An Introduction to Deep Learning: Part II

LASSE AMUNDSEN, HONGBO ZHOU, Statoil, and MARTIN LANDRØ

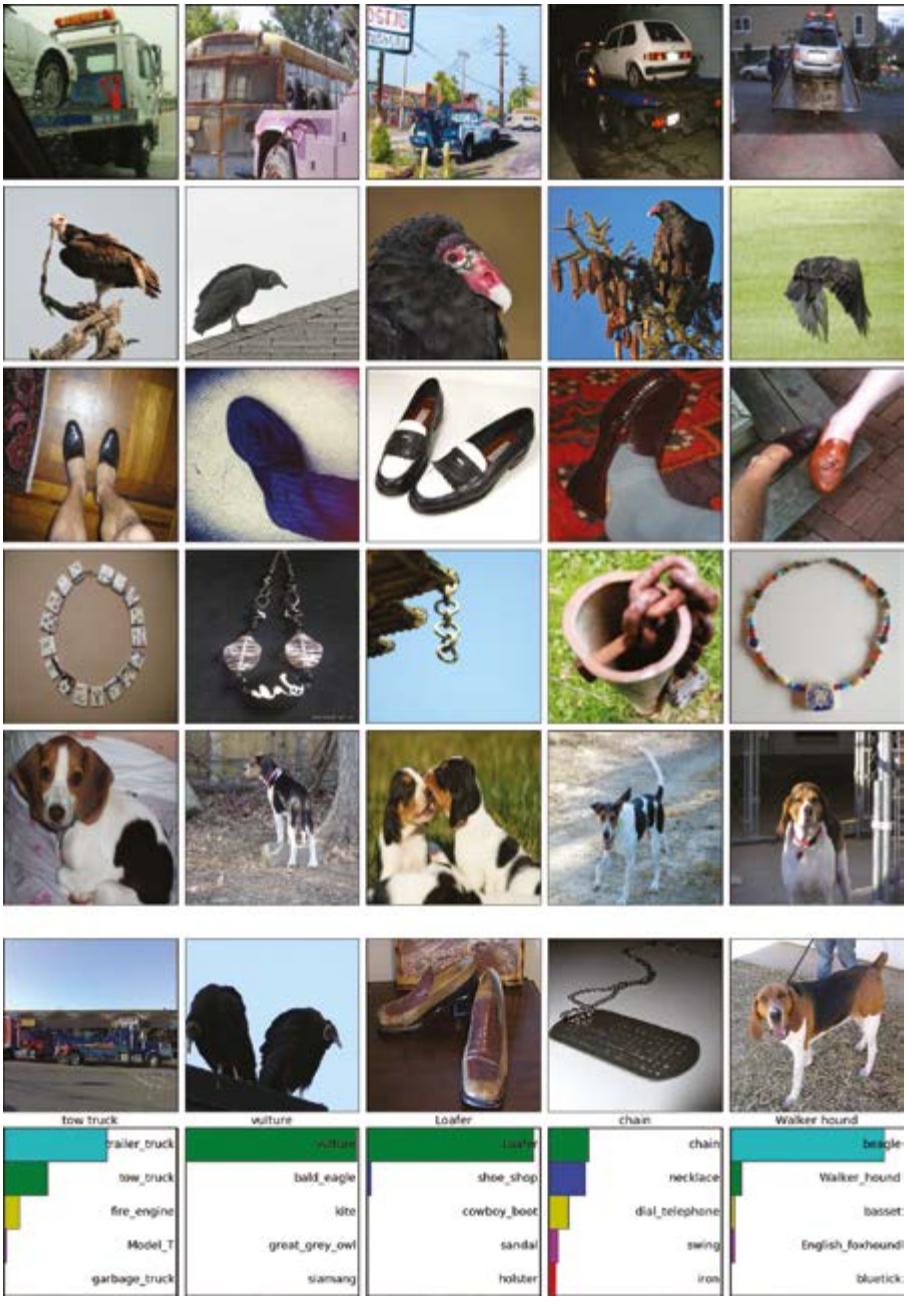
"We need to go deeper."

