

Lecture 1 - Introducing the deep learning revolution

DD2424

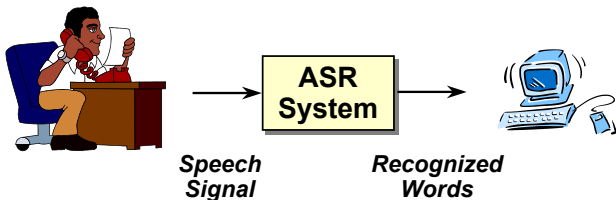
March 20, 2017

Why all the excitement about Deep Learning?

1. Astonishing empirical results.
2. Similar solutions for different tasks in different domains.
3. General formula for improving results:
deeper network + more training data + more computations

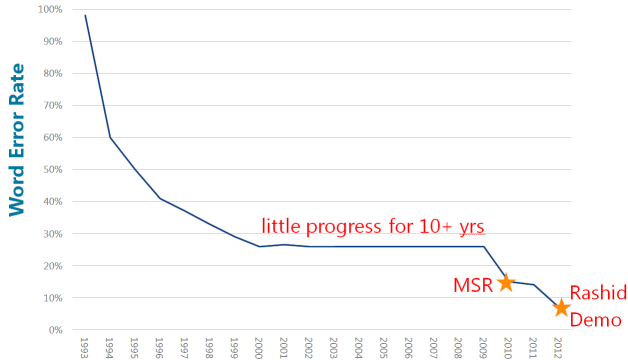
Snapshot of astonishing concrete results

Speech: Spontaneous Speech Recognition



ASR system's challenge is to convert the speech signal into words.

Deep learning \implies better speech recognition



After no improvement for 10+ years by the research community ... deep learning brings large improvements to speech recognition.

Computer Vision: Image Classification

ImageNet: Large Scale Visual Recognition Challenge

Steel drum



Output:

Scale
T-shirt
Steel drum
Drumstick
Mud turtle



Output:

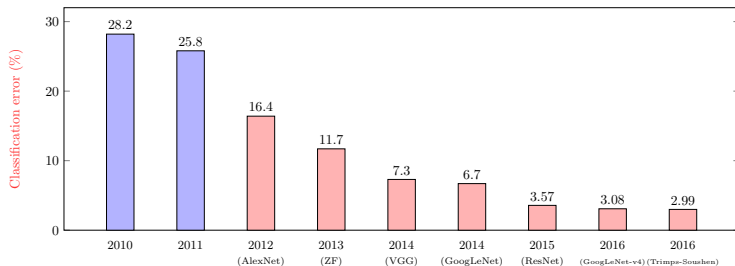
Scale
T-shirt
Giant panda
Drumstick
Mud turtle



$$\text{Error} = \frac{1}{100,000} \sum_{100,000 \text{ images}} 1(\text{incorrect on image } i)$$

Deep Learning → much better image classification

- ImageNet Large Scale Visual Recognition Challenge
- 1000 object classes, 1.4 million labelled training images



High performing systems on the ILSVRC datasets (2010-16).

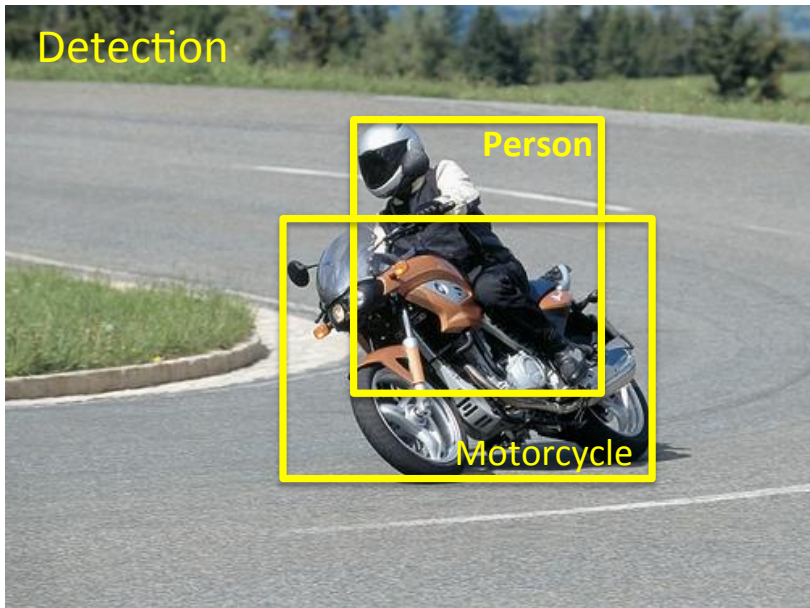
Pink indicates a deep learning based solution.

