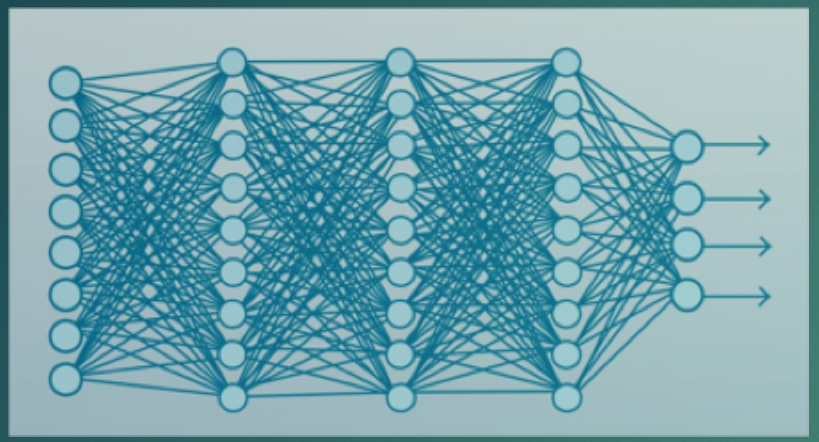


Deep Learning

YANN LECUN, YOSHUA BENGIO, GEOFFREY HINTON

"**DEEP LEARNING**." NATURE 521.7553 (2015): 436-444

RESEARCH REVIEW BY WILLIAM STEIMEL

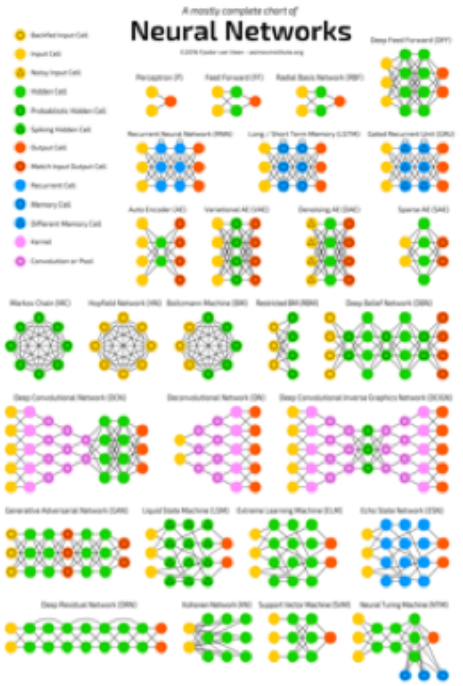
Table of Contents

Motivation

- I am interested in understanding the basics of Deep Learning
- This survey paper was written in 2015 which is relatively new
- Nowadays Deep Learning is considered to be cutting edge and extremely powerful for Machine Learning
- I found this article in a curated list of Deep Learning Papers from beginner to expert.
  - https://github.com/floodsung/Deep-Learning-Papers-Reading-Roadmap
- This paper is recommended as the first paper to read related to Deep Learning

Abstract

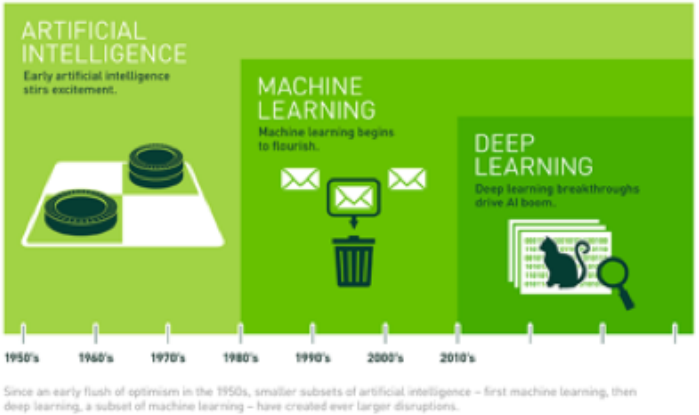
- "Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. "
- Deep Learning has allowed for great improvement in a variety of problem areas including:
  - Speech Recognition
  - Visual Object Recognition
  - Object Detection
  - Drug Discovery
  - Genomics
- Deep Learning utilizes Backpropagation: which tells a Neural network how to optimize its internal parameters/weights until the difference between errors and target are minimized to create the best representation of a function
- Convolutional Neural Networks have been showing promise in datasets related to Speech, Video, Audio, and Images
- Recurrent Neural Networks have been helpful in sequential data like text and speech
- This paper provides a good background on the applications and basics of Deep Learning



