

Introduction to Deep Learning

Principles and applications
in vision and natural language processing

Jakob Verbeek (INRIA)

Slides in collaboration with Laurent Besacier (Univ. Grenoble Alpes)

2018

Introduction

Convolutional Neural Networks

Recurrent Neural Networks

Wrap up

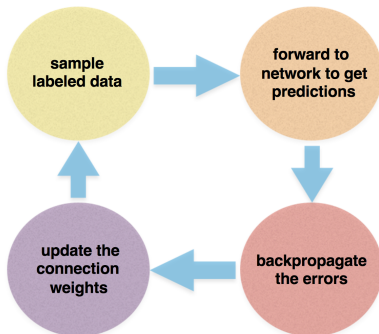
Machine Learning Basics

- ▶ **Supervised Learning:** use of labeled training set
 - ▶ ex: email spam detector with training set of already labeled emails
- ▶ **Unsupervised Learning:** discover patterns in unlabeled data
 - ▶ ex: cluster similar documents based on text content
- ▶ **Reinforcement Learning:** learning sequence of actions based on feedback or reward
 - ▶ ex: machine learn to play a game by winning or losing

What is Deep Learning

- ▶ Part of the ML field of learning representations of data
- ▶ Learning algorithms derive meaning out of data by using a **hierarchy of multiple layers** of units (*neurons*)
- ▶ Each unit computes a weighted sum of its inputs and the weighted sum is passed through a non linear function
 - ▶ each layer transforms input data in more and more abstract representations
- ▶ Learning = find optimal parameters (weights) from data
 - ▶ ex: deep automatic speech transcription or neural machine translation systems have 10-20M of parameters

Supervised Learning Process



- ▶ Learning by generating error signal that measures the differences between network predictions and true values
- ▶ Error signal used to update the network parameters so that predictions get more accurate

Brief History

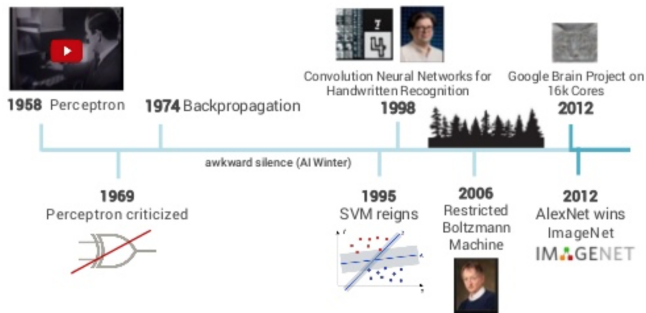


Figure from <https://www.slideshare.net/LuMa921/deep-learning-a-visual-introduction>

- ▶ 2012 breakthrough due to
 - ▶ Data (ex: ImageNet)
 - ▶ Computation (ex: GPU)
 - ▶ Architectures (ex: ReLU)

Success stories of deep learning in recent years

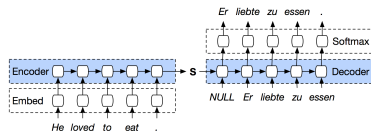
- ▶ Convolutional neural networks (CNNs)
- ▶ For stationary signals such as audio, images, and video
- ▶ Applications: object detection, image retrieval, pose estimation, *etc.*



Figure from [He et al., 2017]

Success stories of deep learning in recent years

- ▶ Recurrent neural networks (RNNs)
- ▶ For variable length sequence data, e.g. in natural language
- ▶ Applications: sequence to sequence prediction (machine translation, speech recognition) . . .



Input sentence:	Translation (PBM):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Images from: <https://smerity.com/media/images/articles/2016/> and <http://www.zdnet.com/article/google-announces-neural-machine-translation-to-improve-google-translate/>

It's all about the features ...

- ▶ With the right features anything is easy ...
- ▶ “Classic” vision / audio processing approach
 - ▶ Feature extraction (**engineered**) : SIFT, MFCC, ...
 - ▶ Feature aggregation (**unsupervised**): bag-of-words, Fisher vec.,
 - ▶ Recognition model (**supervised**): linear/kernel classifier, ...