Deep Learning (ConvNets) \implies great image classification

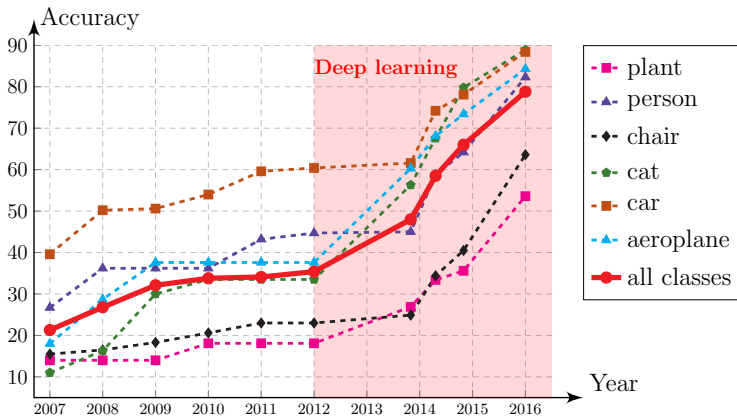
Detection

Person

Motorcycle



Deep Learning → much better object detection



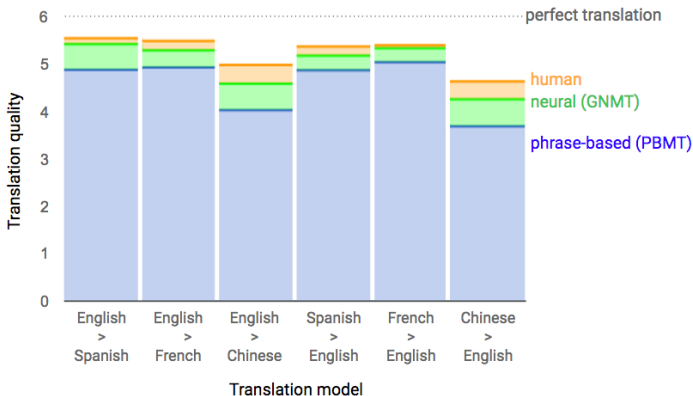
Progress of object detection for the Pascal VOC 2007 challenge.

Natural Language Processing: Text translation

- You have all probably used *Google Translate*
- It's been around for ~ 10 years.
- Up until Autumn 2016 it used **Phrase-Based Machine Translation**.
- But now ...

Deep Learning \implies better machine translation

Google Neural Machine Translation system



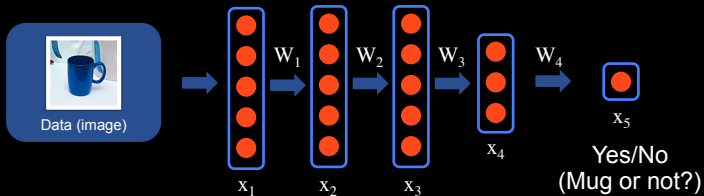
- Human raters compare the quality of translations of a source sentence.
- Scores range from 0 to 6.
 - 0 \equiv nonsense translation
 - 6 \equiv perfect translation

What methods/networks are producing these results?

What methods/networks are producing these results?

Neural Networks trained with lots of labelled data

What is a neural network?



Supervised learning (learning from tagged data)



Data:



Yes



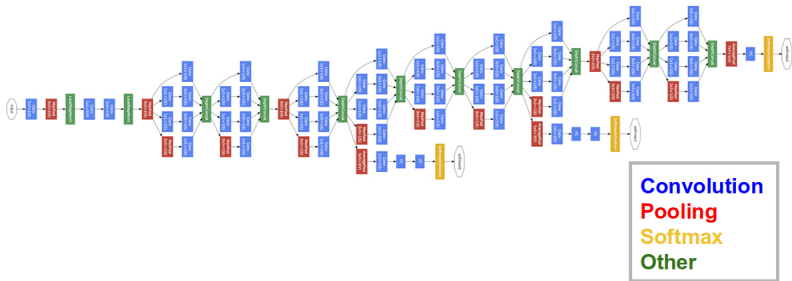
No

What methods/networks are producing these results?

What methods/networks are producing these results?

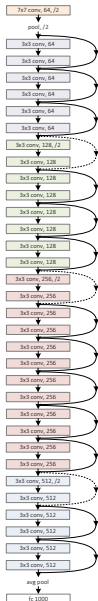
Deep Neural Networks trained with lots of labelled data

GoogLeNet (2014)



Trained using ImageNet to perform image classification.

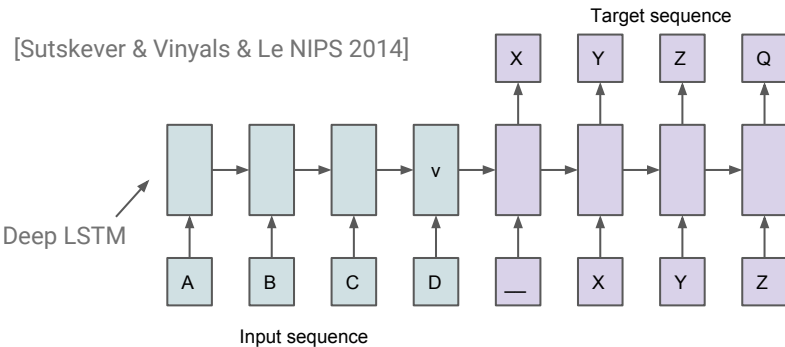
Example modern networks



- **ResNet** - a convolutional neural network with skip connections.
- Introduced in **Deep Residual Learning for Image Recognition**, by He, Zhang, Ren, Sun, CVPR 2016
- Trained for image classification, but similar structures have been transferred to image generation, speech recognition, NLP, ...