Leonardo DiCaprio, in the film Inception (2010). A thief, who steals corporate secrets through the use of dream-sharing technology, is given the inverse task of planting an idea into the mind of a CEO.

Well said, Leo. In Part II of this article, we deepen our discussion on deep learning – a tool for implementing modern machine learning.

Over the past few years, AI has accelerated. Factors that have helped renew progress in AI are faster, cheaper and more powerful computers. Progress came also with Big Data, with exponential growth, availability of data, and growing understanding of the potential value of such data – images, text, mapping data, and so on. With these computing breakthroughs, neural networks were revisited, and they could be made huge.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC), set up to encourage computer-vision breakthroughs, is the world's top computer vision contest. To compare models, ImageNet examines how often the model fails to predict the correct answer in their top five guesses (the top-5 error rate), in descending order of confidence. ILSVRC 2012 brought a small research group led by Geoffrey Hinton at the University of Toronto to



The ImageNet dataset contains 15 million labelled images of objects in around 22,000 categories. ILSVRC, the 'Olympics of computer vision', is an annual competition which uses a subset of ImageNet – roughly 1,000 images in each of 1,000 categories.
The top five rows show five ILSVRC-2012 classes of images for network training and validation. The bottom two rows show the corresponding five test images and their top-5 labels considered most probable by using a variant of the AlexNet model (illustration made for the purposes of demonstration for this article). The correct label is written under each image. We see in the last column that the network found the walker hound incorrectly to be a beagle. The beagle and the walker hound however have similar 'tricolour' coloration. But the net indeed labelled it correctly at the second 'guess'.

everyone's attention. The group had parallelised its convolutional neural network AlexNet on Nvidia GPUs, and won the contest by getting an error rate of 16.4% with the given data for the top-5 guesses, while the error rate of the second-best system was around 26%. ILSVRC 2012 marks the turning point for neural nets research, heralding the abandonment of feature engineering and the adoption of feature learning in the form of deep learning. Since then, remarkable progress has been made by the community and the pinnacle was reached by Microsoft in 2015 when it achieved a top-5 error rate of 3.57%.

Deep Learning

Deep learning is nothing but a rebranded name for a family of deep neural networks – complex mathematical systems that can learn tasks by analysing vast amounts of data. Deep learning thus is a class of learning procedures based on the neural network model. It has facilitated image recognition, object detection, video labelling, and activity recognition, and is making significant progress into other areas of perception, such as audio, speech, language translation and natural language processing (NLP). According to Schmidhuber (2015), Rina Dechter (1986) introduced the expression 'Deep Learning' to the machine learning community at the conference of the Association for the Advancement of Artificial Intelligence (AAAI). Later, it became widely accepted when Igor Aizenberg et al. (2000) introduced it to ANNs. (see Part 1).

Deep learning can circumvent the challenges of feature engineering that are critical for symbolic-based machine learning. The remarkable thing about deep learning is that no human is required to program a computer because the deep learning models are capable of learning the features automatically by themselves. Therefore, programmers just need to feed the computer a learning algorithm, expose it to terabytes of input data to train it, and then allow the computer to figure out itself how to recognise the desired objects. In short, such computers can now learn by themselves. This makes deep learning an extremely powerful tool for modern machine learning. Deep learning methods are beating traditional symbolic-based machine learning approaches on virtually every metric.

Go is a strategy board game for two players, in which the aim is to surround more territory than the opponent. Go is played on a grid of black lines (usually 19×19). Game pieces, the black and white stones, are placed on the lines' intersections.