Source: Asimov Insitute

Introduction

- Machine Learning powers many aspects of modern technology including:
  - Web Search
  - Content Filtering on Social Networks
  - Recommendations on e-commerce websites
  - Smart Phones/Cameras
- Machine Learning is also used for:
  - Object Identification in images
  - Speech to Text Transcription
  - Recommendation of news items or posts with user interests
  - Relevant results for search
- Many of these recent technologies utilize Deep Learning which is a subset of Machine Learning



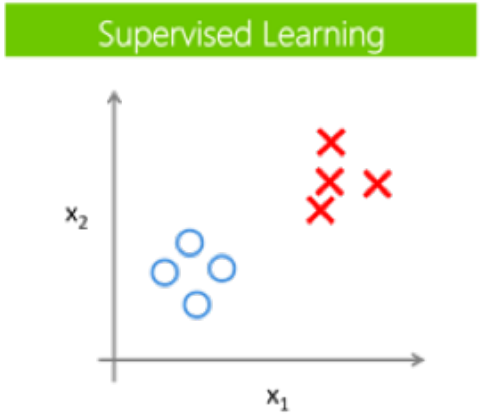
Source: Nvidia

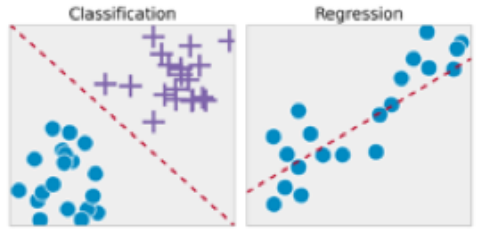
Introduction

- " Conventional ML techniques were limited in their ability to process natural data in their raw form."
  - This means that constructing a pattern recognition/machine learning system required a lot of manual feature engineering and domain expertise
- Representation Learning- a method that allows a machine to learn via the data its fed to discover representations needed for classification or detection.
  - This relates to Deep Learning as they use multiple levels of representation by composing simple but non-linear modules that allows for representation at a higher more abstract level
  - This helps with learning more complex functions
- Deep Learning is recently making many advances in solving problems that have escaped the Artificial Intelligence/Machine Learning community for many years
- Additionally, it is able to handle High Dimensional Data and has many applications in all industries including:
  - Science
  - Business
  - Government
- Recently, Deep Learning has beat many records in Image Recognition and Speech Recognition and has produced hopeful results in a variety of domains.
- The Future of Deep Learning seems to be great as it requires very little engineering, increasing available computing power, as well as new learning algorithms and architectures that are being developed for deep neural networks

Supervised Learning

- Supervised Learning is the most common form of Machine Learning:
  - The training data you feed the algorithm includes the desired solutions/labels
  - Classification/Regression Problems
- Generally when training Machine Learning Models, the goal is to optimize for Cost/Objective Function
  - (Reduction of Errors averaged over all training examples)
- In Deep Learning there are often hundreds of millions of weights (parameters/knobs) and hundreds of millions of labelled examples to train the model

