KAIST Summer Session 2018

Module 3. Deep Learning with PyTorch

Deep Learning 101

KAIST College of Business

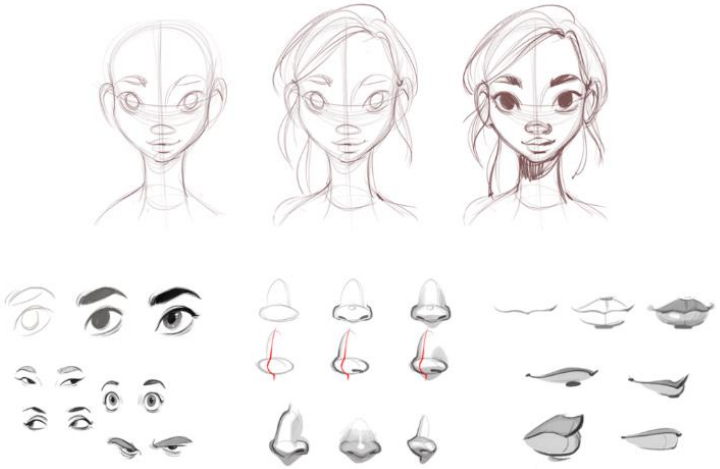
Jiyong Park

6 August, 2018

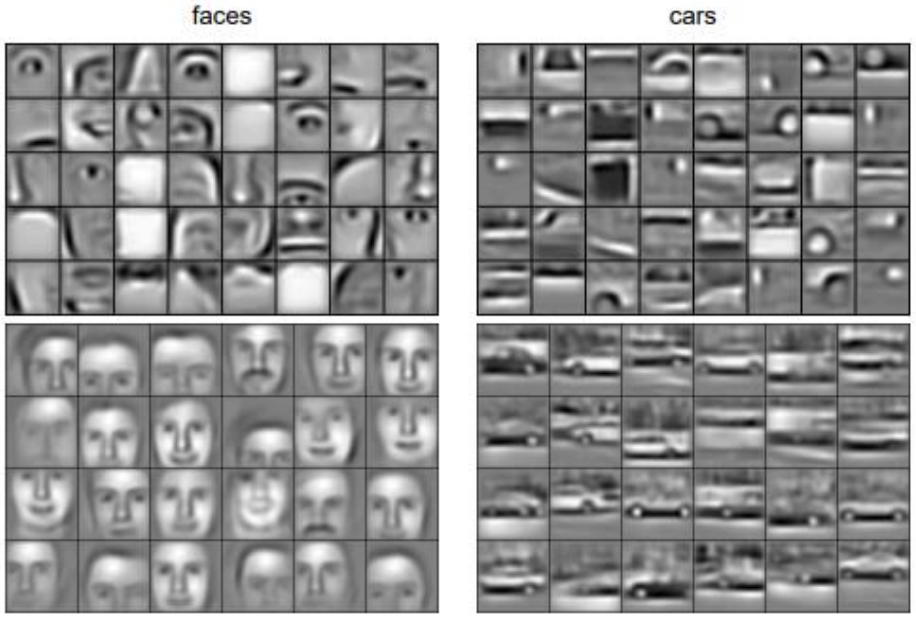
Intuitive Understanding of Deep Learning

- What is deep learning?
  - Deep learning is "**Representation Learning**" to learn and discover how to represent features of data. (Here, representation means a machine-understandable format)

How to draw faces?



How to represent faces?



Intuitive Understanding of Deep Learning

- Why important to well-represent the features of data?

- What needs to well-represent the features of data?

- Then, is deep learning well-suited for data representation?
  - How does it overcome the obstacles?
  - Why is deep learning outstanding at representation learning?

- But, is deep learning a magic bullet?
  - What are the potential limitations?

Structure of This Lecture

- Why important to well-represent the features of data?

- What needs to well-represent the features of data?

- Then, is deep learning well-suited for data representation?
  - How does it overcome the obstacles?
  - Why is deep learning outstanding at representation learning?

- But, is deep learning a magic bullet?
  - What are the potential limitations?

Module 1 – Applications (Computational Social Science & Economining)

Ch1. Artificial Neural Networks

Ch2. History of Deep Learning

Ch3. Unsupervised Deep Learning

Ch4. Deep Learning Algorithms

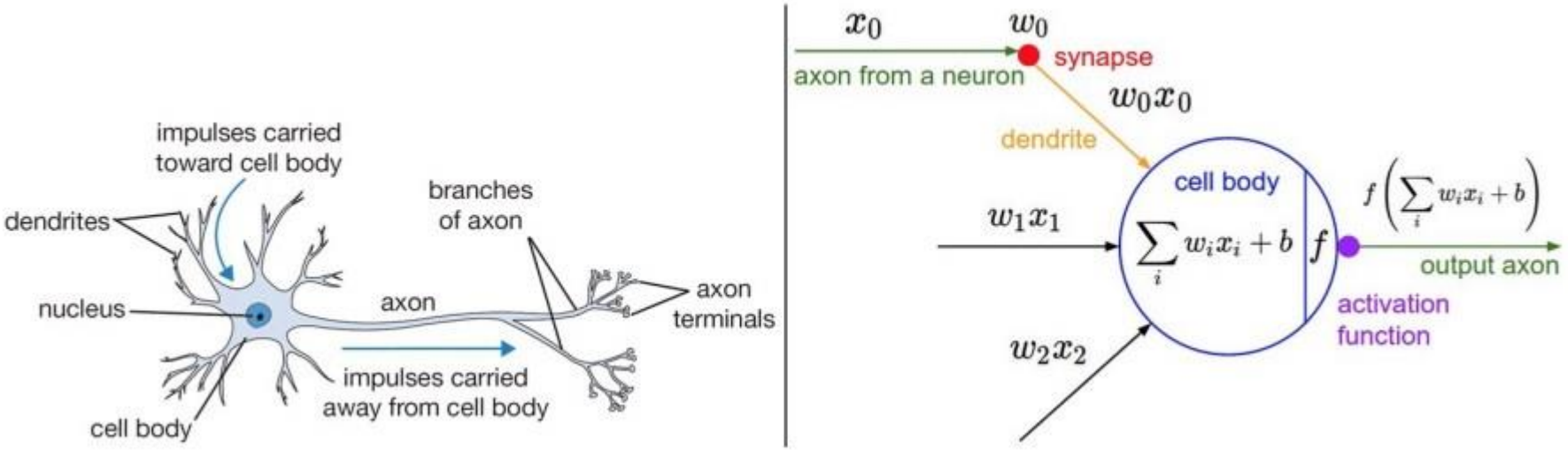
Ch5. Overfitting Issues

Ch6. Black Box Model

Artificial Neural Networks

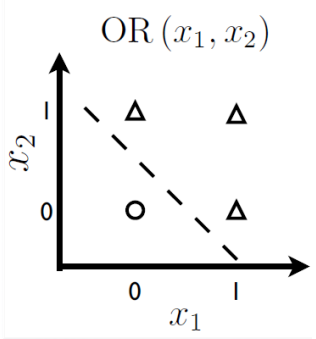
Neural Networks

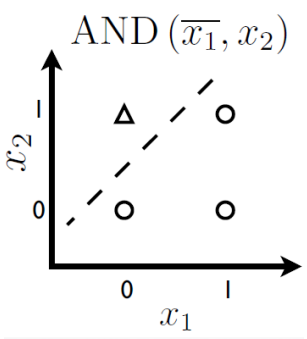
- Artificial neural network mimics the human brain
  - ➢ While neural network-based algorithms for classification or regression may be useful for the purpose of artificial intelligence (AI), neural network itself has nothing to do with.
  - ➢ Single neuron is activated (=1) or not (=0).

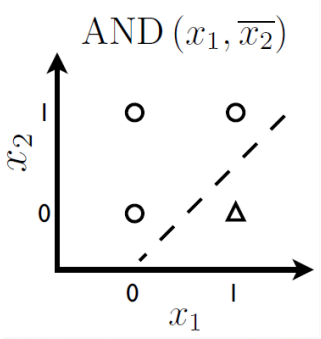


Capacity of Neural Networks

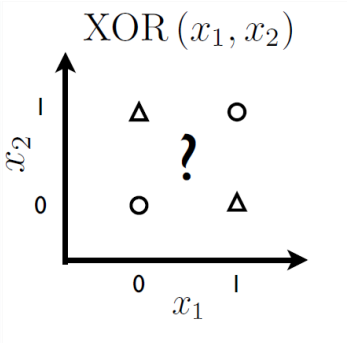
- Artificial neural networks can solve non-linearly separable problems.
  - ➢ That is, neural networks can represent more complex features!
- Linearly separable problems

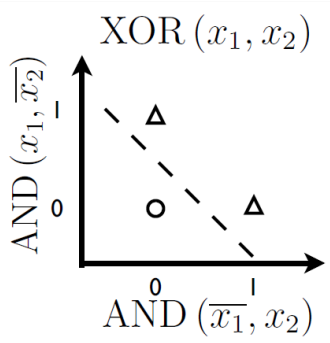






- Non-linearly separable problems



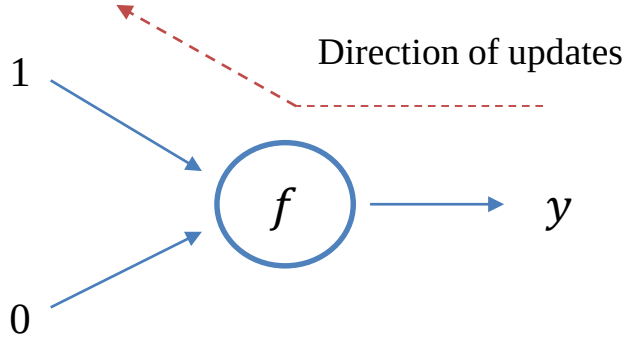


Multi-layer neural networks can solve it by transforming the input in a better representation.

Learning Algorithm of Neural Networks

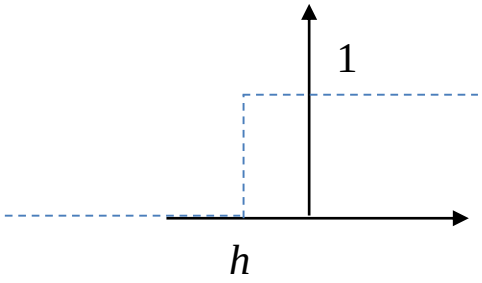
- Back-propagation

  = Learning parameters to minimize the error from the true value in the reverse way



Activation function

$y = f(w_1 * 1 + w_0 * 0)$



(1) If the model outputs 1 and the true value is 0, how should you modify the parameters (w_1, w_0, h)?

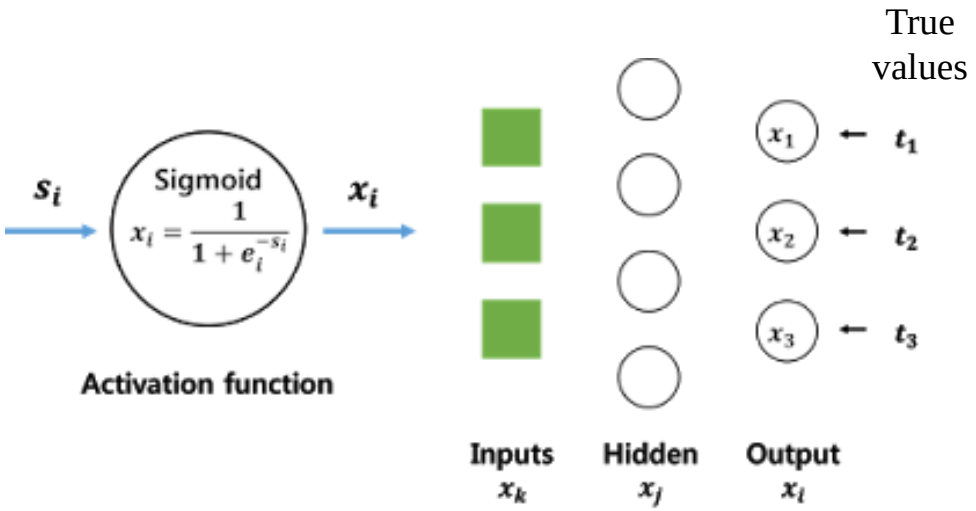
(2) If the model outputs 0 and the true value is 1, how should you modify the parameters (w_1, w_0, h)?