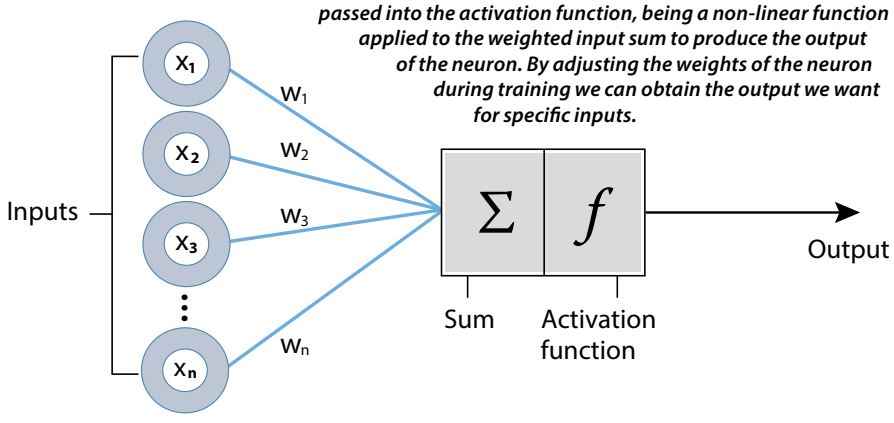
Further evidence that deep learning is on the rise is the amount of capital being invested, the number of people who are choosing it as their area of study, and the number of leading technology companies that are making AI the core of their strategic plans. It is revolutionising many areas of machine perception, with the potential to impact people's everyday experiences. Some even believe that AI could be used to mimic human common sense someday.

Great Achievements

One of the landmark achievements of deep learning is Google DeepMind's AlphaGo beating the world legendary Go champion Lee Sedol four games to one in 2016. The game of Go was invented in China more than 2,500 years ago and is believed to be the oldest board game still played today. Its simple rules and deep strategies have intrigued everyone from emperors to peasants for generations. The goal is to gain more territory than the opponent, but it is very complex and possesses more possibilities than the total number of atoms in the universe. The AlphaGo computer program uses deep neural networks, reinforcement learning and a Monte Carlo tree search to find its moves based on knowledge previously learned by an artificial neural network through extensive training, both from human and computer.

Deep learning can be trained with supervised learning, unsupervised

Model of an artificial neuron. Suppose that the neuron connects with n other neurons and so receives n-many inputs (x1, x2 ,… xn). The inputs are individually weighted, summed together and passed into the activation function, being a non-linear function applied to the weighted input sum to produce the output of the neuron. By adjusting the weights of the neuron during training we can obtain the output we want for specific inputs.