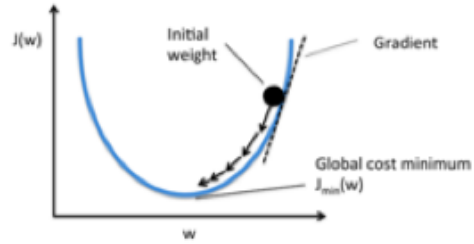




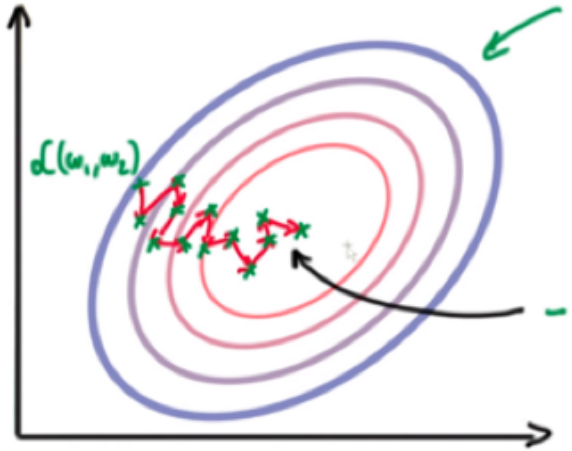
Source: Wikipedia

Supervised Learning

- Objective Function or Cost Function – Can largely be seen as a hill where the minimum is at the bottom.
- Most practitioners use a procedure called Stochastic Gradient Descent (SGD) to find the optimal weights.
- Stochastic Gradient Descent Steps:
  - Show the input vector a few examples
  - Compute the output and the errors
  - Computing the average gradient for those examples
  - Adjusting the weights accordingly
  - The process is repeated on the training set until the objective function stops decreasing



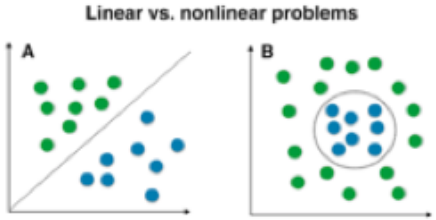
Source: Quora



Source: https://deeplearning4j.org/updater

Supervised Learning

- Many current applications of Machine Learning use linear classifiers and hand-engineered features
- Linear Classifiers – Only classify input space into very similar regions
  - However- Image and speech recognition are susceptible to variations in position, orientation or illumination of an object or variations in pitch/accent of speech
  - A linear classifier could not distinguish against many of these features which is why a good feature extractor is required
- "A deep-learning architecture is a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input–output mappings."
  - Deep Learning through the use of multiple non-linear layers can implement complex functions of its inputs that are sensitive to details and insensitive to irrelevant variations like background, pose, lighting, and surrounding objects

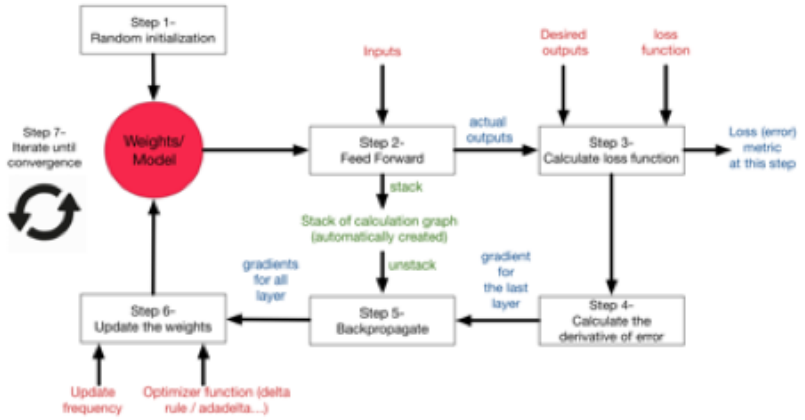


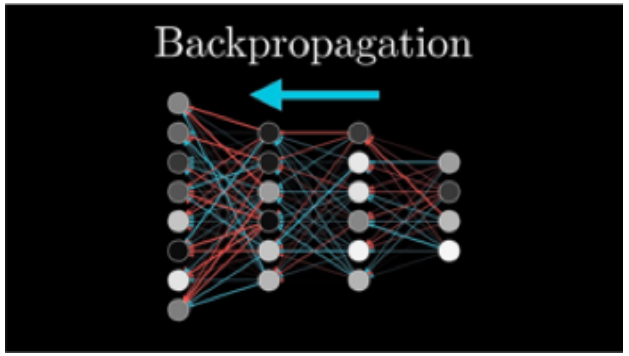
Source: https://i.stack.imgur.com/OrcTJ.png

Backpropagation to train multilayer architectures

- As mentioned previously, Backpropagation tells a Neural network how to optimize its internal parameters/weights until the difference between errors and target are minimized.
- Backpropagation for computing weights to minimize a cost function is simply the chain rule for derivatives.