(3) Until when? Minimizing *the error (loss) function*

(Appendix) Learning Algorithm of Neural Networks

- Gradient descent method (Back-propagation)

  = Learning parameters to minimize the error from the true value in the systematic, reverse way even for multiple hidden layers



The goal is to minimize the following error (loss) function,

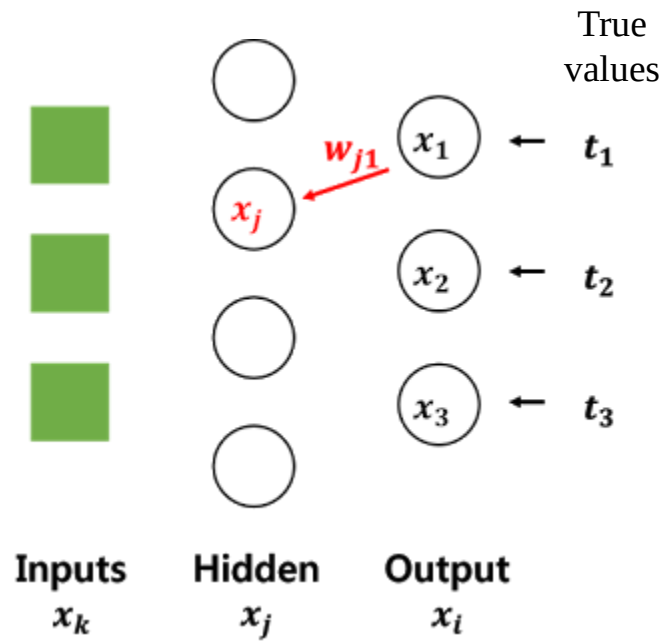
$$E = -\sum_{i=1}^{3}[t_i \log(x_i) + (1 - t_i)\log(1 - x_i))]$$

Source: Backpropagation 설명 예제와 함께 완전히 이해하기 (Jaejun Yoo's Playground)
http://jaejunyoo.blogspot.com/2017/01/backpropagation.html

(Appendix) Learning Algorithm of Neural Networks

- Gradient descent method (Back-propagation)

  = Learning parameters to minimize the error from the true value in the systematic, reverse way even for multiple hidden layers



Firstly, update the parameters between output and hidden layers.

$$\frac{\partial E}{\partial w_{j1}} = \frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial s_i}\frac{\partial s_i}{\partial w_{ji}} \quad \text{(Chain rule)}$$

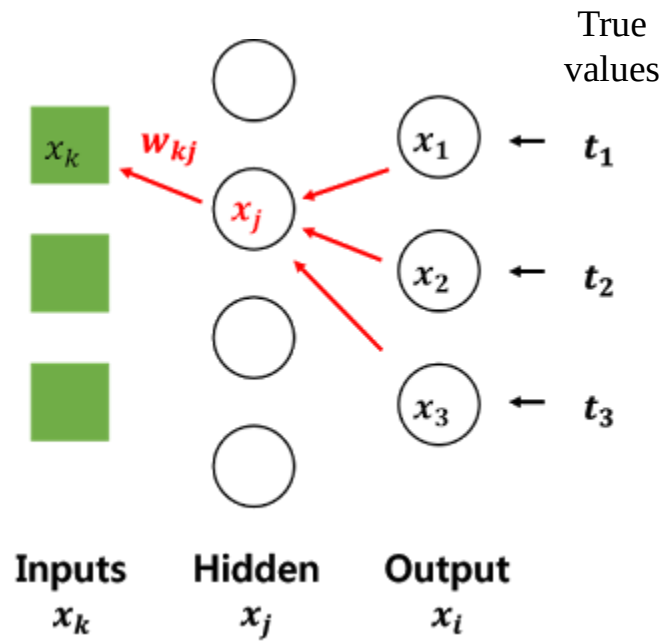
$$\therefore \frac{\partial E}{\partial w_{j1}} = (x_i - t_i)x_j$$

Source: Backpropagation 설명 예제와 함께 완전히 이해하기 (Jaejun Yoo's Playground)
http://jaejunyoo.blogspot.com/2017/01/backpropagation.html

(Appendix) Learning Algorithm of Neural Networks

- Gradient descent method (Back-propagation)

  = Learning parameters to minimize the error from the true value in the systematic, reverse way even for multiple hidden layers



Secondly, update the parameters between hidden and input layers.

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial s_j}\frac{\partial s_j}{\partial w_{ki}}$$

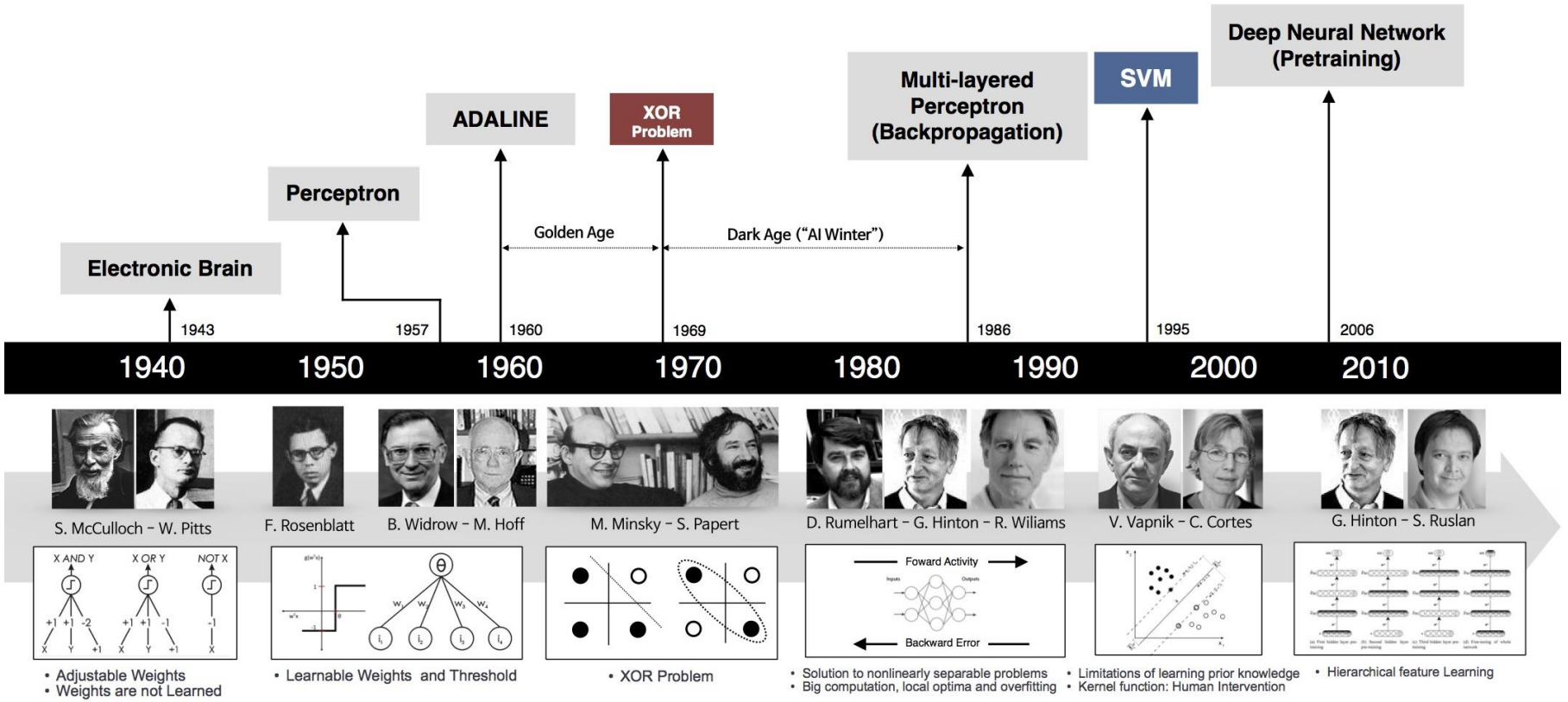
$$\frac{\partial E}{\partial w_{kj}} = \sum_{i=1}^{3} \frac{\partial E}{\partial s_i}\frac{\partial s_i}{\partial x_j}\frac{\partial x_j}{\partial s_j} \times \frac{\partial s_j}{\partial w_{ki}}$$

$$\therefore \frac{\partial E}{\partial w_{kj}} = \sum_{i=1}^{3} (x_i - t_i) w_{ji} (x_j(1 - x_j)) \times x_k$$

Source: Backpropagation 설명 예제와 함께 완전히 이해하기 (Jaejun Yoo's Playground)
http://jaejunyoo.blogspot.com/2017/01/backpropagation.html

History of Deep Learning

Milestones in the Development of Neural Networks



Source: Deep Learning 101 - Part 1: History and Background
https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

Why is Deep Learning Feasible Now?

- There were fundamental problems with deep neural networks.
  - (1) High computational costs (underfitting issues)
  - (2) Overfitting problems

- How does deep learning overcome the obstacles?
  - (1) To mitigate computational costs or improve optimization
    - Computing power (e.g., GPU)
    - Pre-training one layer at a time in a unsupervised way *(Hinton's contribution)*
  - (2) To mitigate overfitting problems (= allowing an white-noise or loose-fit)
    - Denoising algorithms
    - Data preprocessing (e.g., data augmentation)
    - Regularization (e.g., dropout, L2 / L1 regularization)
    - etc.

Breakthroughs of Deep Learning

- Pioneers of CNN, RNN, and LSTM in the 1990s

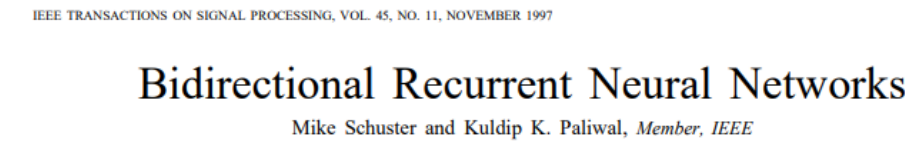
Convolutional Neural Networks (CNN)


PROC. OF THE IEEE, NOVEMBER 1998

Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

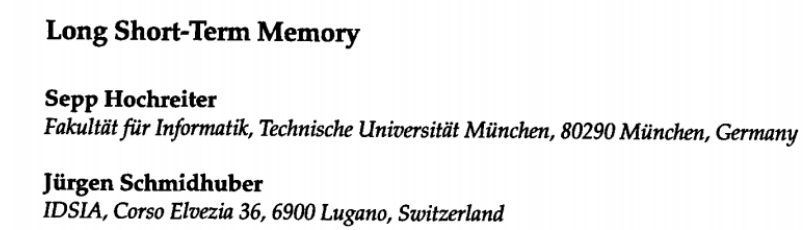
Recurrent Neural Networks (RNN)


IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 45, NO. 11, NOVEMBER 1997

Bidirectional Recurrent Neural Networks

Mike Schuster and Kuldip K. Paliwal, *Member, IEEE*

Long Short-Term Memoery (LSTM)


Long Short-Term Memory

Sepp Hochreiter
Fakultät für Informatik, Technische Universität München, 80290 München, Germany

Jürgen Schmidhuber
IDSIA, Corso Elvezia 36, 6900 Lugano, Switzerland

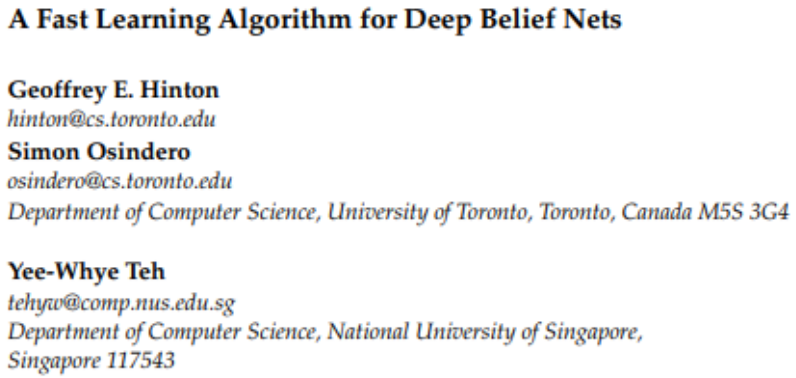
LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11), pp.2278-2324.

Schuster, M. and Paliwal, K.K., 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11), pp.2673-2681.

Hochreiter, S. and Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation*, 9(8), pp.1735-1780.

Breakthroughs of Deep Learning

- Opening the current deep learning era



A Fast Learning Algorithm for Deep Belief Nets

Geoffrey E. Hinton
hinton@cs.toronto.edu
Simon Osindero
osindero@cs.toronto.edu
Department of Computer Science, University of Toronto, Toronto, Canada M5S 3G4

Yee-Whye Teh
tehyw@comp.nus.edu.sg
Department of Computer Science, National University of Singapore,
Singapore 117543

REPORT

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton*, R. R. Salakhutdinov
+ See all authors and affiliations

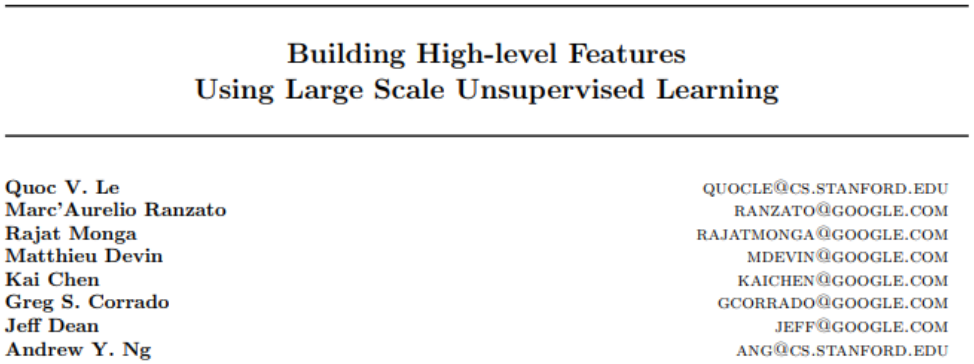
Science 28 Jul 2006:
Vol. 313, Issue 5786, pp. 504-507
DOI: 10.1126/science.1127647

Hinton, G.E., Osindero, S. and Teh, Y.W., 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), pp.1527-1554.

Hinton, G.E. and Salakhutdinov, R.R., 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), pp.504-507.