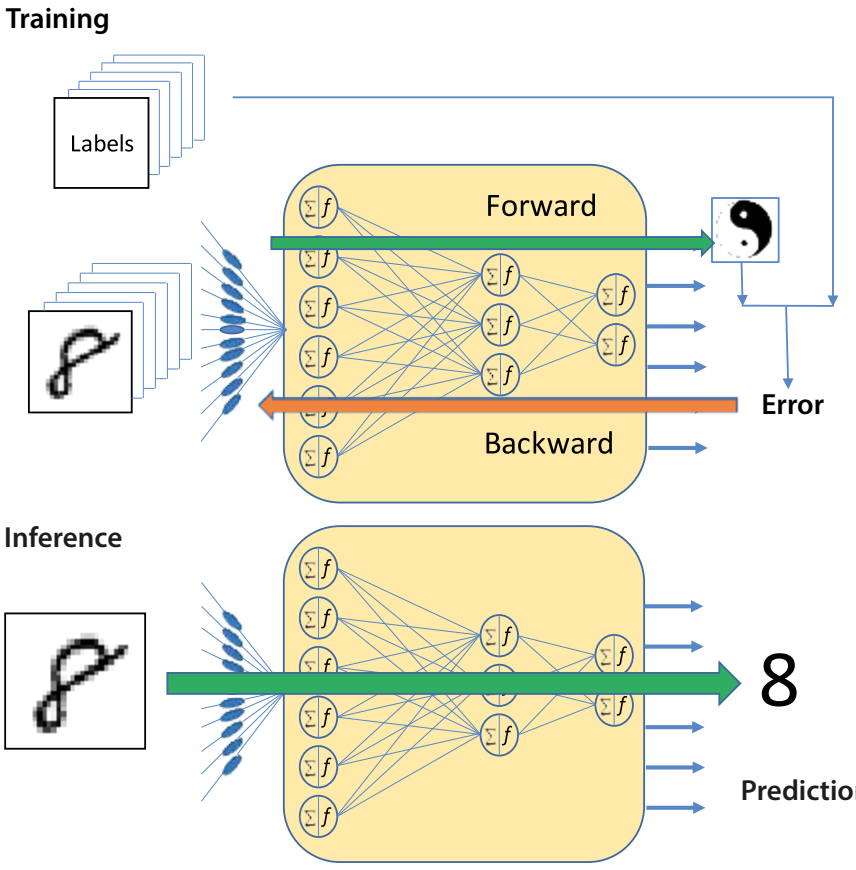


learning, or reinforcement learning. Unsupervised learning is the ultimate goal for the future and reinforcement learning is gaining more ground, especially in gaming and robotics. But supervised learning is the champion today: it works by showing the network a bunch of things with labels saying what they are, and getting the network to learn and classify future things without labels.

Deep neural networks generally work as a two-stage process. First, a neural network is trained with its parameters determined using labelled examples of inputs and desired output and then the network is deployed to run inference, using its previously trained parameters, to classify, recognise, and process new inputs. This is illustrated in deep forward neural nets on the right. When receiving an input image, the network translates it into a hierarchical level of features, and the neurons in each layer of the network are tuned to recognise certain patterns in the features. Low-level neurons recognise things like edges or basic shapes, then pass the data to the next layer. This layer of neurons does its own task, and passes processed data on. Neurons in high-level layers can 'see' objects – say, a cat or a dog. Each layer communicates forward with the one next to it, and as information travels down the network, some feature extraction processes take place automatically. At the end, the network comes up with an output – a prediction of what is in the image.

Return to the Wrong Way sign example in Part I: the image is split into a number of tiles that are inputted into the first layer of the neural network. The neurons examine each tile's attributes: for the Wrong Way sign, its rectangular shape, red colour, eight letters, and its size. Each neuron assigns a weighting to its input, where the weight tells how correct or incorrect it is relative to the task being performed. As data are passed forwards, the neural network's task is to predict with some probability, based on the total of the weightings, whether the sign is Wrong Way or not. Perhaps the network is 90% confident the image is a Wrong Way sign, 7% confident it is a Bicycle Wrong Way sign, and 2% confident it is a Danish flag on a flagpole, and so on.

While the neural network is being trained, the odds are in favour of it predicting incorrectly. Therefore, it needs lots of training, using millions of images, until the weightings of the neuron inputs are tuned so precisely that it gets the answer right almost every time. At that point the neural network has taught itself what a Wrong Way sign looks like.



Deep learning training and inference. In training (top), many inputs, often in large batches, are used to train a deep neural network. In inference (bottom), the trained network is used to discover information within new inputs that are fed through the network in smaller batches.