Neural Network Learning Process

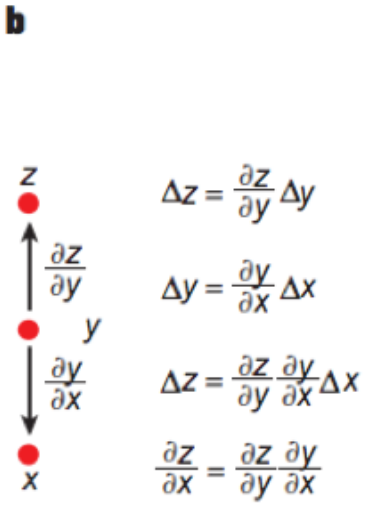




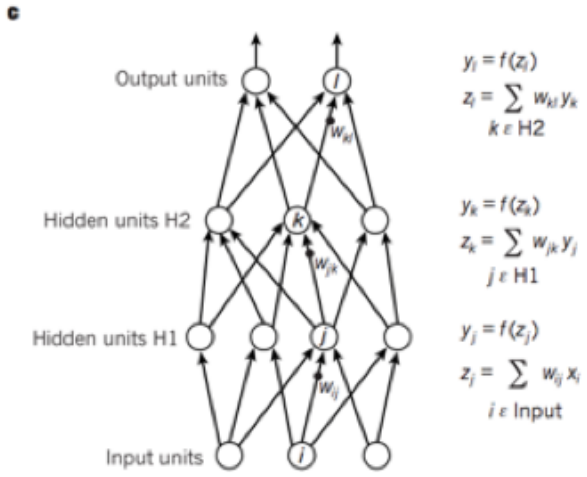
Source: https://datathings.com/blog/post/neuralnet/

Figure 1- Multilayer Neural Networks and Backpropagation

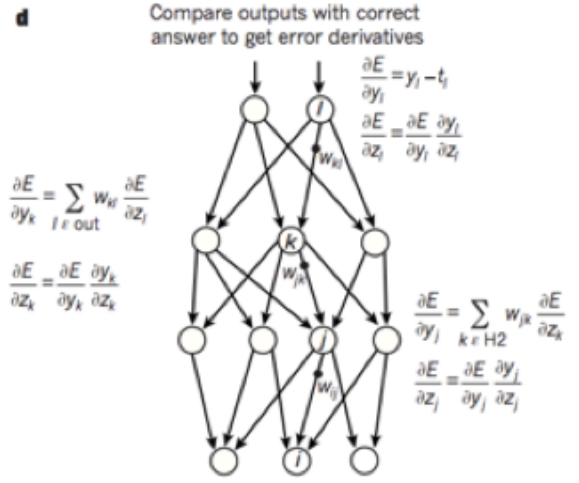
Chain Rule for Derivatives



Forward Propagation



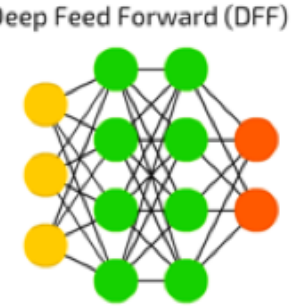
Backpropagation



(Image source: LeCun, Bengio, and Hinton, 2015; Fig.1)

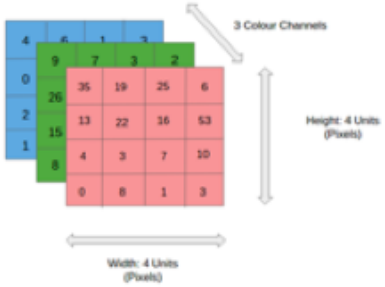
Backpropagation to train multilayer architectures

- In the 1990s – neural nets were largely ignored by the machine learning, computer vision, and speech—recognition communities.
- 2006 – Interest in deep feedforward networks was revived by researchers at Canadian Institute for Advanced Research (CIFAR)
  - Researchers created unsupervised learning methods that could create layers of feature detectors without requiring labelled data
  - This worked well with recognizing handwritten digits and detecting pedestrians despite the limited amount of labelled data
- One type of deep, feedforward network that achieved success when Neural networks were out of favor was the Convolutional Neural Network (ConvNet)
  - It was also much easier to train and generalize and has recently been used by the computer vision community