Sequence-to-Sequence Model

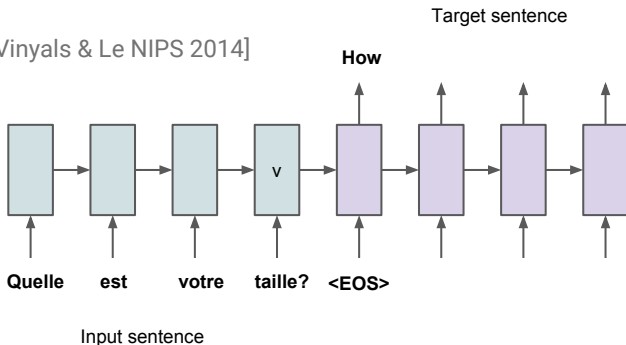
[Sutskever & Vinyals & Le NIPS 2014]



$$P(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots, y_{t-1})$$

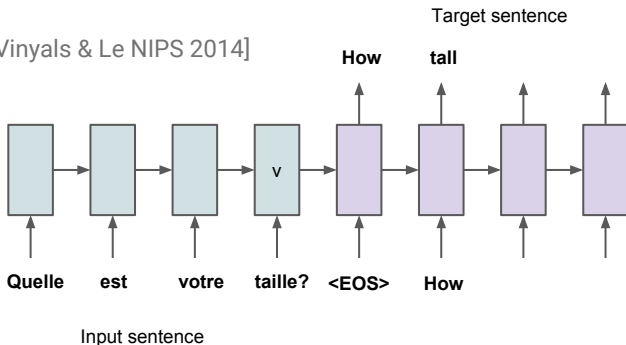
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



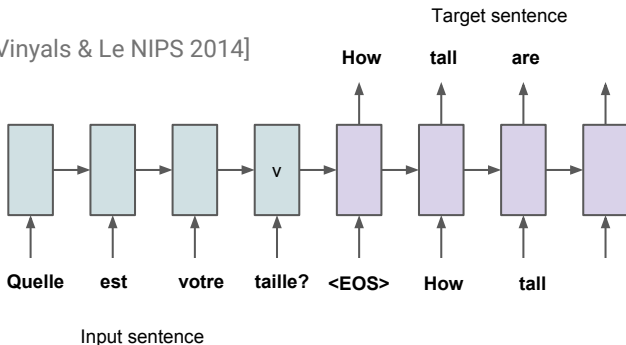
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



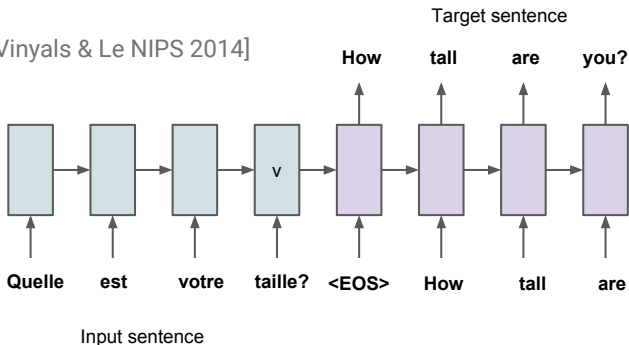
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



Sequence-to-Sequence Model: Machine Translation

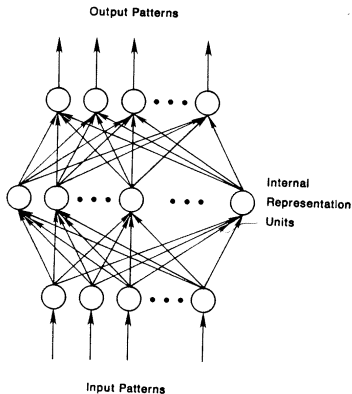
[Sutskever & Vinyals & Le NIPS 2014]



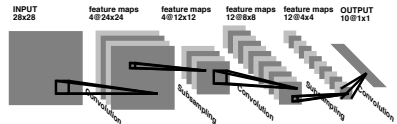
Why the great successes now?

Question:

Haven't these basic networks and their training algorithms been around for decades?



Back-prop Rumelhart in '86



LeCun's LeNet-1 '90

Why the great successes now?

Question:

Haven't these basic networks and their training algorithms been around for decades?

Answer:

Yes but now have

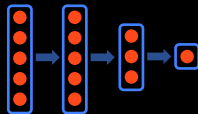
1. Explosion of labelled digital data available for training.
2. Deeper networks (networks with more layers)
3. Better understanding & procedures for learning the parameters of the networks.
4. GPUs \implies can exploit above to train deep networks in a "reasonable" time.

Why is Deep Learning taking off?



Engine

Fuel



Large neural networks

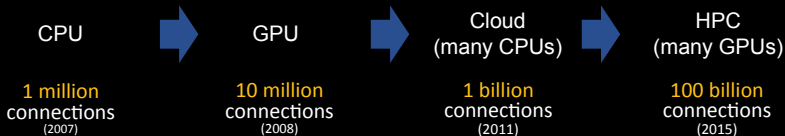


Data

Why is Deep Learning taking off?

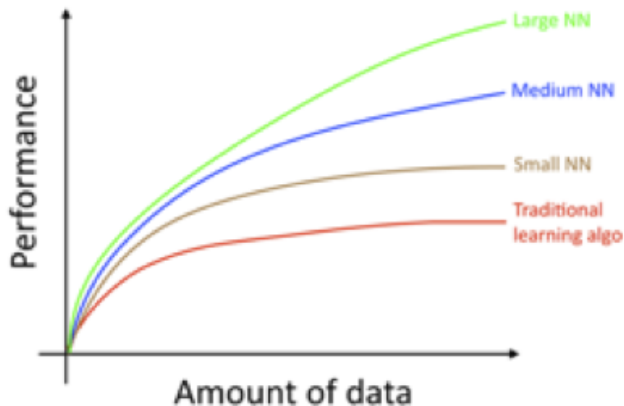


Rocket engines: Deep Learning driven by scale



Why didn't other approaches exploit these developments?

Other methods couldn't take advantage of large datasets



Reason 1: Deep Learning does end-to-end training

- Learn everything from the raw input data to desired output.
- Train hierarchical representations



as opposed to **prior approach** of hand engineering features



- Engineers job transferred from signal processing to network architecture & training algorithms.