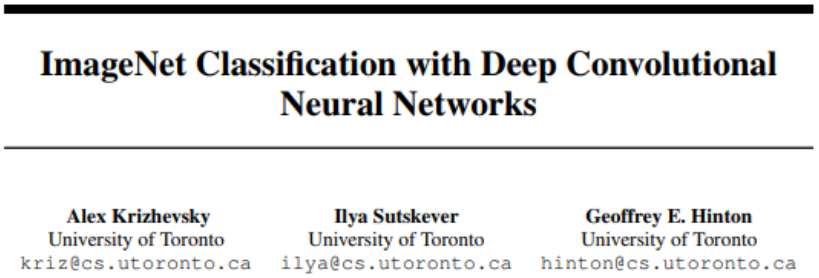
Breakthroughs of Deep Learning

- Showing the possibility of deep learning in practice on a large scale



Building High-level Features
Using Large Scale Unsupervised Learning

Quoc V. Le QUOCLE@CS.STANFORD.EDU
Marc'Aurelio Ranzato RANZATO@GOOGLE.COM
Rajat Monga RAJATMONGA@GOOGLE.COM
Matthieu Devin MDEVIN@GOOGLE.COM
Kai Chen KAICHEN@GOOGLE.COM
Greg S. Corrado GCORRADO@GOOGLE.COM
Jeff Dean JEFF@GOOGLE.COM
Andrew Y. Ng ANG@CS.STANFORD.EDU



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Le, Q.V., Ranzato, M.A., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J. and Ng, A.Y., 2012. Building High-Level Features using Large Scale Unsupervised Learning. *In Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML).*

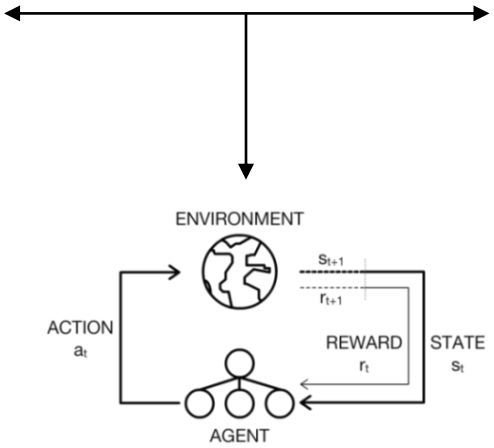
Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet Classification with Deep Convolutional Neural Networks. *In Advances in Neural Information Processing Systems (NIPS).*

Unsupervised Deep Learning

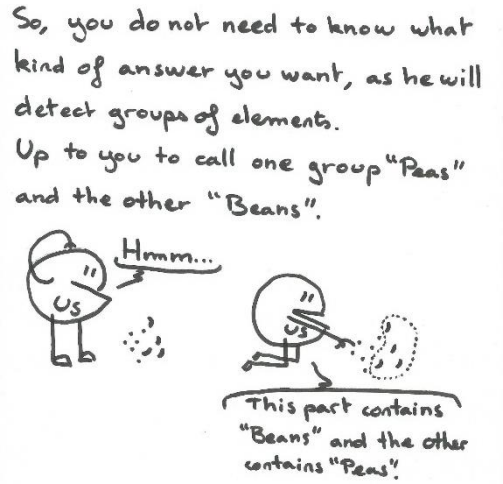
Supervised and Unsupervised Learning

- How to teach the machine

Supervised learning



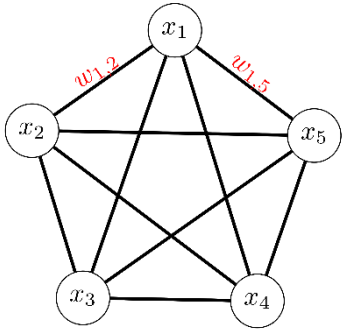
Reinforcement learning

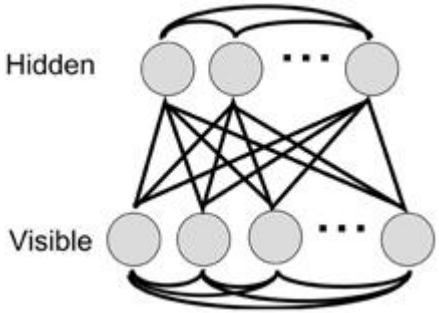


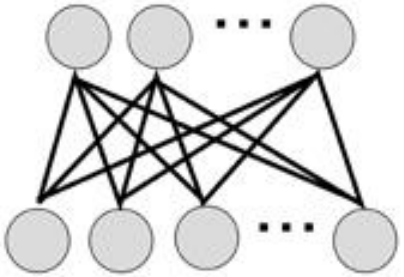
Unsupervised learning

Source: https://triskell.github.io/2016/11/13/Supervised-Learning-and-Unsupervised-Learning.html

Restricted Boltzmann Machine (RBM)







p(h | x)

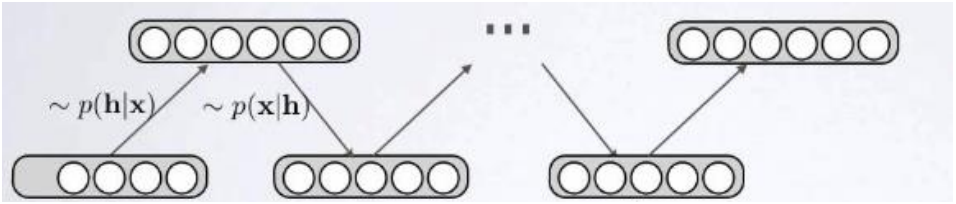
p(x | h)

Hopfield network

Boltzmann machine
(= stochastic Hopfield network with hidden units)

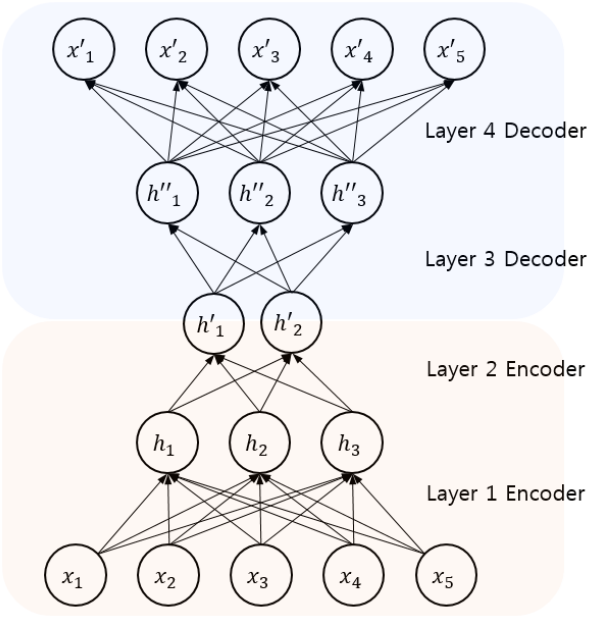
Restricted Boltzmann Machine
(= Boltzmann machines, with the restriction that there are no connections between nodes within a group)

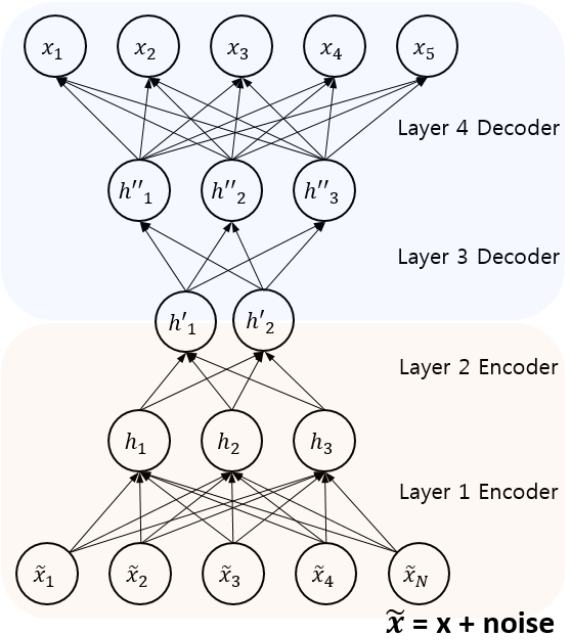
- RBM is unsupervised learning using Gibbs sampling



Autoencoder

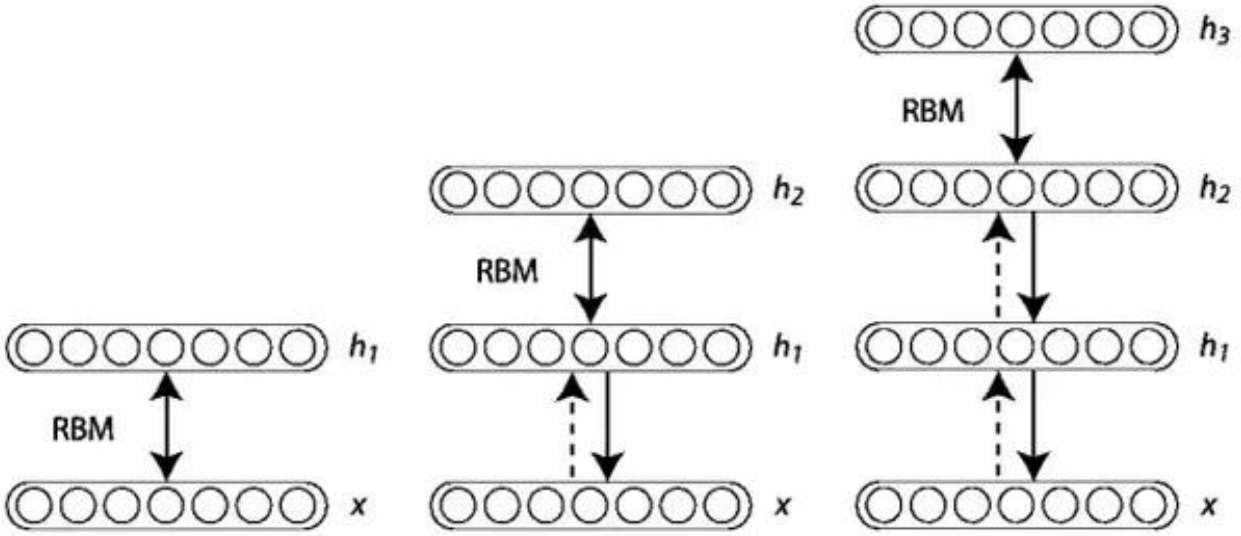
- Autoencoder is a variant of neural networks whose output is the input.
  - ➢ It is to learn a representation (encoding) for a set of data, typically for the purpose of dimensionality reduction.
  - ➢ Denoising autoencoder is to intentionally introduce noises into the input, allowing the model to encode features robust to noises.





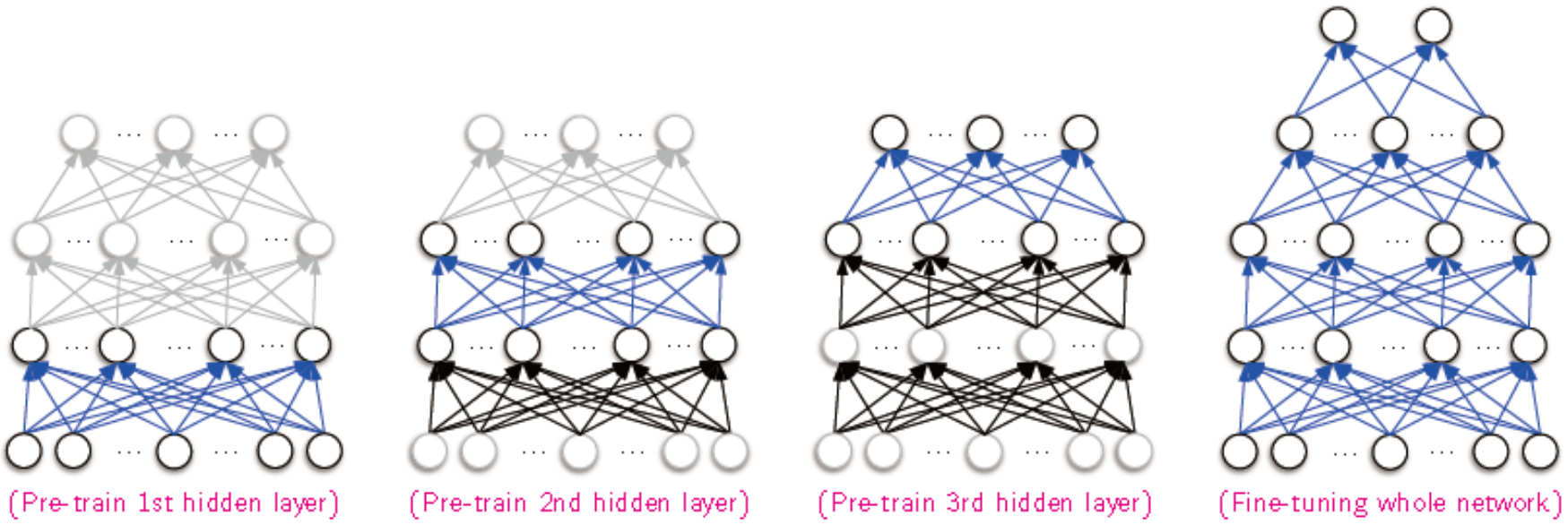
Deep Belief Networks

- Deep belief networks (DBNs) are generative models that are trained using a series of stacked Restricted Boltzmann Machines (or sometimes Autoencoders) with an additional layers.