Source: Asimov Insitute

Convolutional Neural Networks



- The paper then summarizes Convolutional Neural Networks including the types of data they process, their architecture, and practical applications.
- Convolutional neural networks are designed to process data that come in the form of multiple arrays
  - Ex: Color image composed of three 2 D arrays with pixel values for three color channels
- ConvNets Architecture – A series of stages:
  - Convolutional layers – Images organized in feature maps, Convolution is a process used to extract features from an input image
  - Relu – Non-Linear Activation Function – It is used to introduce non-linearity into a neural network and outputs a rectifier map
  - Pooling layers – The Pooling Layer then takes the rectified feature map and uses a max, sum, or average function over the defined patch size.
- Convolution, Relu, and Pooling are all vital to CNN architecture and function in conjunction to extract features to help produce a classification result.
- Backpropagation functions similarly through a conv net as normal neural nets to find optimal weights
- Numerous Applications of Convolutional Neural Networks going back to the 90s
  - Time-Delay neural networks for Speech Recognition/document reading
  - Character Recognition/handwriting recognition (Microsoft)
  - Experiments with object detection/facial recognition

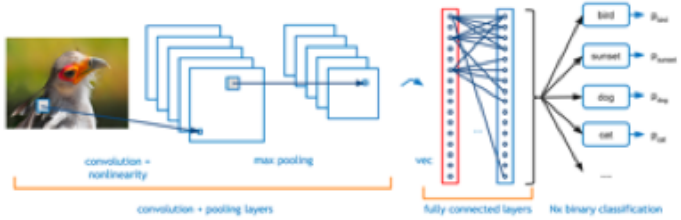
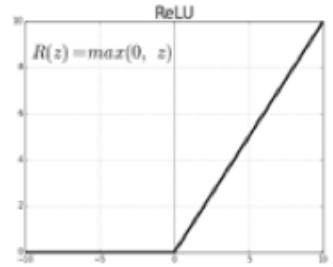


Image understanding with deep convolutional networks

- Since the early 2000s, ConvNets have been very successful at object segmentation/recognition
  - Traffic Sign Recognition
  - Face Recognition
  - Segmentation of Biological Images
  - Other Areas include natural language understanding and speech recognition
- What Started the Computer Vision Revolution ?
  - ImageNet Competition in 2012
  - Deep Convolutional Neural Networks were applied and achieved results halving the error rates of the best approaches at the time
  - This was due to: stronger GPUs, ReLUs (Activation Function), a regularization technique called dropout, and techniques to generate more training examples by deforming existing ones
- Training of recent ConvNet architectures originally took weeks but now can be done in a few hours with new advances in hardware, software, and algorithm parallelization
- The success of ConvNets has lead to the adaption of them through technology companies like Google, Facebook, Microsoft, IBM, Yahoo as well as startups and others
  - Self Driving Cars
  - Facial Recognition