Reason 2: Deep networks have efficient high capacity

- A neural network is a function f

$$f : \text{input space} \rightarrow \text{output space}$$

- **Universal approximation** (both shallow & deep):

Given a sufficient number of hidden nodes a 2-layer network can approximate any function.

- **Shallow networks not efficient representation:**

Some functions compactly represented with k layers in a network may require exponential size with 2 layers.

- \implies Deep networks are frequently a much more efficient representation of complicated functions than their shallow counterparts.

Reason 2: Deep networks efficient high capacity

- Deep network exploit **Compositionality**.
Complicated features are combinations of smaller, simpler features.
- Compositional features give an exponential gain in representational power over shallow representations.
- Compositionality is useful to describe the world around us efficiently.

First Success Story of Deep Learning: **Speech Recognition**

Problems in speech processing & recognition

- Speech (continuous time series) → Speech (continuous time series)
 - Speech Enhancement, Voice Conversion
- Speech (continuous time series) → Text (discrete symbol sequence)
 - Automatic speech recognition (ASR), Voice Activity Detection (VAD)
- Text (discrete symbol sequence) → Speech (continuous time series)
 - Text-to-speech synthesis (TTS)
- Text (discrete symbol sequence) → Text (discrete symbol sequence)
 - Machine translation (MT)

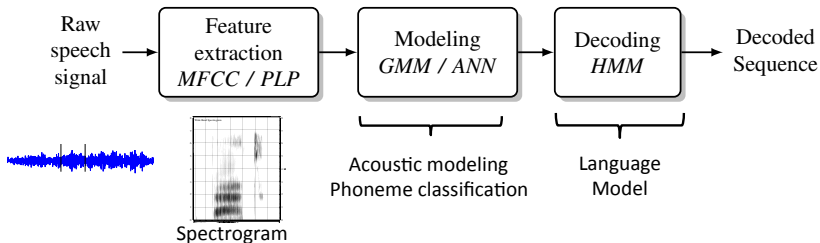
Problems in speech processing & recognition

- Speech (continuous time series) → Speech (continuous time series)
 - Speech Enhancement, Voice Conversion
- Speech (continuous time series) → Text (discrete symbol sequence)
 - Automatic speech recognition (ASR), Voice Activity Detection (VAD)
- Text (discrete symbol sequence) → Speech (continuous time series)
 - Text-to-speech synthesis (TTS)
- Text (discrete symbol sequence) → Text (discrete symbol sequence)
 - Machine translation (MT)

Will present a short history of recent developments in ASR.

Brief Aside - Speech

- Also huge impact by neural nets
- Traditional approach (pre-2009):



- Very incremental gains in performance

Progress of spontaneous speech recognition

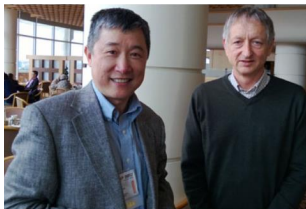


Best word error rates on the *Switchboard* dataset.

2400 two-sided phone conversations among 543 speakers (302 male, 241 female) from all areas of US.

Pivotal Academic-Industrial Collaboration

- Geoff Hinton collaboration with MSR, Redmond 2009-2010.

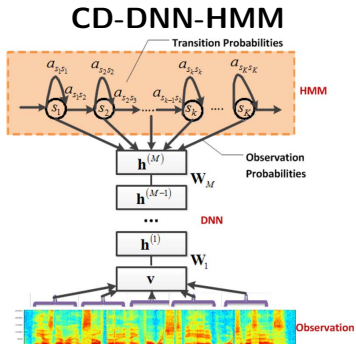


Deng (MSR, Redmond) & Hinton (University of Toronto)

- Mutually beneficial academic-industrial collaboration
 - *Automatic Speech Recognition (ASR)* industry looking for new approaches as progress had stalled.
 - Hinton had developed deep learning tools (Deep belief networks 2006) to train deep networks looking for applications and data.

- **Computing power and data available**
 - Advent of GPU computing. (Nvidia CUDA library released 2007/08)
 - Large labelled training sets data in speech were already available.
- **Algorithms/Approaches existed to train deep networks**
 - Layer-wise training of DBNs
 - Added supervised training, classic back-prop, to Hinton's deep generative models to train Deep Neural Networks.

Error rate on *Switchboard* dataset down to 18.5% from 27.4%




- Replace GMM of GMM-HMM with a deep neural network (DNN)
- For input to DNN use longer MFCC/filter-bank windows with no transformation.

Dahl, Yu, Deng, and Acero, *Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition*, IEEE Trans. ASLP, Jan. 2012 (also ICASSP 2011)

Seide, Li and Yu, *Conversational Speech Transcription Using Context-Dependent Deep Neural Network*, 2011

Deep Learning Technical Revolution

- First resurgence
 - A. Mohamed, G. Dahl and G. Hinton "*Deep belief networks for phone recognition*," In NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. 2009
 - DNNs for Large Scale Tasks
 - F. Seide, G. Li, and D. Yu, "*Conversational Speech Transcription Using Context-Dependent Deep Neural Networks*," in Proc. Interspeech 2011.
 - CNNs for Large Scale Tasks
 - T. N. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Deep Convolutional Neural Networks for LVCSR," in Proc. ICASSP, 2013.
 - LSTMs for Large Scale Tasks
 - H. Sak, A. Senior and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in Proc. Interspeech, 2014.
- 

[Slide: Tara Sainath, Google, Advancements in Deep Learning, SLT Keynote, Dec 2014.]