Pre-Training using Unsupervised Learning

- Greedy layer-wise training of deep networks (Hinton 2006; Bengio et al. 2007)
  - Feed-forward, pre-training one layer at a time in a unsupervised way
  - Fine-tuning whole networks using supervised back propagation

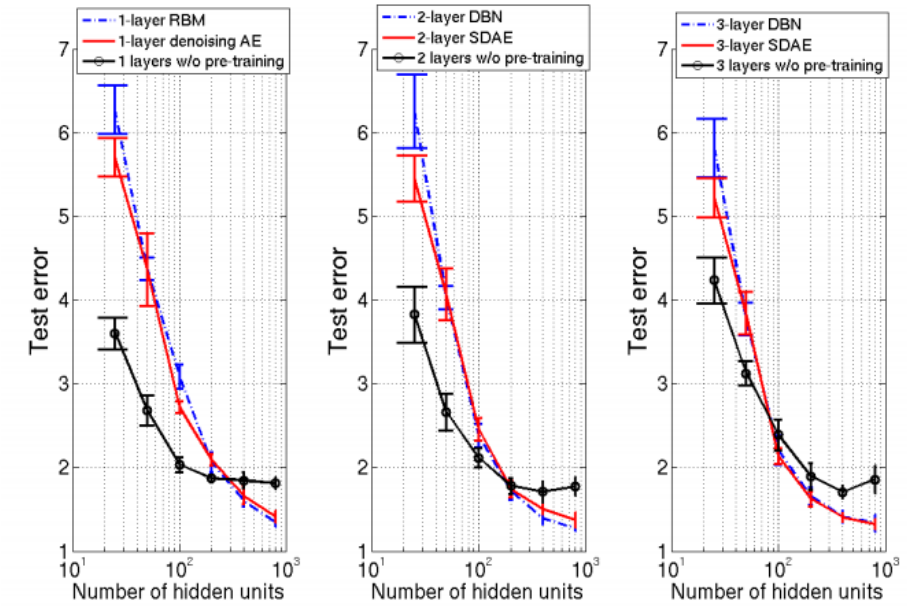


Hinton, G.E., Osindero, S. and Teh, Y.W., 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), pp.1527-1554.

Bengio, Y., Lamblin, P., Popovici, D. and Larochelle, H., 2007. Greedy Layer-wise Training of Deep Networks. *In Advances in Neural Information Processing Systems (NIPS).*

Pre-Training using Unsupervised Learning

- Unsupervised pre-training helps supervised deep learning
  - ➢ Many studies suggest that denoising autoencoders performs generally well at pre-training. (Vincent et al. 2008)



DBN: Deep Belief Net
SDAE: Stacked Denoising Auto Eocoder

(Erhan et al. 2010)

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P. and Bengio, S., 2010. Why Does Unsupervised Pre-training Help Deep Learning?. *Journal of Machine Learning Research*, 11, pp.625-660.

Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.A., 2008. Extracting and Composing Robust Features with Denoising Autoencoders. In Proceedings of the 25th International Conference on Machine Learning (ICML)

Deep Learning Algorithms

Representation Learning + Deep Architecture

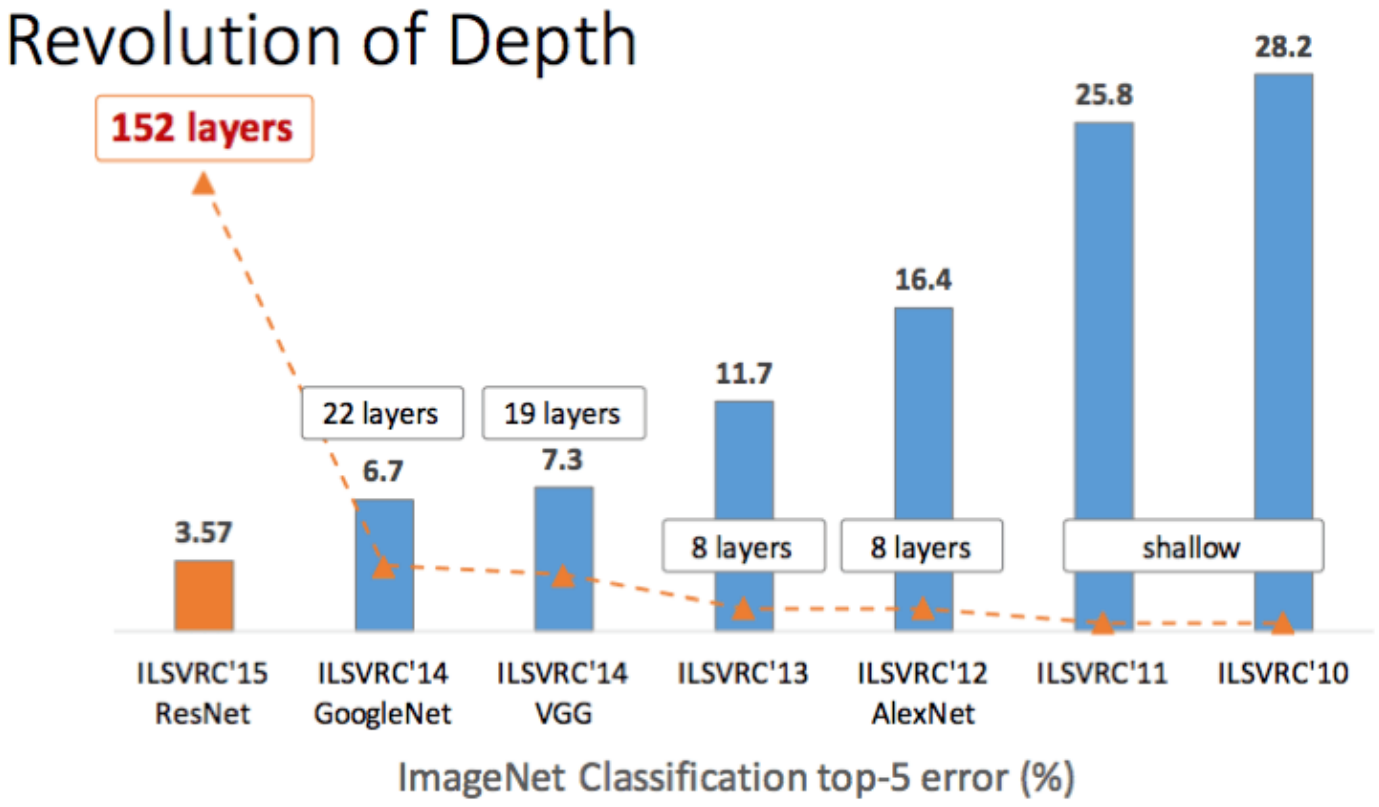
- "Deep Learning can be considered as special case of representation learning algorithms which learn representations of the data in a Deep Architecture with multiple levels of representations." (Najafabadi et al. 2015, p. 5)

- Recent deep learning algorithms are mostly based on neural networks.
  - Neural network + Deep architecture
    = Feed-forward networks with many hidden layers
    = Multi-layer perceptron (MLP)
    = Deep neural network (DNN)
  - More specialized algorithms: CNN, RNN, LSTM, etc.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep Learning Applications and Challenges in Big Data Analytics. *Journal of Big Data*, *2*(1), 1.

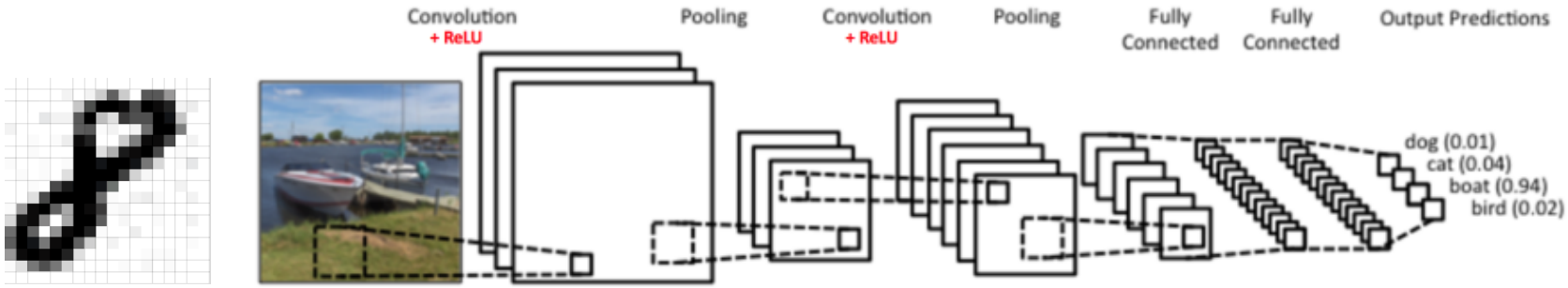
Representation Learning + Deep Architecture

- Then, how deep?



Convolutional Neural Network (CNN)

- A typical example of CNN



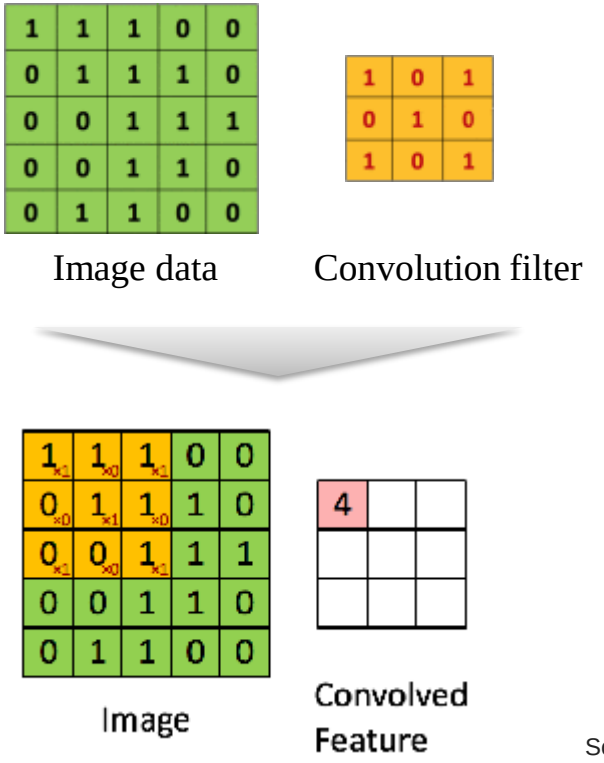
  - Input layer → [Convolution -> Pooling] → … → [Convolution -> Pooling] → Fully connected layer → Output layer (prediction)

- CNN has been considered as a basic deep learning algorithm. Why is CNN superior at feature representations?
  - Effectively reducing computational complexity (Convolution filter)
  - Hierarchical feature representation (multiple convolution layers)
  - Non-linearity and less overfitting (Rectified Linear Unit (ReLU) and Pooling)

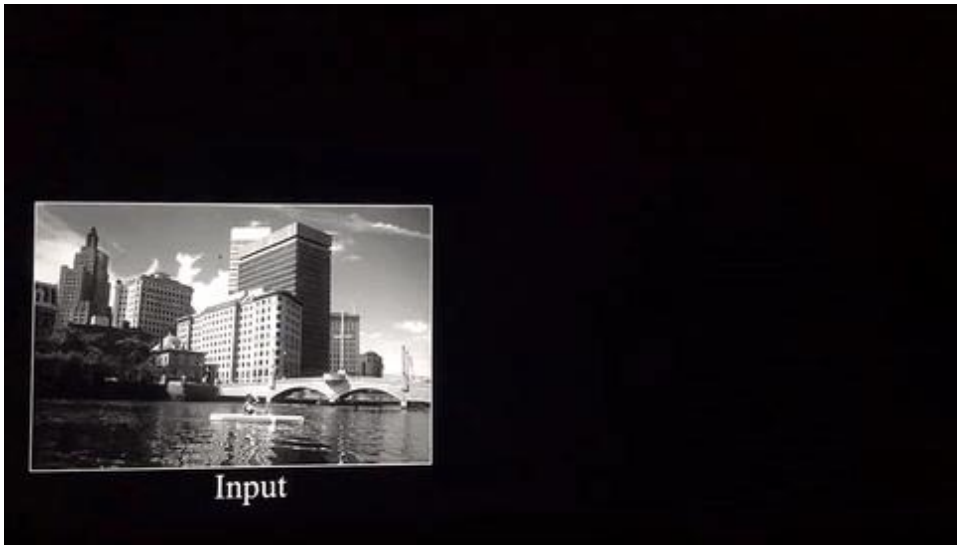
Convolutional Neural Network (CNN)

- Convolution is to filtering higher level of abstractions or features.
  - For images, this is already used in filter cameras.

