly that rules for every scenario could not be derived. Finally, as every data centre has a different architecture a generic solution cannot be used. An energy consumption reduction of 40% was realised by training ANNs on historical data centre operating data. These ANN models could be used to calculate recommended actions to optimise the power usage efficiency [23].

Robotics is another area where deep learning is being explored. Programming manufacturing robots to carry out pick & place operations is a difficult and time-consuming task. Recently, Fanuc showed that this task can be sped up and simplified by using deep RL. Using this AI technique, robots learn by themselves from trial & error to pick & place parts. Within eight hours of learning the robot has achieved a similar performance level as if it where programmed by a human expert [24].

Conclusion

New developments in the area of deep learning are being used in manufacturing and control applications. Major companies including GE, Siemens, Bosch, Kuka and Fanuc are all investing in this technology, with the aim to reduce cost, improve quality and create new applications.

While access to AI algorithms, hardware and software is open to everybody, successful industrial AI applications depend on having domain-specific datasets and talent that understands both AI and manufacturing and control engineering. The latter is a rare breed because both AI and engineering are a specialism requiring deep knowledge. The way to go forwards then is that engineers team up with AI specialists, so that in a next AI-themed *Mikroniek* issue high-end AI applications for mechatronic control can be presented in detail.

REFERENCES

[1] Russell, S., & Norvig, P., *Artificial Intelligence: A Modern Approach*, 3rd edition, 2015.

[2] Singh, H., et al, "Real-Life Applications of Fuzzy Logic", *Advances in Fuzzy Systems*, Volume 2013.

[3] Van Breemen, A.J.N., "Neural Adaptive Control", 1997, *aiblog.nl/neural-adaptive-control*

[4] Van Breemen, A.J.N., & De Vries, T.J.A., "Design and implementation of a room ", *Control Engineering Practice*, 9(3):233–248, 2001.

[5] Imagenet Large Scale Visual Recognition Challenge (ILSVRC), *image-net.org/challenges/LSVRC*

[6] Krizhevski, A., et al, "ImageNet Classification with Deep Convolutional Neural Networks", *papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf*

[7] DeepMind, *deepmind.com/research/alphago*

[8] Nvidia autonomous driving, *devblogs.nvidia.com/deep-learning-self-driving-cars*

[9] Google TPU, *cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu*

[10] Scikit-Learn, *scikit-learn.org*

[11] Tensorflow, *www.tensorflow.org*

[12] PyTorch, *pytorch.org*

[13] Haykin, S., *Neural Networks – A Comprehensive Foundation*, Pearson Prentice Hall, 1999.

[14] Blaus, B., Blausen.com staff (2014), "Medical gallery of Blausen Medical 2014", *WikiJournal of Medicine* 1 (2). *doi:10.15347/wjm/2014.010*, ISSN 2002-4436.

[15] *software.intel.com/en-us/articles/hands-on-ai-part-16-modern-deep-neural-network-architectures-for-image-classification*

[16] *www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge*

[17] Sutton, R.S., & Barto, A.G., *Reinforcement Learning*, MIT Press, 2nd edition, 2018.

[18] *deepmind.com*

[19] Bojarski, M., et al., "End to end learning for self-driving cars", *arxiv.org/abs/1604.07316*, 2016.

[20] *spectrum.ieee.org/automaton/robotics/drones/drone-uses-ai-and-11500-crashes-to-learn-how-to-fly*

[21] GE, digital twin, *www.ge.com/digital/applications/digital-twin*

[22] *www.siemens.com/innovation/en/home/pictures-of-the-future/digitalization-and-software/simulation-and-virtual-reality-simulations-gas-turbines.html*

[23] *deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40*

[24] *www.technologyreview.com/s/601045/this-factory-robot-learns-a-new-job-overnight*