DNN Acoustic Modeling Results

- DNNs provide between a **8-25% relative improvement** in word error rate over GMM/HMM systems across a variety of tasks and languages
- Results confirmed by many, many research labs

	300 hour SWB Conversational Telephony	400 hour Broadcast News	2000 hour Voice Search
GMM/HMM	14.3	16.5	16.0
DNN	12.2	15.2	12.2
% Relative Improvement	14.7	7.9	23.8

[Slide: Tara Sainath, Google, Advancements in Deep Learning, SLT Keynote, Dec 2014.]

CNN vs DNN Results

- CNNs trained with vtlb-warped log-mel fb features offer between a **4-12% relative improvement** over DNNs trained with speaker-adapted features (VTLN +fMLLR)

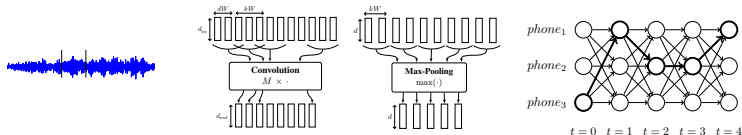
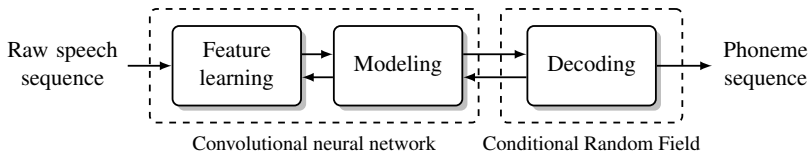
Model	BN-50	BN-400	SWB-300
Baseline GMM/HMM	18.1	13.8	14.5
DNN	15.8	13.3	12.2
CNN	15.0	12.0	11.5

[Sainath et al, ICASSP 2013]

[Slide: Tara Sainath, Google, Advancements in Deep Learning, SLT Keynote, Dec 2014.]

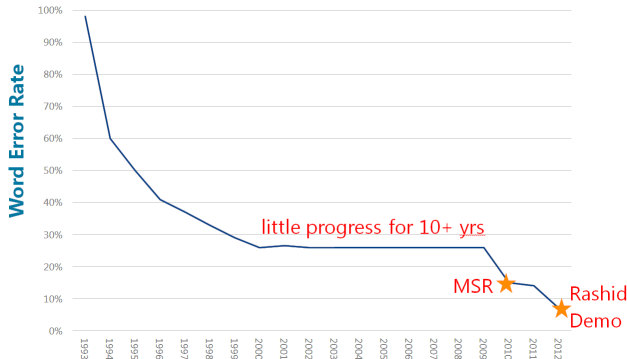
End-to-End Recognition

- Go directly from raw waveform



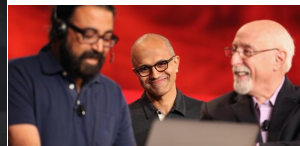
- Convolutional Neural Networks-based Continuous Speech Recognition using Raw Speech Signal, Palaz, Magimai-Doss, Collobert, ICASSP 2015.
- Superior results on TIMIT (phoneme recog), comparable results on WSJ.

Progress of spontaneous speech recognition



After no improvement for 10+ years by the research community ... MSR reduced error from $\sim 27\%$ to $< 13\%$ (and under $< 7\%$ for Rick Rashid's demo in 2012)!

Impact of deep learning in speech technology



Skype to get 'real-time' translator



Next Success Story of Deep Learning:
Image Recognition & Computer Vision

ImageNet Large Scale Visual Recognition Challenge

Steel drum



Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle



Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle



$$\text{Error} = \frac{1}{100,000} \sum_{100,000 \text{ images}} 1(\text{incorrect on image } i)$$

How well would a human perform on ImageNet?

- Andrej Karpathy, Stanford, set himself this challenge.
- Replicated the 1000 way classification problem for a human.
 - Person shown image on the left of figure.
 - On the right shown 13 examples from each of the 1000 classes.
 - Must pick 5 of these classes as the potential ground truth label.

The screenshot displays the ImageNet challenge interface. On the left is a large target image of a hotdog. Below it are two buttons: "Show answer" and "Show google prediction". Under these buttons, the Google prediction is shown: "hotdog, hot dog, red hot". Below the prediction are three input fields for selecting the top 5 classes. The first field contains "hotdog, hot dog, red hot" and has a "1" in a box. The second field contains "cheeseburger" and has a "1" in a box. The third field is empty. To the right of the input fields are 13 small images for each of the 1000 classes. The classes are grouped by color-coded headers: "consomme" (light green), "snack food" (pink), "hotdog, hot dog, red hot" (light blue), "hamburger, beefburger, burger" (pink), "cheeseburger" (light green), "course" (light green), "entree, main course" (pink), "plate" (light green), and "dessert, sweet, after" (pink). The "hotdog, hot dog, red hot" class is highlighted in light blue.

consomme

snack food

hotdog, hot dog, red hot

hamburger, beefburger, burger

cheeseburger

course

entree, main course

plate

dessert, sweet, after

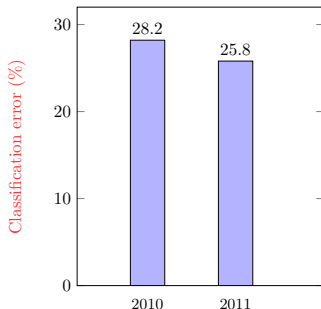
GoogleNet predictions:
hotdog, hot dog, red hot
ice cream, icecream
buckeye, horse chestnut, conker
French loaf
cheeseburger

How well would a human perform on ImageNet?