Dropout

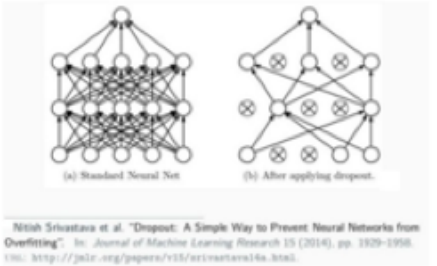
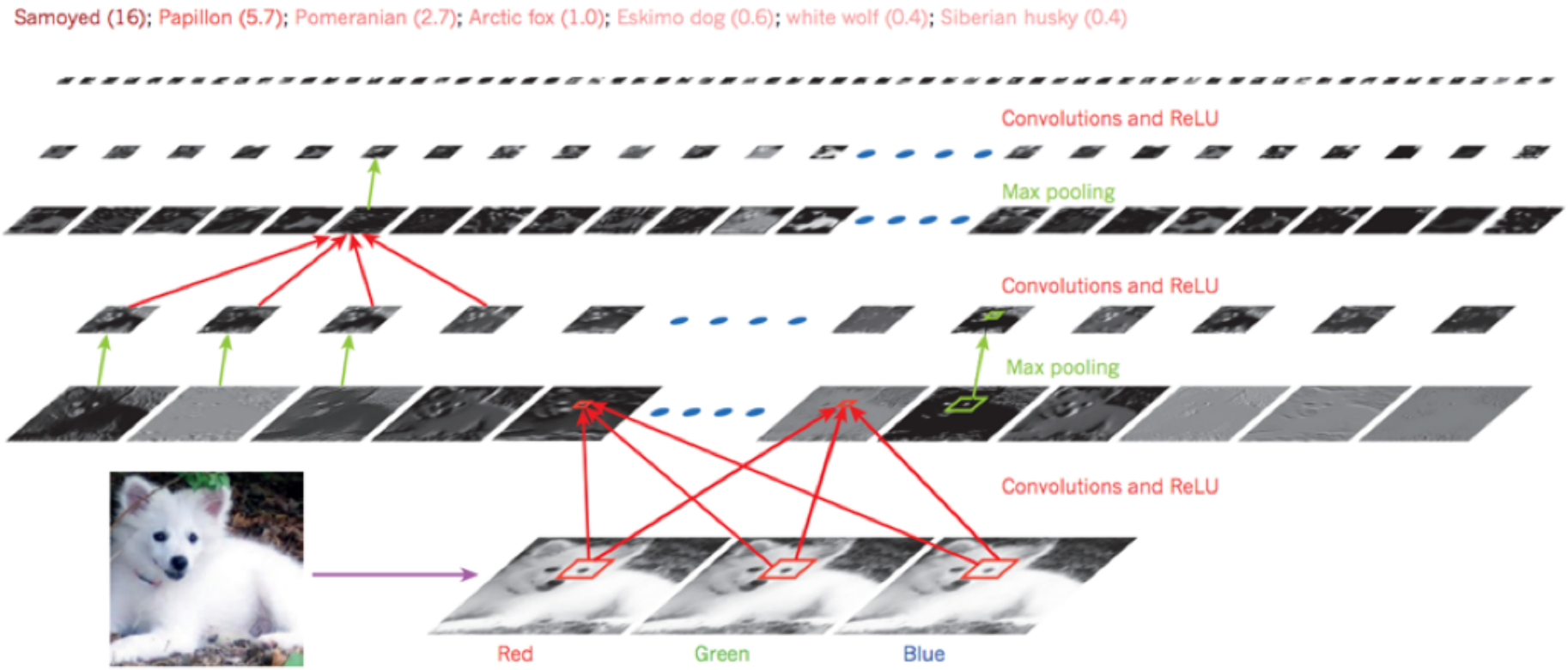


Figure 2- Inside a Convolutional Network



(Image source: LeCun, Bengio, and Hinton, 2015; Fig.2)

Distributed representations and language processing

- This section discusses the advantages of Neural Networks in Natural Language Processing and how they compare to traditional methods
- Deep Learning has two advantages over many traditional learning algorithms who do not use distributed representations
  - Learning distributed representations allows for generalization of values seen beyond training
  - Layers of representation – more depth
- Traditional NLP before Neural Networks did not exploit distributions and often used short symbol sequences called N-grams
  - "N-grams treat each word as an atomic unit, so they cannot generalize across semantically related sequences of words, whereas neural language models can because they associate each word with a vector of real valued features, and semantically related words end up close to each other in that vector space "