Example (1)



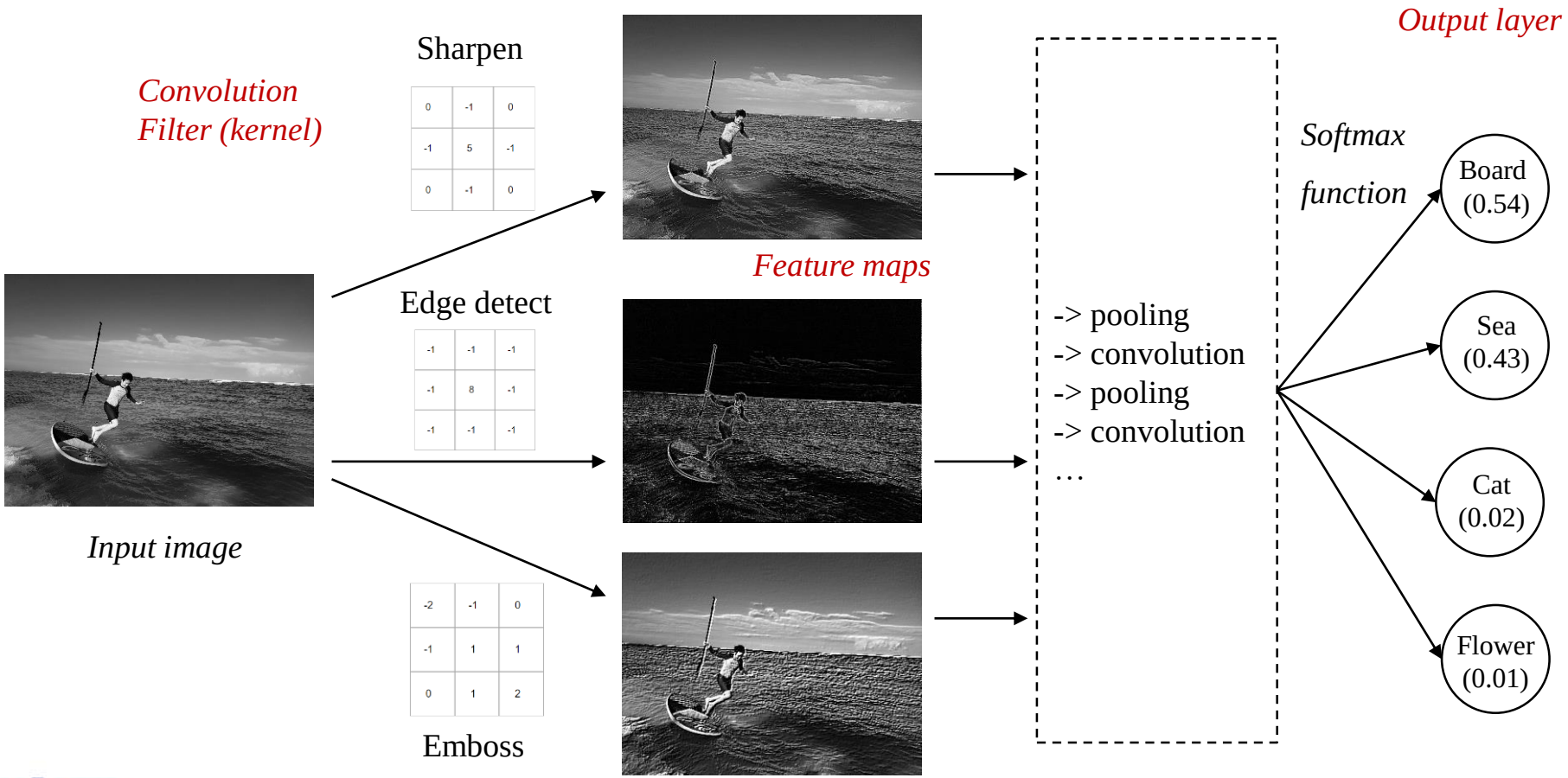
Example (2)



Source: An Intuitive Explanation of Convolutional Neural Networks (Data Science Blog)
https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Convolutional Neural Network (CNN)

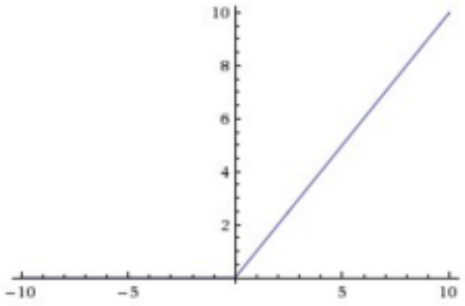
- CNN automatically learns the best filters to extract the feature maps for image classifications through back propagation.



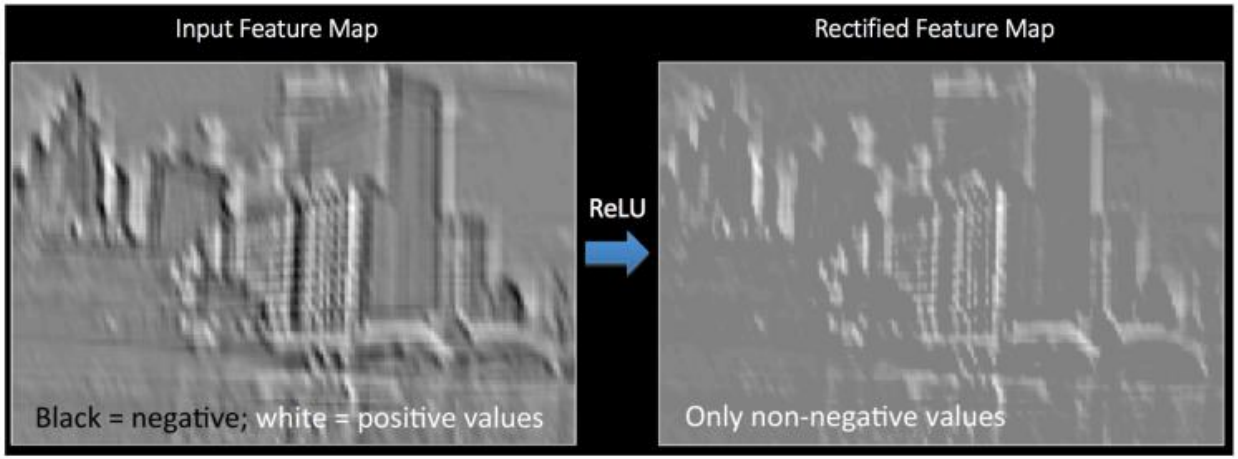
Source: Image Kernels demo, http://setosa.io/ev/image-kernels/

Convolutional Neural Network (CNN)

- Introducing non-linearity (ReLU)
  - The purpose of ReLU is to introduce non-linearity in CNN.
  - Other non-linear functions such as tanh or sigmoid can also be used, but ReLU has been found to perform better in most situations.



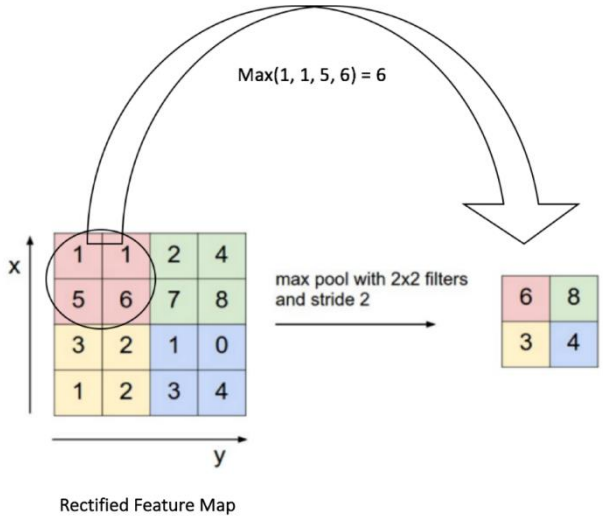
Output = Max(zero, Input)

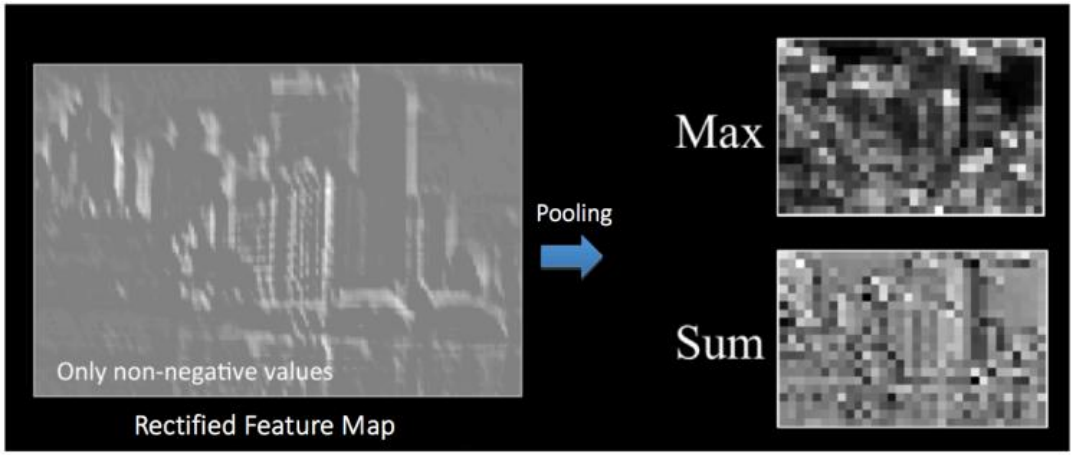


Source: An Intuitive Explanation of Convolutional Neural Networks (Data Science Blog)
https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Convolutional Neural Network (CNN)

- Reducing the dimensionality of each feature map (Pooling)
  - The purpose of pooling is to reduce the number of parameters and computations in the network, therefore, controlling overfitting.
  - Another purpose is to make the network invariant to small transformations, distortions and translations in the input image.
  - Pooling can be of different types: Max, Average, Sum, etc.

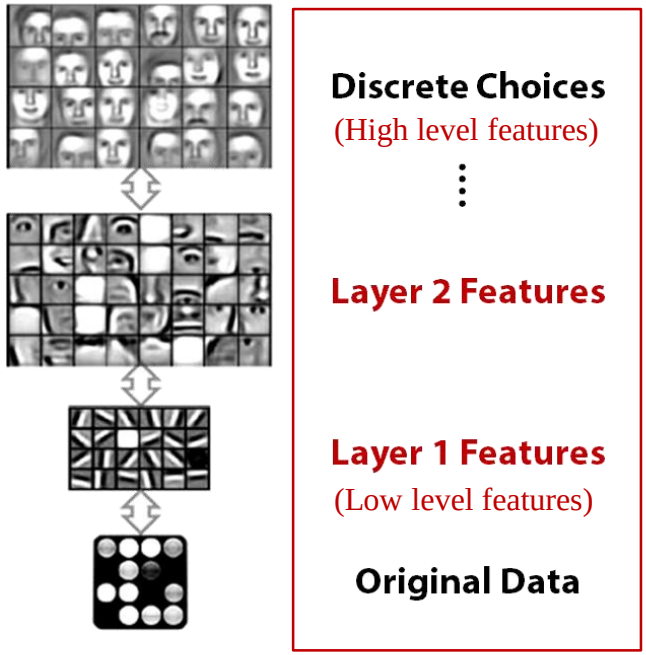


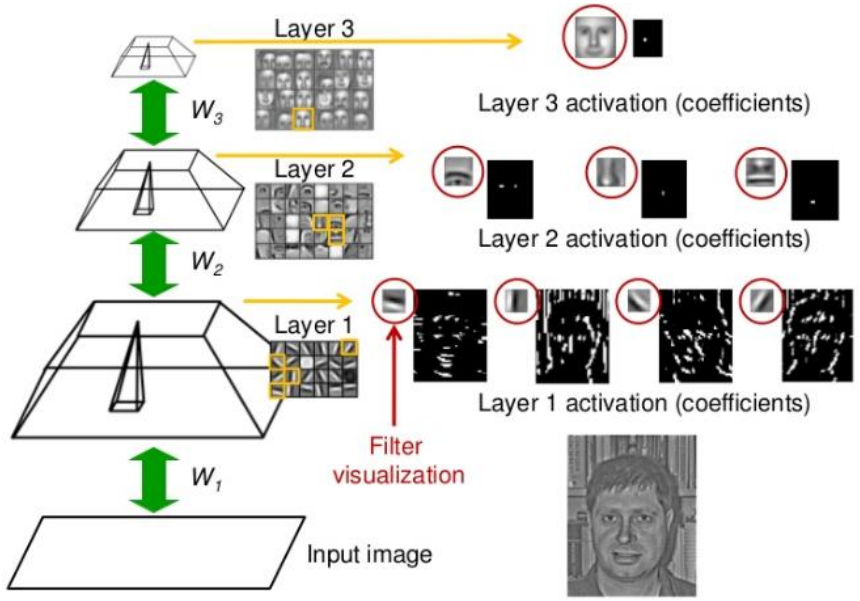


Source: An Intuitive Explanation of Convolutional Neural Networks (Data Science Blog)
https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Convolutional Neural Network (CNN)

- In short, deep learning is to learn hierarchical representations of data through multiple stages of non-linear feature transformations.
  - ➢ The hierarchical learning from low to high level features is processed through convolution filters.



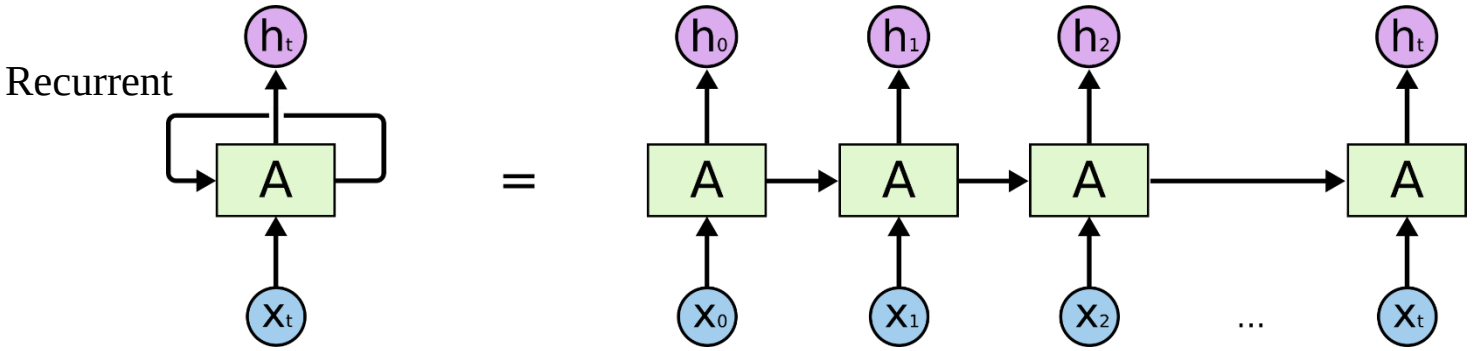


Source: https://cs.stackexchange.com/questions/16545/what-is-the-difference-between-a-neural-network-a-deep-learning-system-and-a-de

Recurrent Neural Network (RNN)

- Unlike CNN, RNN can use past information to learn the present task.
  - Example: Natural Language Processing (NLP)
    - "The clouds are in the ( )."
    - "I grew up in France …… I speak fluent ( )."
  - Vanishing gradient problem
    - As that gap grows, RNN becomes unable to learn to connect the information.