- Efforts and results reported on [his blog post](#).
- Estimated his own accuracy on ImageNet as 5.1%. (After some training period.)
- Later conjectured (Feb 2015) a dedicated and motivated human classifier capable of error rate in the range of 2%–3%.

State-of-the-art performance in 2011

- ImageNet Large Scale Visual Recognition Challenge
- 1000 object classes, 1.4 million labelled training images



Performance of winning entry in ILSVRC competitions (2010-11) prior to deep learning.

A whirlwind review of computer vision

The Birth of Computer Vision



Marvin Minsky



Gerald Sussman

1966

Marvin Minsky (MIT) asked his undergraduate student Gerald Jay Sussman to

“spend the summer linking a camera to a computer and getting the computer to describe what it saw.”

The Birth of Computer Vision



Marvin Minsky



Gerald Sussman

1966

Marvin Minsky (MIT) asked his undergraduate student Gerald Jay Sussman to

“spend the summer linking a camera to a computer and getting the computer to describe what it saw.”

Now know the problem was much more difficult.... ¹

¹ “You’ll notice that Sussman never worked in vision again!” – Berthold Horn

Why is it so hard? Consider object recognition

Challenges:

View Point Variation



Illumination Variation



Occlusion



Deformation



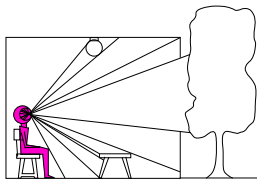
Intra-class variation



and haven't even mentioned clutter or NLP.

Common high-level road map for object recognition:

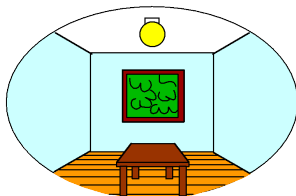
3D world



Point of observation



2D image



“Exploit the physics and geometry of imaging.”

- **Training:** Build 3D models of objects from 2D images of them.
- **Test time:**
 - Estimate object's *pose* in the image relative to camera's position.
 - Synthesize how the model would appear in the image.
 - Compare the synthetic image to the real image.

**Successes driven by this line of
research**



(a) Matier data set (7 images)



However progress on problem of object detection stalled

- Unclear how to mathematically model object categories.
- Unclear how best to spot and distinguish between instances of these models in images.

**Shift to learning based methods in the
naughties**

What we mean by learning based methods

Solve problem by referencing to training data

Have a face image if it looks like an image which I know is a face.

What we mean by learning based methods

Solve problem by referencing to training data

*Have a face image if it **looks like** an image which I know is a face.*

Trend fueled by the

- rapid growth of computational power,
- rapid growth of memory and
- the abundance of digital images and video and the web.

Recognition in Computer Vision 101

You see this



34	45	53	55	69	79	91	95	105	197	254	250	254	254	254	254	254	254	254	254
0	11	20	39	59	58	62	73	67	92	213	255	254	254	254	254	254	254	254	254
5	5	0	11	30	16	39	87	67	27	167	255	254	254	254	254	254	254	254	254
0	0	10	12	8	5	73	172	172	140	204	255	254	254	254	254	254	254	254	254
5	0	17	0	0	20	123	237	249	255	246	250	254	254	254	254	254	254	254	254
0	16	9	0	0	48	200	255	242	255	255	255	255	255	254	252	251	252	253	254
7	0	0	5	23	175	234	250	243	250	253	254	251	251	251	252	253	253	254	254
0	0	17	0	17	198	255	248	250	246	255	245	254	255	255	255	255	253	252	250
0	16	2	14	69	125	247	255	255	247	255	249	253	253	253	254	253	253	252	251
26	15	1	109	181	102	148	235	254	240	249	252	250	250	250	250	251	252	254	255
0	0	44	203	249	169	69	208	255	255	248	255	255	255	255	255	255	253	251	250
4	47	156	232	255	245	115	166	244	253	249	245	244	247	252	255	255	254	251	249
114	193	251	253	247	255	191	88	153	185	207	182	200	209	224	240	251	255	255	255
193	255	255	213	147	131	97	63	59	77	86	81	88	110	123	112	156	199	250	245
228	178	151	113	9	4	17	40	43	32	42	68	65	53	36	74	70	75	121	151
171	52	33	0	13	0	0	31	44	29	32	55	61	72	71	107	91	55	62	165
47	0	17	11	40	28	22	33	52	68	76	80	78	101	119	110	124	63	74	170
21	22	19	30	26	45	59	60	64	77	85	84	93	101	125	120	117	30	56	213
35	40	27	51	52	51	57	66	66	55	48	49	92	108	108	101	52	0	18	195
27	19	52	89	56	31	19	34	45	41	40	47	67	66	39	15	18	45	51	151