Figure 4- Visualizing the learned word vectors

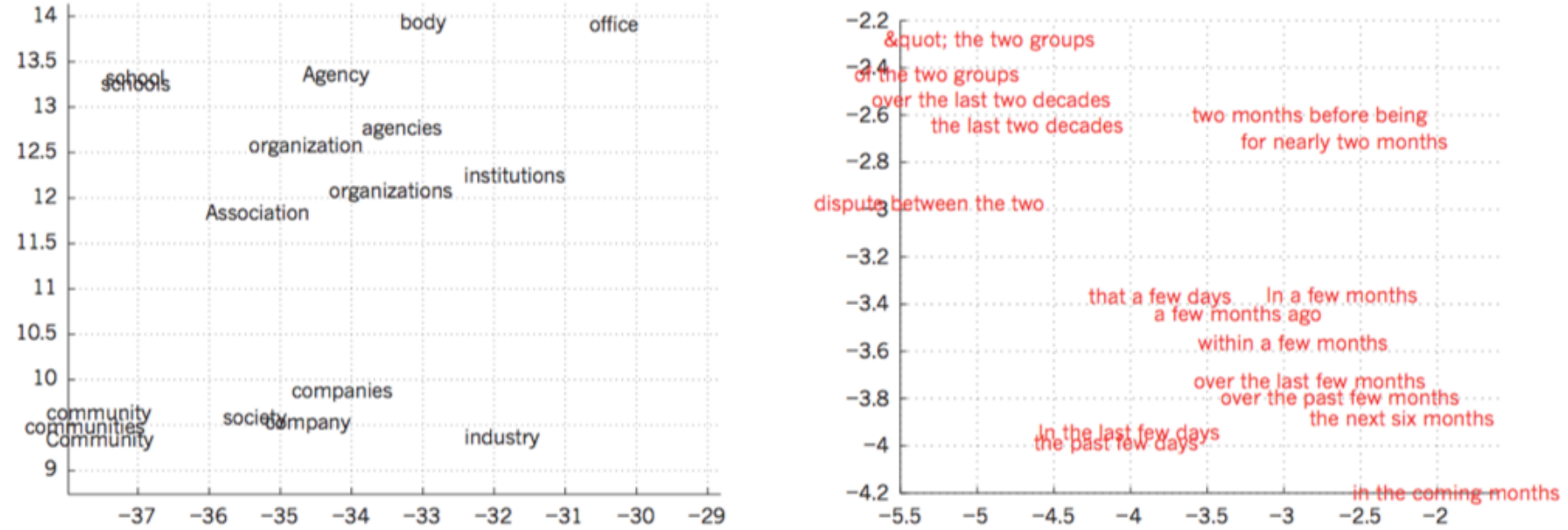


Figure 4 | Visualizing the learned word vectors. On the left is an illustration of word representations learned for modelling language, non-linearly projected to 2D for visualization using the t-SNE algorithm[103]. On the right is a 2D representation of phrases learned by an English-to-French encoder–decoder recurrent neural network[75]. One can observe that semantically similar words or sequences of words are mapped to nearby representations. The distributed representations of words are obtained by using backpropagation to jointly learn a representation for each word and a function that predicts a target quantity such as the next word in a sequence (for language modelling) or a whole sequence of translated words (for machine translation)[18,75].

(Image source: LeCun, Bengio, and Hinton, 2015; Fig.4)

Recurrent neural networks

- The next section introduces the applications of Recurrent Neural Networks
- RNN's are powerful at tasks that require sequential inputs:
  - Speech
  - Language
- "RNNs process an input sequence one element at a time, maintaining in their hidden units a 'state vector' that implicitly contains information about the history of all the past elements of the sequence."
  - Predict next word, next character
- Advances in architecture have lead to very good predictions in next character in text/ next word in sequence – they can also be used for more complex tasks
  - English to French Machine Translation
  - RNN's have also been used in translation of images into English Sentences (Figure 3)

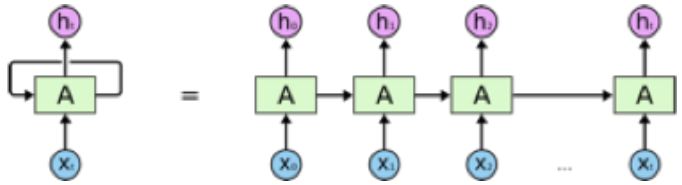
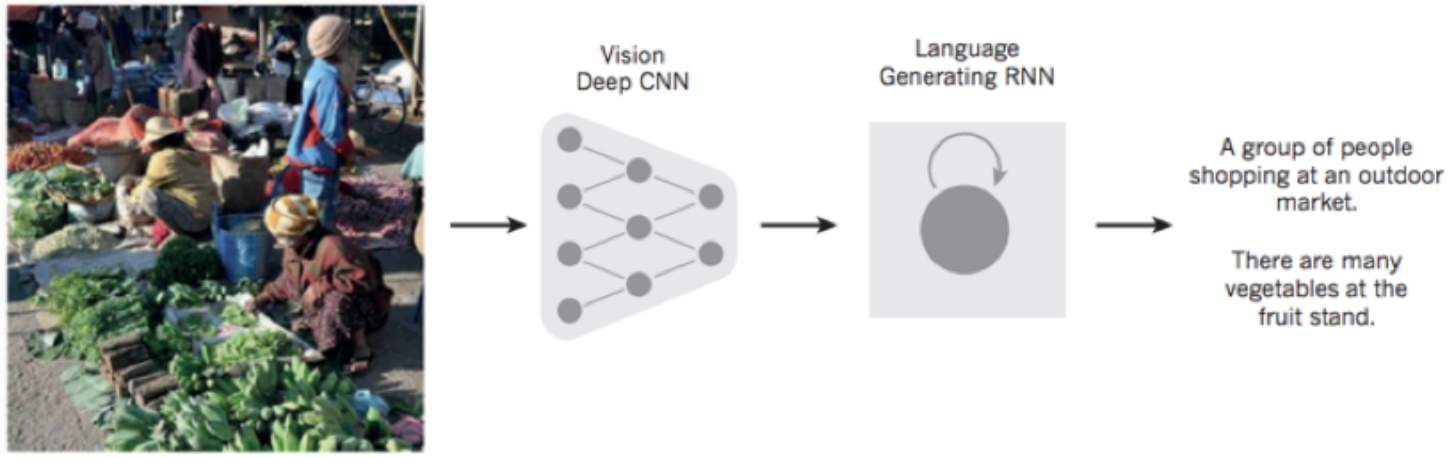