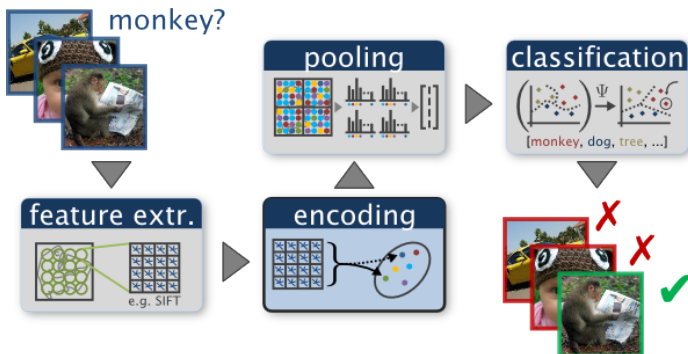
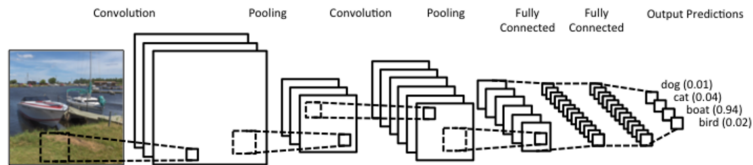


Image from [Chatfield et al., 2011]

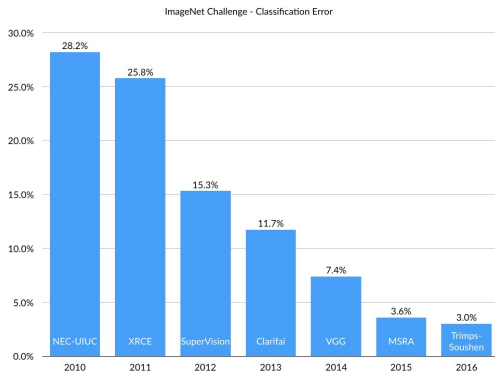
It's all about the features ...

- ▶ Deep learning blurs boundary feature / classifier
 - ▶ Stack of simple non-linear transformations
 - ▶ Each one transforms signal to more abstract representation
 - ▶ Starting from raw input signal upwards, e.g. image pixels
- ▶ Unified training of all layers to minimize a task-specific loss
- ▶ Supervised learning from lots of labeled data



Convolutional Neural Networks for visual data

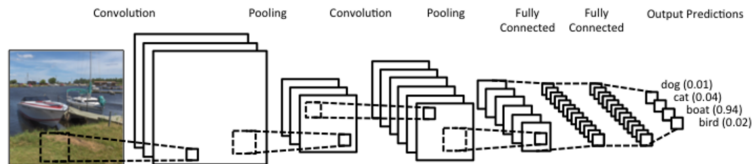
- ▶ Ideas from 1990's, huge impact since 2012 (roughly)
 - ▶ Improved network architectures
 - ▶ Big leaps in data, compute, memory
- ▶ ImageNet: 10^6 images, 10^3 labels



[LeCun et al., 1990, Krizhevsky et al., 2012]

Convolutional Neural Networks for visual data

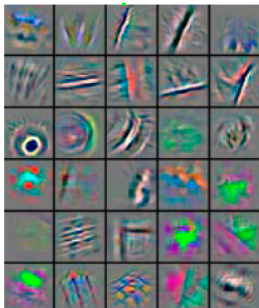
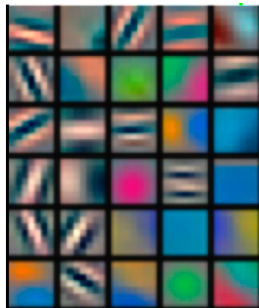
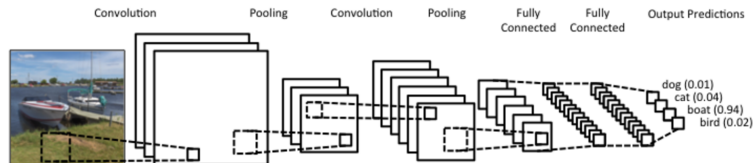
- ▶ Organize “neurons” as images, 2D grid
- ▶ Convolution computes activations from one layer to next
 - ▶ Translation invariant (stationary signal)
 - ▶ Local connectivity (fast to compute)
 - ▶ Nr. of parameters decoupled from input size (generalization)
- ▶ Pooling layers down-sample the signal every few layers
 - ▶ Multi-scale pattern learning
 - ▶ Degree of translation invariance



Example: image classification

Hierarchical representation learning

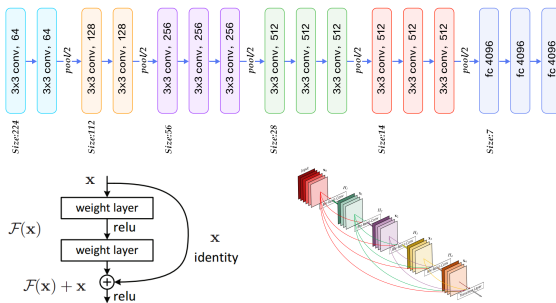
► Representations learned across layers



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Applications: image classification