      (the past information would be vanishing or exploding)



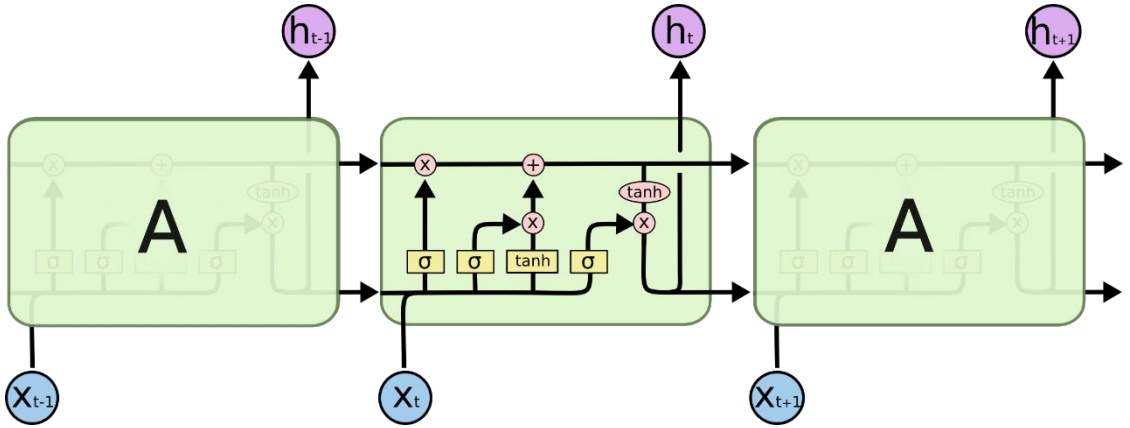
Source: Understanding LSTM Networks (Colah's Blog), http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Long Short-Term Memory (LSTM)

- LSTM allows RNN to learn how much past information would pass to the next
  - It can overcome the vanishing gradient problem.

- Introducing three gates into RNN
  - Forget gate (f_i): It decides what information we're going to throw away
  - Input gate (i_i): It decides which values we'll update
  - Output gate: It decides which parts we only output

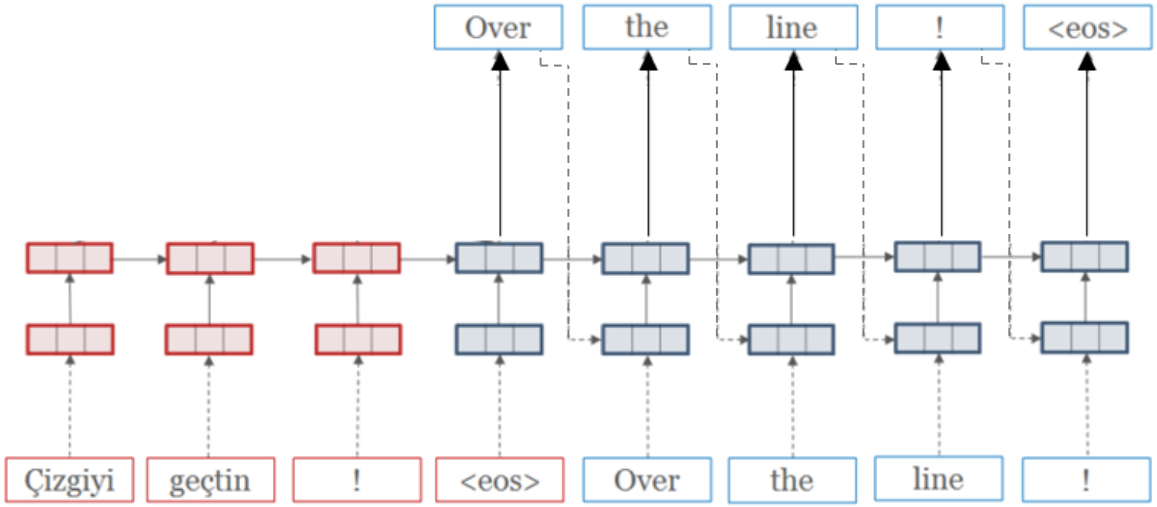
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



Source: Understanding LSTM Networks (Colah's Blog), http://colah.github.io/posts/2015-08-Understanding-LSTMs/

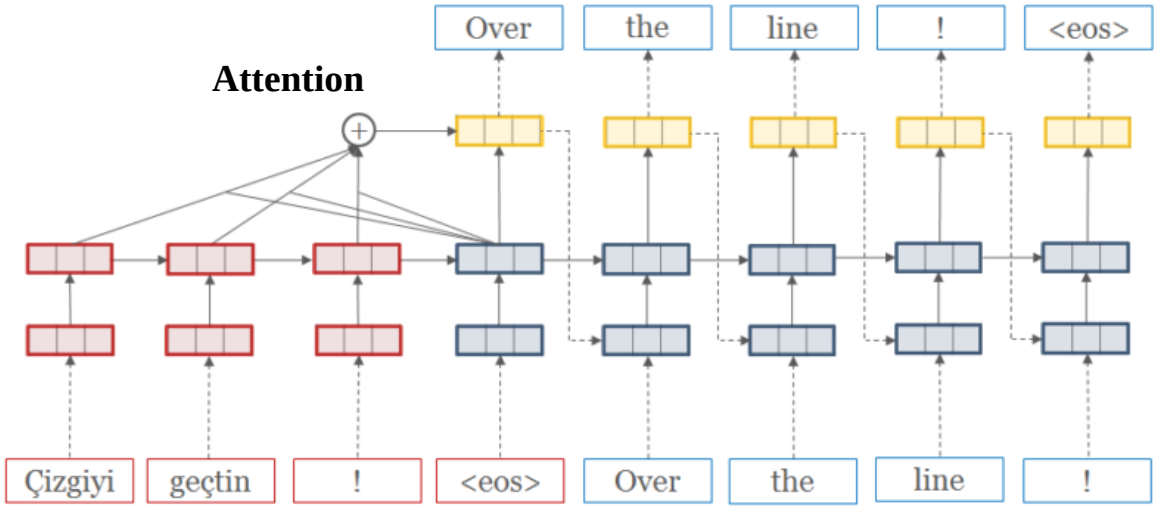
Sequence-to-Sequence (Seq2Seq)

- Encoder-Decoder architecture
  - **Encoder**: A source encoder RNN maps each source word to a word vector, and processes these to a sequence of hidden vectors.
  - **Decoder**: The target decoder combines an RNN hidden representation of previously generated words with source hidden vectors to predict scores for each possible next word.



Sequence-to-Sequence with Attention (attn-Seq2Seq)

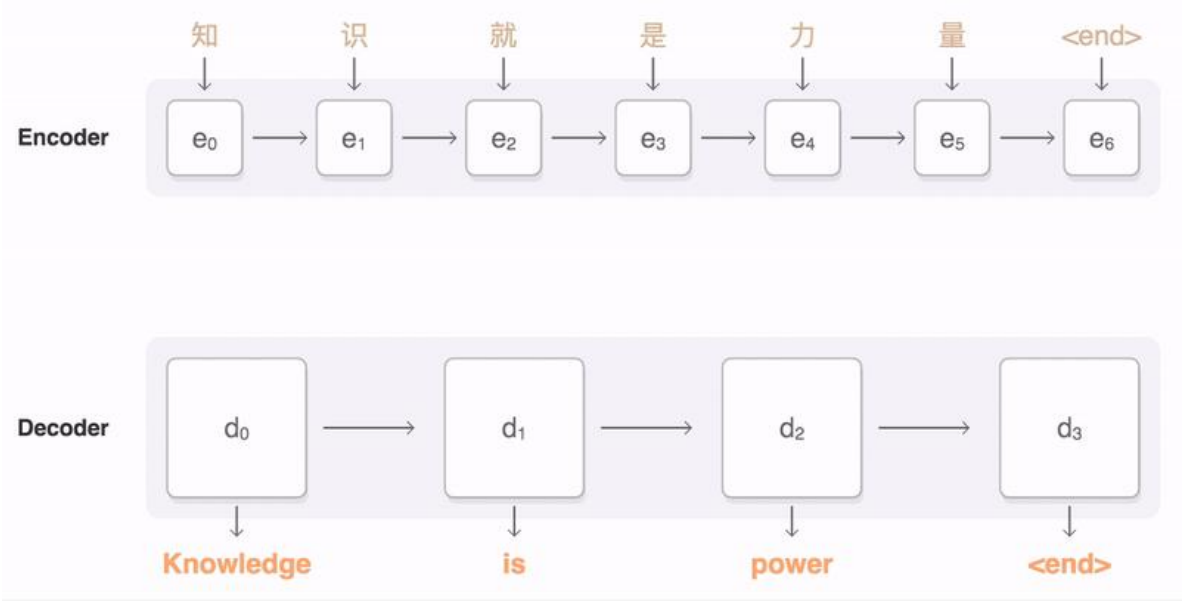
- Attention-based encoder-decoder architecture allows different weights each source word relative to its expected contribution to the target prediction.
  - ➢ As long as the problem can be phrased as encoding input data in one format and decoding it into another format, seq2seq models can be applied (e.g., machine translation, text summarization, style transfer).



Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

Sequence-to-Sequence with Attention (attn-Seq2Seq)

- Attention-based encoder-decoder architecture allows different weights each source word relative to its expected contribution to the target prediction.
  - ➢ As long as the problem can be phrased as encoding input data in one format and decoding it into another format, seq2seq models can be applied (e.g., machine translation, text summarization, style transfer).

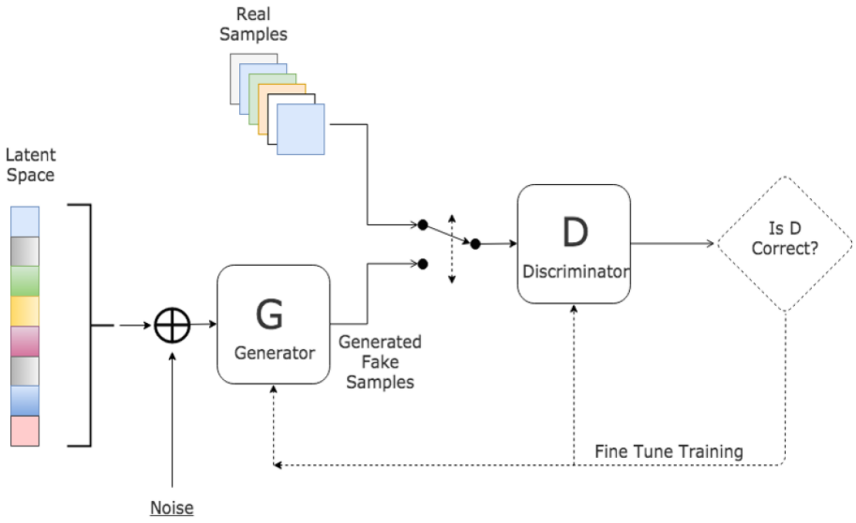


Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

Generative Adversarial Networks (GAN)

- GAN models aim to yield the generated samples to be indistinguishable from real data, using two competing neural networks.
  - **Generator**: Generating samples by taking noise as input
  - **Discriminator**: Distinguishing between the two sources from both the generator and the training data

- These two networks play a continuous game, where the generator is learning to produce more and more realistic samples, and the discriminator is learning to get better at distinguishing generated data from real data.



Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative Adversarial Nets. *In Advances in Neural Information Processing Systems (NIPS)*.

Transfer Learning

- Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned.