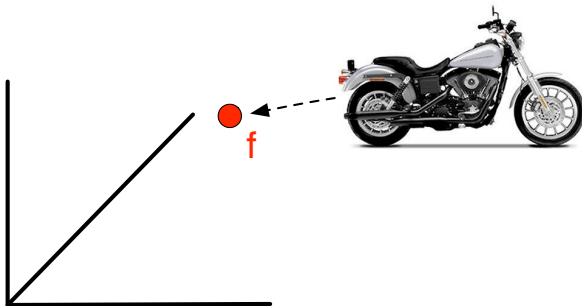
but a computer sees this

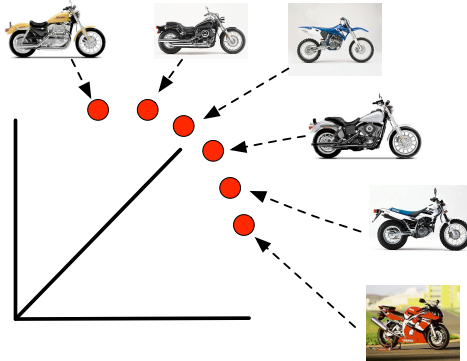
34	45	53	55	69	79	91	95	105	197	254	250	254	254	254	254	254	254	254	254
0	11	20	39	59	58	62	73	67	92	213	255	254	254	254	254	254	254	254	254
5	5	0	11	30	16	39	87	67	27	167	255	254	254	254	254	254	254	254	254
0	0	10	12	8	5	73	172	172	140	204	255	254	254	254	254	254	254	254	254
5	0	17	0	0	20	123	237	249	255	246	250	254	254	254	254	254	254	254	254
0	16	9	0	0	48	200	255	242	255	255	255	255	255	254	252	251	252	253	254
7	0	0	5	23	175	234	250	243	250	253	254	251	251	252	252	253	253	254	254
0	0	17	0	17	198	255	248	250	246	255	245	254	255	255	255	255	253	252	250
0	16	2	14	69	125	247	255	255	247	255	249	253	253	253	254	253	253	252	251
26	15	1	109	181	102	148	235	254	240	249	252	250	250	250	250	251	252	254	255
0	0	44	203	249	169	69	208	255	255	248	255	255	255	255	255	255	253	251	250
4	47	156	232	255	245	115	166	244	253	249	245	244	247	252	255	255	254	251	249
114	193	251	253	247	255	191	88	153	185	207	182	200	209	224	240	251	255	255	255
193	255	255	213	147	131	97	63	59	77	86	81	88	110	123	112	156	193	250	245
228	178	151	113	9	4	17	40	43	32	42	68	65	53	36	74	70	75	121	215
171	52	33	0	13	0	0	31	44	29	32	55	61	72	71	107	91	55	62	165
47	0	17	11	40	28	22	33	52	68	76	80	78	101	119	110	124	63	74	170
21	22	19	30	26	45	59	60	64	77	85	84	93	101	125	120	117	30	56	213
35	40	27	51	52	51	57	66	66	55	48	49	92	108	108	101	52	0	18	195
27	19	52	89	56	31	19	34	45	41	40	47	67	66	39	15	18	45	51	159

$$\Rightarrow \mathbf{f} = (f_1, f_2, \dots)$$

Convert pixel data to a feature vector.

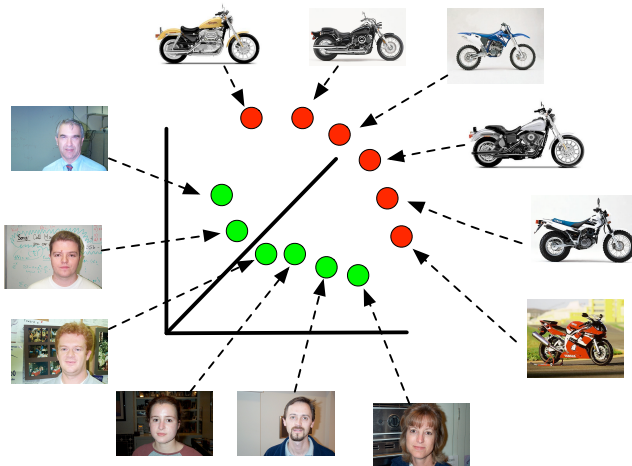
Feature extraction turns image into a point