- ▶ Output a single label for an image:
 - ▶ Object recognition: car, pedestrian, *etc.*
 - ▶ Face recognition: john, mary, ...
- ▶ Test-bed to develop new architectures
 - ▶ Deeper networks (1990: 5 layers, now >100 layers)
 - ▶ Residual networks, dense layer connections
- ▶ Pre-trained classification networks adapted to other tasks



Applications: Locate instances of object categories

- ▶ For example, find all cars, people, *etc.*
- ▶ Output: object class, bounding box, segmentation mask, ...



[He et al., 2017]

Applications: Scene text detection and reading

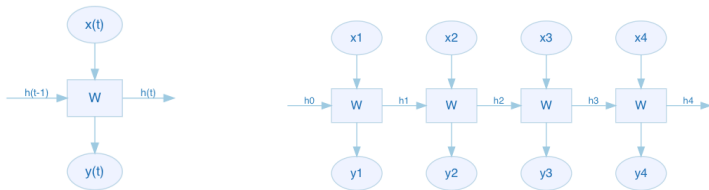
- ▶ Extreme variability in fonts and backgrounds
- ▶ Trained using synthetic data: real image + synth. text



Synthetic training data generated by [Gupta et al., 2016]

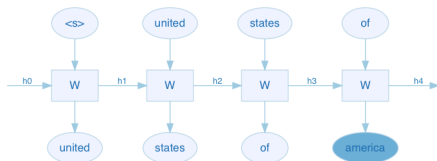
Recurrent Neural Networks (RNNs)

- ▶ Not all problems have fixed-length input and output
- ▶ Problems with sequences of variable length
 - ▶ Speech recognition, machine translation, etc.
- ▶ RNNs can store information about past inputs for a time that is not fixed a priori



Recurrent Neural Networks (RNNs)

- ▶ Example for language modeling
- ▶ Generative power of RNN language models



- ▶ Example of generation after training on Shakespeare

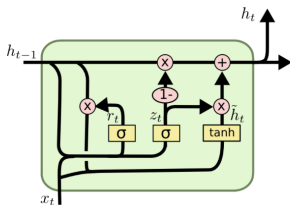
KING LEAR:

O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Figure from <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Handling Long Term Dependencies

- ▶ Problems if sequences are too long
 - ▶ Vanishing / exploding gradient
- ▶ Long Short Term Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997]
 - ▶ Learn to remember / forget information for long period of time
 - ▶ Gating mechanism
 - ▶ Now widely used (LSTMs or **GRUs**)



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

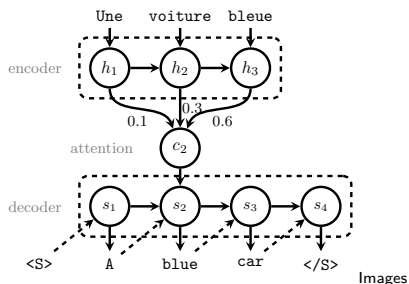
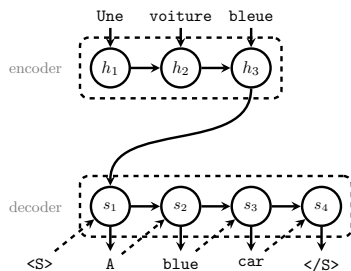
$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure from <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Applications: Neural Machine Translation

- ▶ End-to-End translation
- ▶ Most online machine translation systems (Google, Systran, DeepL) now based on this approach
- ▶ Map input sequence to a fixed vector, decode target sequence from it [Sutskever et al., 2014]
- ▶ Models later extended with attention mechanism [Bahdanau et al., 2014]



from Alexandre Berard's thesis

Applications: End-to-end Speech Transcription

- Architecture similar to neural machine translation
- Speech encoder based on CNNs or pyramidal LSTMs [Chorowski et al., 2015]