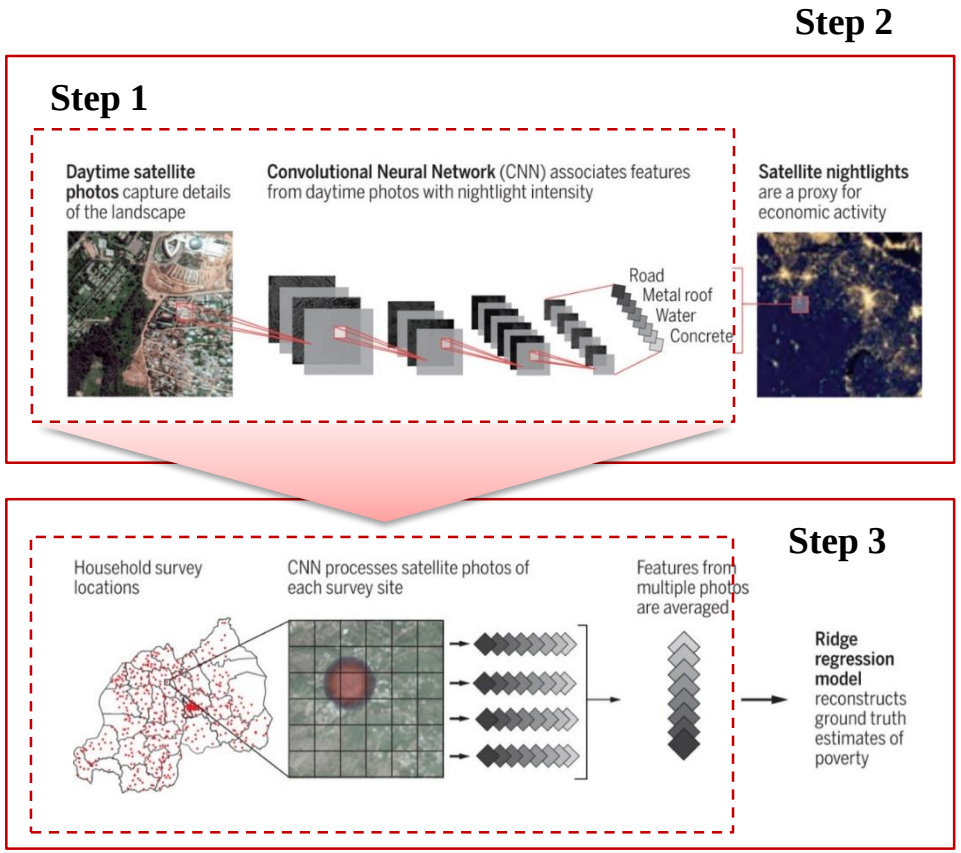
- Example: Jean et al. (2016)

Step 1. Pre-training a CNN model on ImageNet

Step 2. Fine-tuning the CNN to predict satellite nightlights (on the globe where nightlights are abundant)

Transfer the learned model to a new task

Step 3. Using this learned model as a feature extractor for daytime satellite images to predict the poverty level (on Africa where nightlights are scarce)

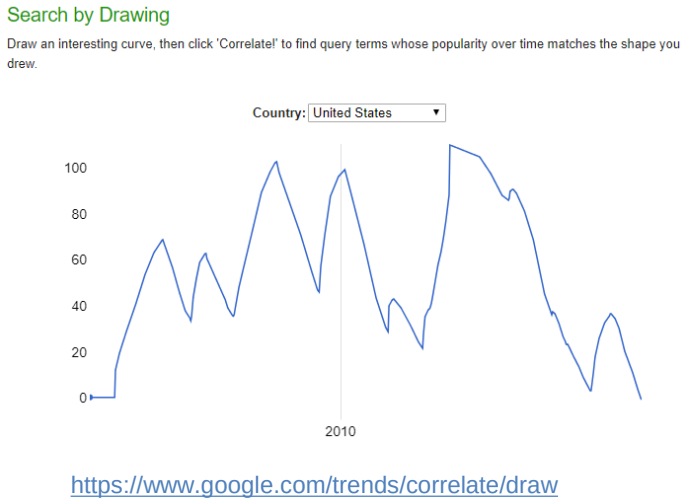


Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B. and Ermon, S., 2016. Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science*, 353(6301), pp.790-794.

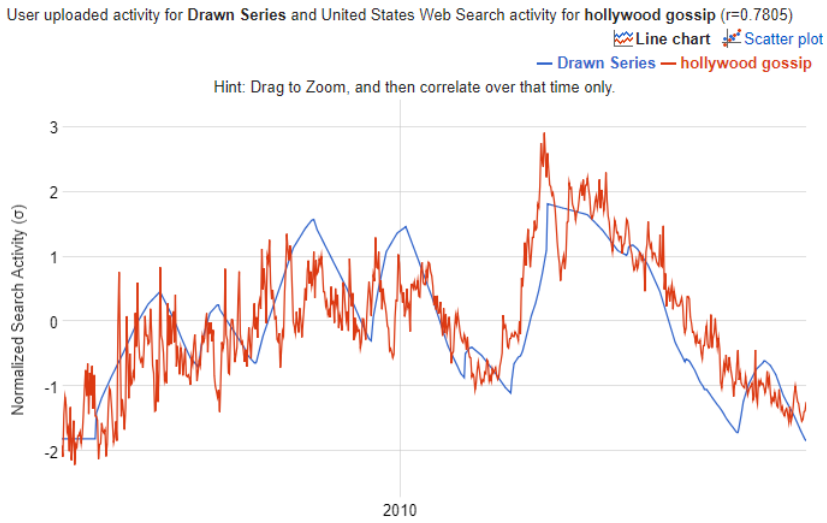
Overfitting Issues

Overfitting is Easy

- When I draw a random curve of Google search trends...



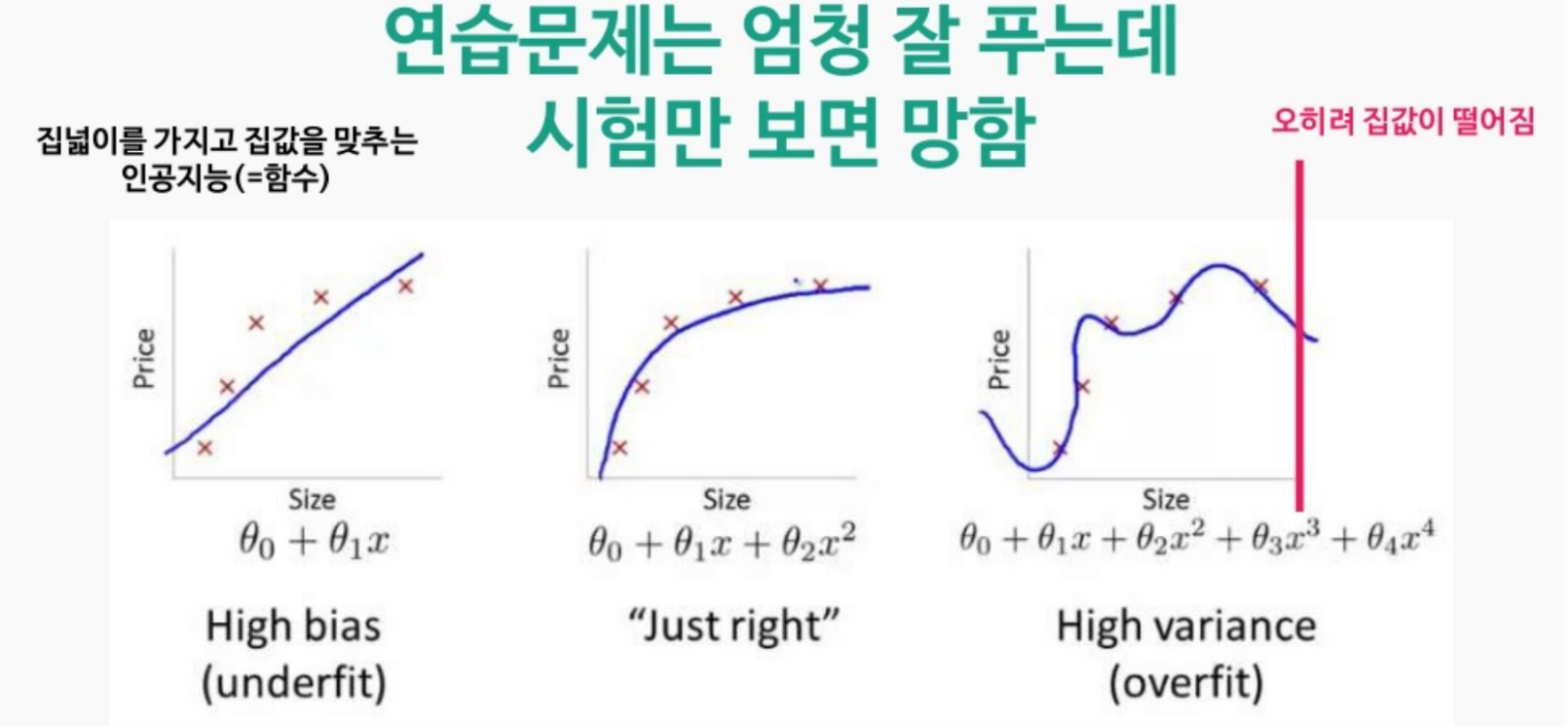
https://www.google.com/trends/correlate/draw



- "Eureka! I accidently succeed in predicting how people search about 'Hollywood gossip' on Google. Its correlation is 0.78!"
- … Really?

Overfitting is Common in Prediction

- Overfitting often results from (i) too complex models and (ii) too few data.



Source: https://www.slideshare.net/modulabs/2-cnn-rnn

Occam's Razor

- Occam's razor is still valid (that is, simple is best).
  - For out-of-sample predictions, simpler models are more likely to hold up on future observations than more complex ones, all else being equal (Dhar 2013).
- Sometimes, machine learning algorithms might have poorer performances than a simple logistic regression.
  - It may be especially when (1) there is relatively simple relationships, or (2) there are too few data to learn about the relationships.
  - Example: Predicting movie success (Lee et al. 2018)

Logistic Regression		Deep NN		Random Forest		Support Vector Machine	
LR		*NN (MLP)*		*RF*		*SVC*	
Bingo	1-Away	Bingo	1-Away	Bingo	1-Away	Bingo	1-Away
36.0 %	85.3 %	48.0 %	86.7 %	46.7 %	84.0 %	26.7 %	64.0 %
45.3 %	93.3 %	40.0 %	88.0 %	56.0 %	90.7 %	30.7 %	62.7 %
61.3 %	88.0 %	38.7 %	84.0 %	53.3 %	90.7 %	26.7 %	56.0 %
54.7 %	86.7 %	50.7 %	88.0 %	56.0 %	86.7 %	26.7 %	48.0 %
56.0 %	90.7 %	42.7 %	81.3 %	62.7 %	89.3 %	26.7 %	61.3 %
54.7 %	89.3 %	36.0 %	82.7 %	49.3 %	88.0 %	41.3 %	74.7 %
50.7 %	90.7 %	49.3 %	85.3 %	57.3 %	81.3 %	29.3 %	44.0 %
45.3 %	86.7 %	34.7 %	75.7 %	49.3 %	86.7 %	28.0 %	65.3 %
42.7 %	89.3 %	45.3 %	86.7 %	50.7 %	90.7 %	25.3 %	53.3 %
50.7 %	82.7 %	38.7 %	81.3 %	49.3 %	76.0 %	25.3 %	60.0 %
49.7 %	**88.3 %**	**42.4 %**	**84.0 %**	**53.1 %**	**86.4 %**	**28.7 %**	**58.9 %**
7.5 %	3.1 %	5.7 %	3.9 %	4.9 %	4.8 %	4.8 %	8.9 %

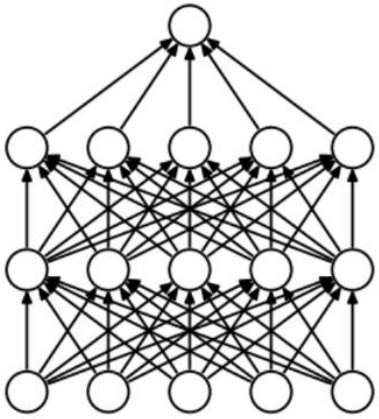
Accuracy

Dhar, V., 2013. Data Science and Prediction. *Communications of the ACM*, 56(12), pp.64-73.
Lee, K., Park, J., Kim, I. and Choi, Y., 2018. Predicting Movie Success with Machine Learning Techniques: Ways to Improve Accuracy. *Information Systems Frontiers*, 20(3), pp.577-588.

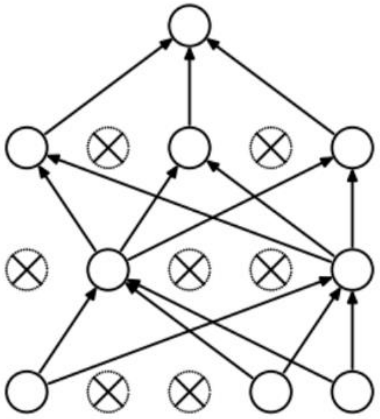
How Does Deep Learning Overcome Overfitting?

- Allowing white-noise or loose-fit (= 적당히 빈틈을 허용하자)

- Remind previous slides…
  - Debiasing autoencoder
  - Data augmentation
  - Pooling in CNN
  - Taking noise as input in GAN



(a) Standard Neural Net



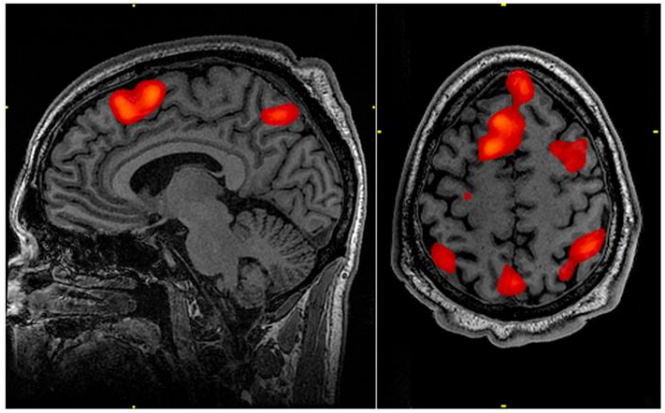
(b) After applying dropout.