Figure 3- From Image to Text



(Image source: LeCun, Bengio, and Hinton, 2015; Fig.3)

Recurrent neural networks

- Weakness of RNN's – The authors state that although their main purpose is to learn long-term dependencies it is difficult for them to learn to store information for long.
- Long Short-Term Memory (LSTM) - Have proven to be more effective than conventional RNN's in this regard
- Although the research does not go to in-detail on LSTM architecture they are designed naturally to learn inputs for a long time
- LSTM is commonly used for encoder and decoder networks that perform well at machine translation
  - Machine translation
  - Speech recognition/acoustic recognition
- Other proposals to augment RNN's have included Neural Turing Machines and memory networks

Figure 5- A Recurrent Neural Network unfolding in time of the computation involved in its forward computation

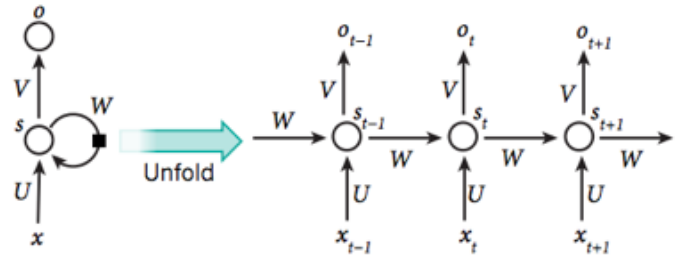


Figure 5 | A recurrent neural network and the unfolding in time of the computation involved in its forward computation. The artificial neurons (for example, hidden units grouped under node s with values s_t at time t) get inputs from other neurons at previous time steps (this is represented with the black square, representing a delay of one time step, on the left). In this way, a recurrent neural network can map an input sequence with elements x_t into an output sequence with elements o_t, with each o_t depending on all the previous $x_{t'}$ (for $t' \leq t$). The same parameters (matrices U, V, W) are used at each time step. Many other architectures are possible, including a variant in which the network can generate a sequence of outputs (for example, words), each of which is used as inputs for the next time step. The backpropagation algorithm (Fig. 1) can be directly applied to the computational graph of the unfolded network on the right, to compute the derivative of a total error (for example, the log-probability of generating the right sequence of outputs) with respect to all the states s_t and all the parameters.

(Image source: LeCun, Bengio, and Hinton, 2015; Fig.5)

The Future of Deep Learning

- Unsupervised Learning- Although unsupervised learning was largely not discussed in this research review its likely that it will become far more important in the long term in regards to deep learning.
- Human Vision- Authors assert that much of the progress will come from a combination of Convolutional Neural Networks with RNNs that use reinforcement learning
  - Systems that combine Deep Learning and Reinforcement are still new but already produce very good results in things like learning to play different video games.
- Natural Language Processing- Deep Learning will likely have huge impacts on NLP in the future
  - Author's expect RNN's to be able to better understand sentences and whole documents
- Artificial Intelligence- The authors assert that major progress will be made in Artificial Intelligence with systems that combine representation learning and complex reasoning

Additional Resources

- Amazing Applications of Deep Learning
  - http://www.yaronhadad.com/deep-learning-most-amazing-applications/
- History of Neural Nets
  - http://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/
- CNN's
  - https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/
- What is a Neural Network?
  - https://www.youtube.com/watch?v=aircAruvnKk&t=1019s

Discussion

- What do you think of the future of Deep Learning/Neural Networks?
- Which applications of Deep Learning/Neural Networks do you find interesting?
  - Vision, Natural Language Processing, Music, Art. Etc.

ありがとうございます！