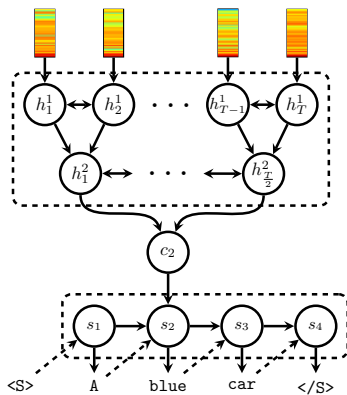
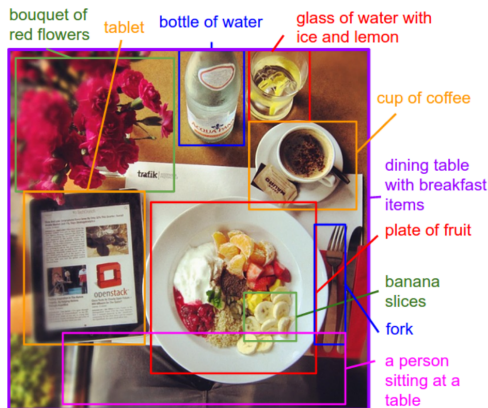


Image from Alexandre Berard's thesis

Applications: Natural language image description

- ▶ Beyond detection of a fixed set of object categories
- ▶ Generate word sequence from image data
- ▶ Image search, visually impaired, etc.



Example from [Karpathy and Fei-Fei, 2015]

Wrap-up — Take-home messages

- ▶ Core idea of deep learning
 - ▶ Many processing layers from raw input to output
 - ▶ Joint learning of all layers for single objective
- ▶ A strategy that is effective across different disciplines
 - ▶ Computer vision, speech recognition, natural language processing, game playing, *etc.*
- ▶ Widely adopted in large-scale applications in industry
 - ▶ Face tagging on Facebook over 10^9 images per day
 - ▶ Speech recognition on iPhone
 - ▶ Machine translation at Google, Systran, DeepL, *etc.*
- ▶ Open source development frameworks available (pytorch, tensorflow and the like)
- ▶ Limitations: compute and data hungry
 - ▶ Parallel computation using GPUs
 - ▶ Re-purposing networks trained on large labeled data sets

Outlook — Some directions of ongoing research (1/2)

- ▶ Optimal architectures and hyper-parameters
 - ▶ Possibly under constraints on compute and memory
 - ▶ Hyper-parameters of optimization: learning to learn (meta learning)
- ▶ Irregular structures in input and/or output
 - ▶ (molecular) graphs, 3D meshes, (social) networks, circuits, trees, etc.
- ▶ Reduce reliance on supervised data
 - ▶ Un-, semi-, self-, weakly- supervised, etc.
 - ▶ Data augmentation and synthesis (e.g. rendered images)
 - ▶ Pre-training, multi-task learning
- ▶ Uncertainty and structure in output space
 - ▶ For text generation tasks (ASR, MT): many different plausible outputs

Outlook — Some directions of ongoing research (2/2)

- ▶ Analyzing learned representations
 - ▶ Better understanding of black boxes
 - ▶ Explainable AI
 - ▶ Neural networks to approximate/verify long standing models and theories (link with cognitive sciences)
- ▶ Robustness to adversarial examples that fool systems
- ▶ Introducing prior knowledge in the model
- ▶ Biases issues (GenderShades and the like¹)
- ▶ Common sense reasoning
- ▶ etc.

¹Bolukbasi & al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520

References I

- [Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014).
Neural machine translation by jointly learning to align and translate.
CoRR, abs/1409.0473.
- [Chatfield et al., 2011] Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. (2011).
The devil is in the details: an evaluation of recent feature encoding methods.
In *BMVC*.
- [Chorowski et al., 2015] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015).
Attention-based models for speech recognition.
In *NIPS*.
- [Gupta et al., 2016] Gupta, A., Vedaldi, A., and Zisserman, A. (2016).
Synthetic data for text localisation in natural images.
In *CVPR*.
- [He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017).
Mask r-cnn.
arXiv, 1703.06870.
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Identity mappings in deep residual networks.
In *ECCV*.

References II

- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997).
Long short-term memory.
Neural Comput., 9(8):1735–1780.
- [Huang et al., 2017] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. (2017).
Densely connected convolutional networks.
In *CVPR*.
- [Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015).
Deep visual-semantic alignments for generating image descriptions.
In *CVPR*.
- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012).
Imagenet classification with deep convolutional neural networks.
In *NIPS*.
- [LeCun et al., 1990] LeCun, Y., Denker, J., and Solla, S. (1990).
Optimal brain damage.
In *NIPS*.
- [Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015).
Very deep convolutional networks for large-scale image recognition.
In *ICLR*.

References III

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. (2014).
Sequence to sequence learning with neural networks.
In *NIPS*.