- Regularization
  - Dropout
  - Penalizing model complexity (e.g., ridge regression)

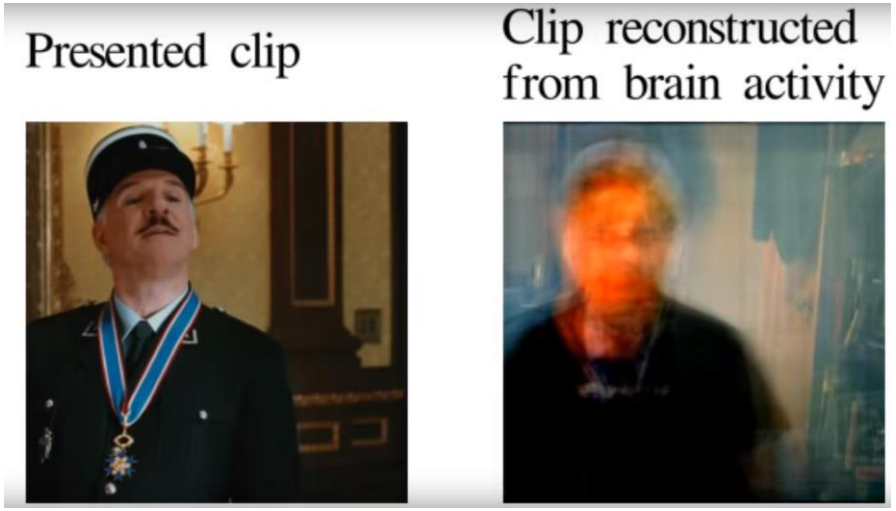
Black Box Model

Deep Learning is a Particularly Dark Black Box

- Pros and Cons of deep learning (as of 2018)
  - (Pros) Strangely, it works well :)
  - (Cons) Strangely, it does not work :(
- Although we can know the activities of neurons and even predict the meaning of their activities, we do not yet know the mechanism how our brain works.



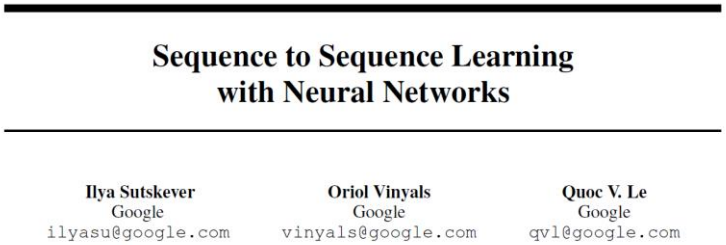
Functional MRI (fMRI)



Source: Movie reconstruction from human brain activity, https://youtu.be/nsjDnYxJ0bo

Deep Learning Does not Care about Why it Works

- Recent deep learning research has been advanced based on *speculations* and *experiments*.
  - (Example) Reversing the source sentences in seq2seq (Sutskever et al. 2014)



Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

"While we do not have a complete explanation to this phenomenon, we believe that it is caused by the introduction of many short term dependencies to the dataset…

Initially, we believed that reversing the input sentences would only lead to more confident predictions in the early parts of the target sentence and to less confident predictions in the later parts. However, LSTMs trained on reversed source sentences did much better on long sentences than LSTMs trained on the raw source sentences, which suggests that reversing the input sentences results in LSTMs with better memory utilization." (p. 4)

Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to Sequence Learning with Neural Networks. *In Advances in Neural Information Processing Systems (NIPS)*.

Deep Learning Does not Care about Why it Works

- Recent deep learning research has been advanced based on *speculations* and *experiments*.
  - ➢ (Example) Transformer model with self-attention (Vaswani et al. 2017)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

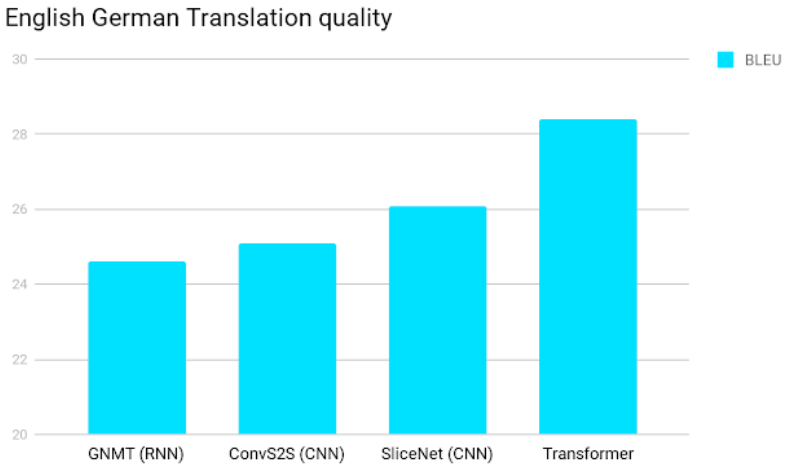
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is All You Need. *In Advances in Neural Information Processing Systems (NIPS).*

Deep Learning Does not Care about Why it Works

- Recent deep learning research has been advanced based on *speculations* and *experiments*.
  - ➢ (Example) Applying seq2seq into chemistry reactions (Schwaller et al. 2017)

"Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions using Neural Sequence-to-Sequence Models

Philippe Schwaller,* Théophile Gaudin,* Dávid Lányi, Costas Bekas, Teodoro Laino
IBM Research, Zurich
{phs,tga,dla,bek,teo}@zurich.ibm.com

"One way to view the reaction prediction task is to cast it as a translation problem… Intuitively, there is an analogy between a chemist's understanding of a compound and a language speaker's understanding of a word… The immediate consequence of this discovery is that the vocabulary of organic chemistry and human language follow very similar laws…

This has strengthened our belief that the methods of computational linguistics can have an immense impact on the analysis of organic molecules and reactions." (p. 2)

Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. and Laino, T., 2017. "Found in Translation": Predicting Outcome of Complex Organic Chemistry Reactions using Neural Sequence-to-Sequence Models. *In Advances in Neural Information Processing Systems (NIPS)*.

Deep Learning Does not Automatically Yield Best Results

- Researchers should specify a range of hyper-parameters and may need a pre-training in applying deep learning algorithms.
  - (Example) Document embedding using doc2vec (Schwaller et al. 2017)

An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation

Jey Han Lau[1,2] and Timothy Baldwin[2]
[1] IBM Research
[2] Dept of Computing and Information Systems,
The University of Melbourne

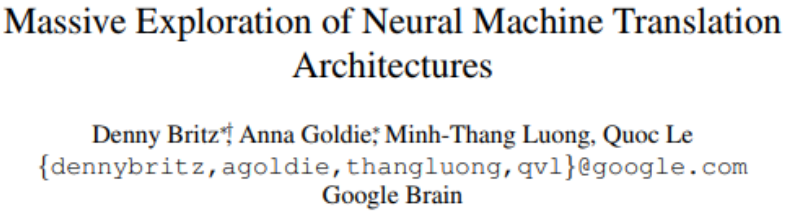
"While Le and Mikolov (2014) report state-of-the art results over a sentiment analysis task using *doc2vec*, others have struggled to replicate this result.

Given this background of uncertainty regarding the true effectiveness of *doc2vec*…, we aim to shed light on a number of empirical questions: (1) how effective is *doc2vec* in different task settings?; (2) which is better out of dmpv and dbow?; (3) is it possible to improve *doc2vec* through careful hyper-parameter optimization or with pre-trained word embeddings?; and (4) can *doc2vec* be used as an off-the-shelf model like *word2vec*?" (p. 1)

Lau, J.H. and Baldwin, T., 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *arXiv preprint arXiv:1607.05368.*

Deep Learning Does not Automatically Yield Best Results

- Searching the optimal hyper-parameters requires huge resources and long time for experiments. (too expensive..)
  - ➢ (Example) Exploration for neural machine translation (Britz et al. 2017)



Massive Exploration of Neural Machine Translation Architectures

Denny Britz*† Anna Goldie* Minh-Thang Luong, Quoc Le
{dennybritz,agoldie,thangluong,qvl}@google.com
Google Brain

"One major drawback of current architectures is that they are expensive to train, typically requiring days to weeks of GPU time to converge.

We report empirical results and variance numbers for several hundred experimental runs, corresponding to over 250,000 GPU hours on the standard WMT English to German translation task." (p. 1)

Britz, D., Goldie, A., Luong, T. and Le, Q., 2017. Massive Exploration of Neural Machine Translation Architectures. arXiv preprint arXiv:1703.03906.

Deep Learning Could be Biased, Even if not Intended

- Machines may learn not only human language, but also absorb ingrained prejudices concealed within the patterns of language use in the training data.

COGNITIVE SCIENCE

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,[1]* Joanna J. Bryson,[1,2]* Arvind Narayanan[1]*

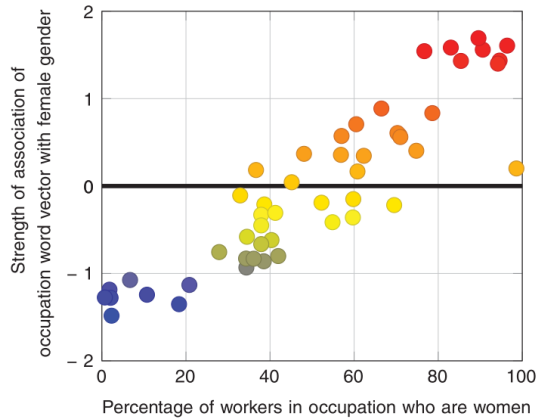


Fig. 1. Occupation-gender association. Pearson's correlation coefficient ρ = 0.90 with $P < 10^{-18}$.

White-sounding name

```
text_to_sentiment("My name is Emily")
```

```
2.2286179364745311
```

```
text_to_sentiment("My name is Heather")
```

```
1.3976291151079159
```

```
text_to_sentiment("My name is Yvette")
```

```
0.98463802132985556
```

```
text_to_sentiment("My name is Shaniqua")
```

```
-0.47048131775890656
```

Black-sounding name

Caliskan, A., Bryson, J.J. and Narayanan, A., 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, 356(6334), pp.183-186.

"How to make a racist AI without really trying," https://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

Fairness, Accountability, and Transparency in Machine Learning

- Fairness, accountability, and transparency (FAT) in machine learning is a very active research area.
  - (Example) the FATML workshop (https://www.fatml.org/)

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

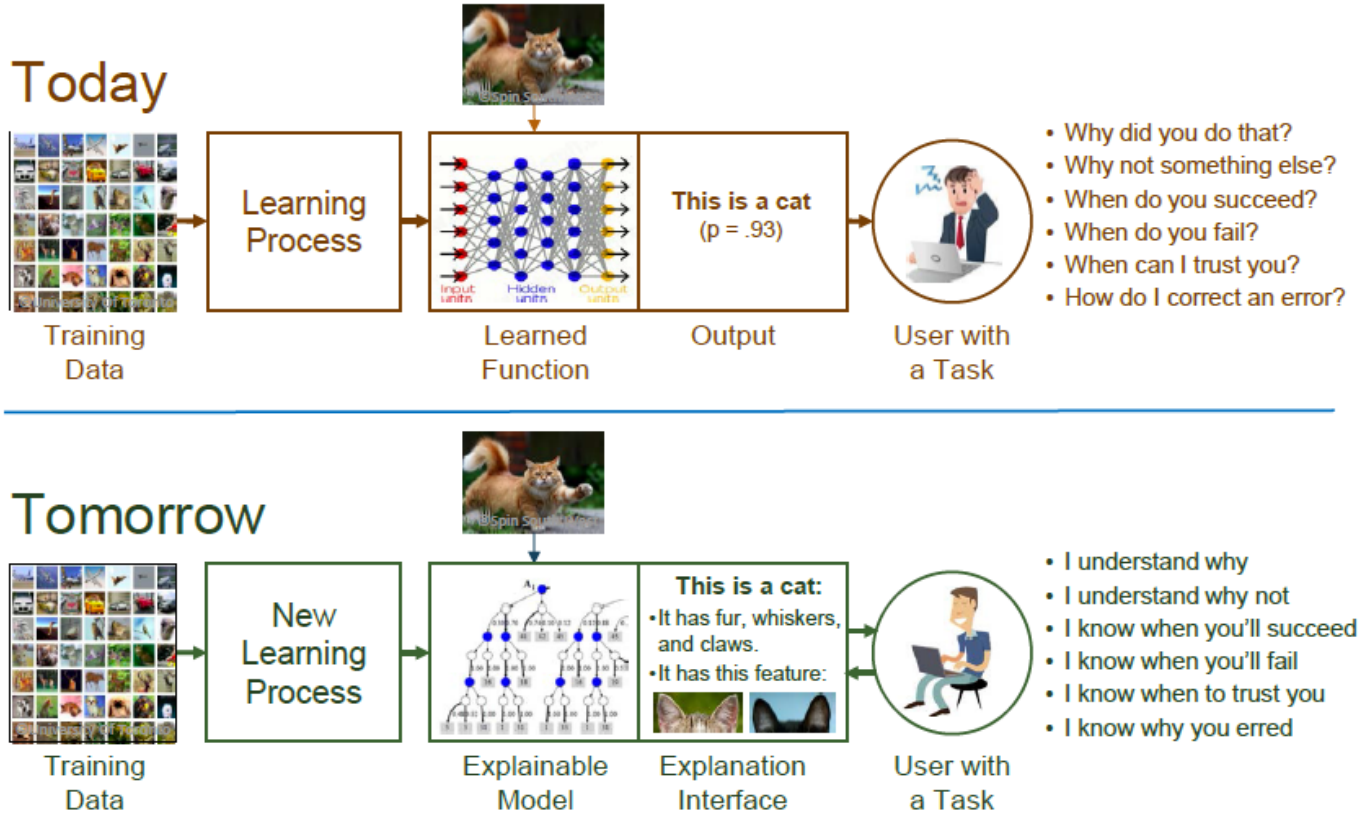
The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to "the algorithm made me do it."

The annual event provides researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods.

Explainable Artificial Intelligence (XAI)

- We don't want to just rely on a black box model.



Source: http://explainablesystems.comp.nus.edu.sg/wp-content/uploads/2018/03/XAI%20for%20IUI%202018.pdf

End of Document