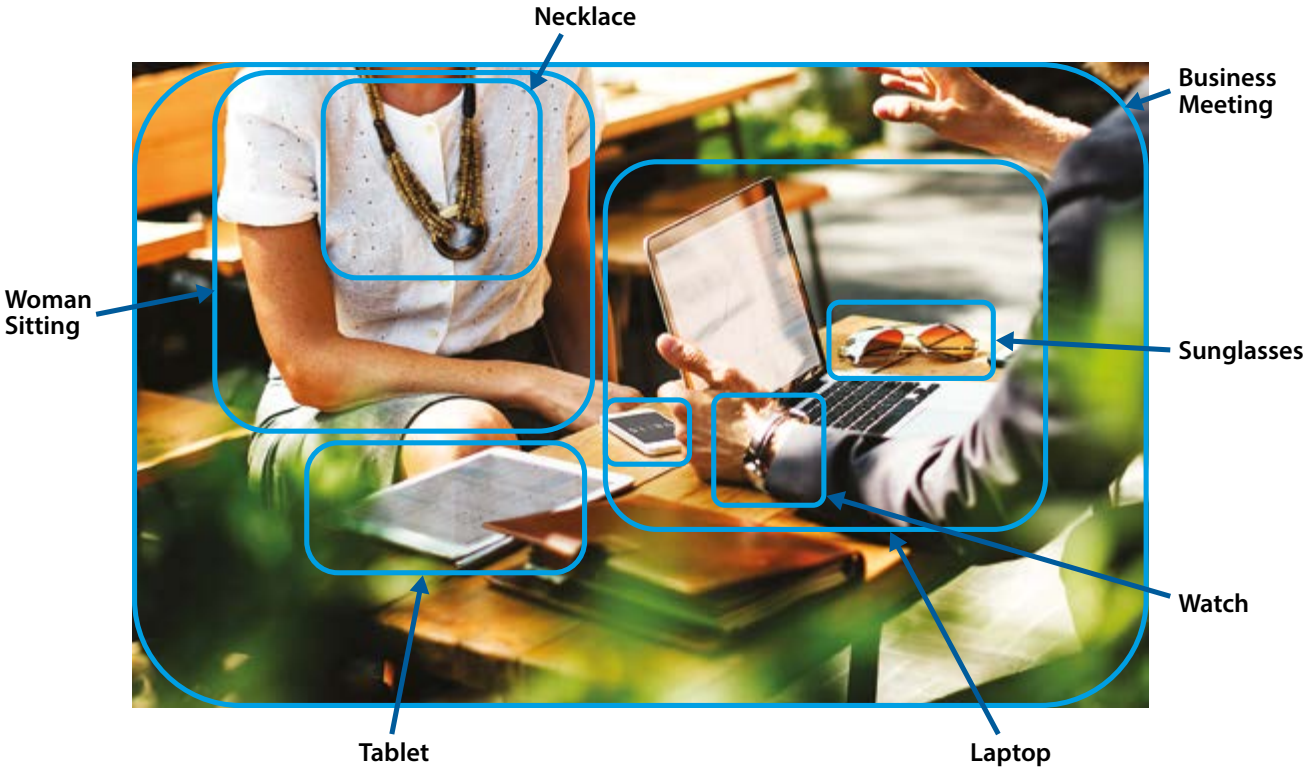
Classification, Localisation, Detection, and Segmentation

The examples in the introductory image address image classification, which is the task of taking an input image and outputting its class (cat, dog, etc.) or a probability of classes that best describes the image. It works well when the image contains only one object.

For humans, a quick glance at an image is sufficient to point out and describe an immense amount of details about the visual scene. When we look at an image we are immediately able to characterise the objects and give each a label. These skills at quickly recognising patterns, generalising from prior knowledge, and adapting to different image environments are ones that computers do not easily share with us. However, the success of the AlexNet in 2012 spurred research and great achievements in object localisation, detection and segmentation.

Object localisation not only produces a class label but also a bounding box that describes where the object is in the image. In the task of object detection, localisation needs to be done on all of the objects in the picture, resulting in multiple bounding boxes and multiple class labels. Finally, we also have object segmentation where the task is to output a class label as well as an outline of every object in the input image.

We end by referring the reader to Karpathy and Li (2015), who combine convolutional neural networks (CNNs) and bidirectional Recurrent Neural Networks (RNNs) to generate natural language descriptions of different image regions. Basically, their model is able to take in an image, and output a concept, as demonstrated in the image on the next page. ■



With conventional CNNs, a single label is associated with each image in the training data. Karpathy and Li (2015) presented a model that generates natural language descriptions of images and their regions. Their model has training examples that have a sentence associated with each image. This type of label is called a weak label, where segments of the sentence refer to (unknown) parts of the image. Using this training data, a deep neural network 'infers the latent alignment between segments of the sentences and the region that they describe'. Another neural net takes in the image as input and generates a description in text. The illustration above is a conceptual example of such an output.