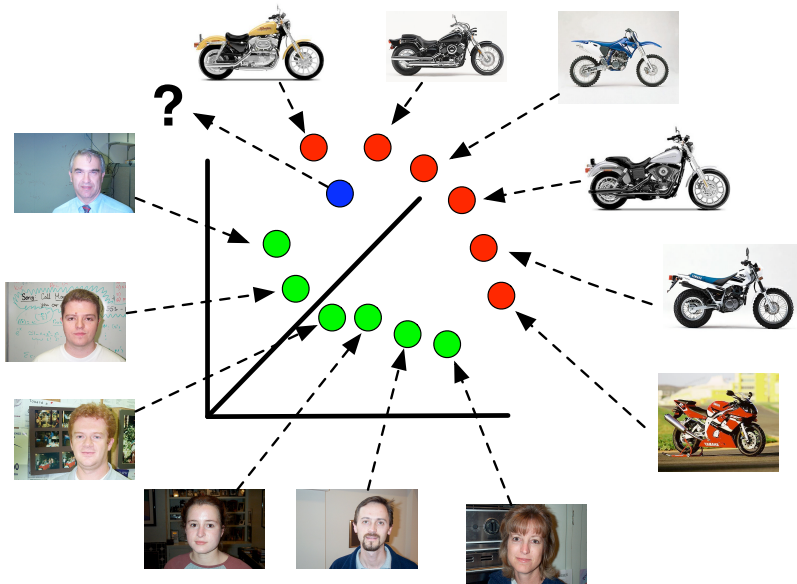
Many feature vectors from a category

Learning from examples

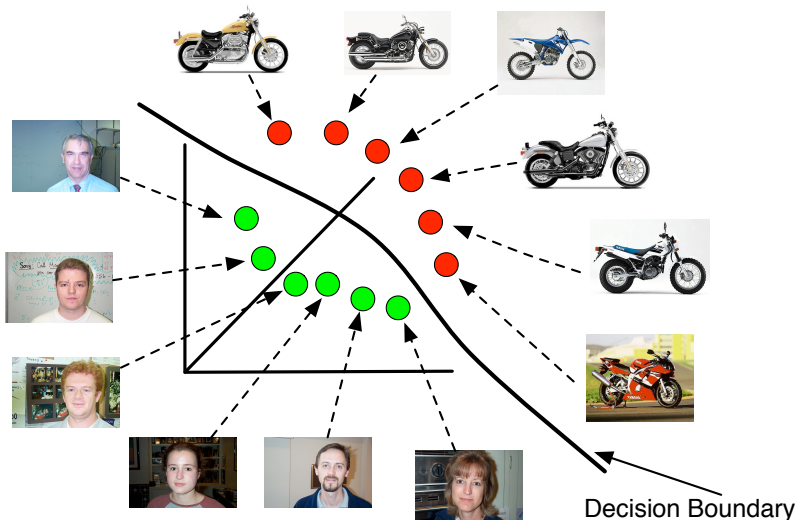


Want different categories to occupy different volumes.

Is it a bike or a face ?



Construct a decision boundary



1. Training Phase

- Gather labelled training data.
- Extract a feature representation for each training example.
- Construct a decision boundary.

2. Test Phase

- Extract feature representation from the test example.
- Compare to the learnt decision boundary.

1. Training Phase

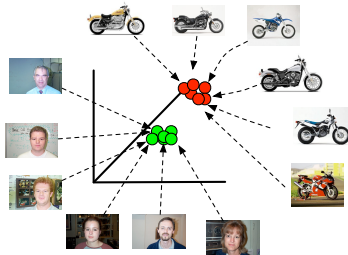
- Gather labelled training data.
- Extract a feature representation for each training example.
- Construct a decision boundary.

2. Test Phase

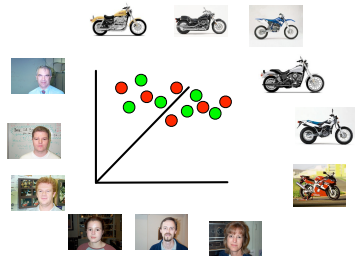
- Extract feature representation from the test example.
- Compare to the learnt decision boundary.

This is supervised learning.

Success depends mainly on quality of feature extraction



Ideal features



Far from ideal

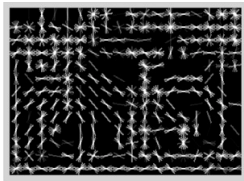
Naughties focused on hand-crafted features

- Engineer/researcher designing and constructing ingenious features.
- Let machine learning do feature selection and refining of features.

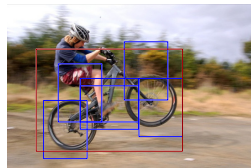
Popular features



intensity template

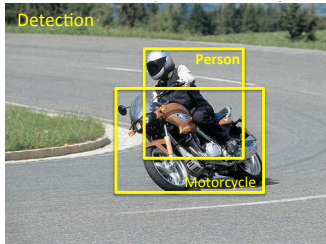


HOG

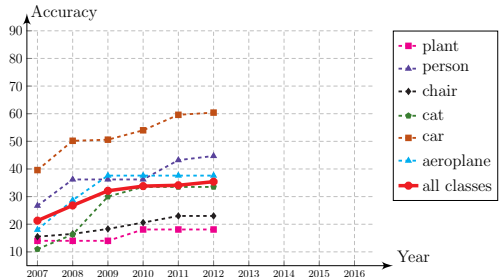


deformable part models

Progress at first but then stagnation



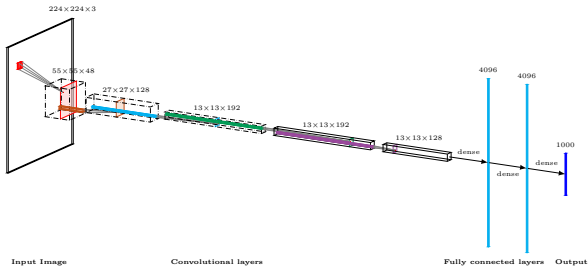
Task of object detection



But then ...

ImageNet 2012: Most exciting CV workshop ever

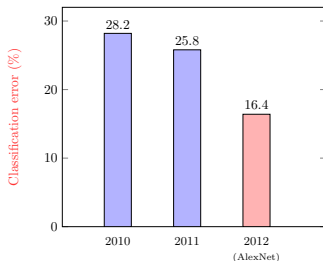
- Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton from University of Toronto present **AlexNet**.



- First modern deep Convolutional Network trained using Backprop to solve a hard computer vision problem.
- Outperforms all competitors by a large margin.

Impact of AlexNet on state-of-the-art performance

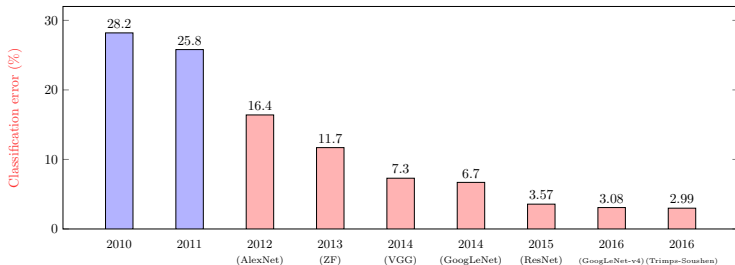
- ImageNet Large Scale Visual Recognition Challenge
- 1000 object classes, 1.4 million labelled training images



Performance of winning entry in ILSVRC competitions (2010-12).

AlexNet was only the start

- ImageNet Large Scale Visual Recognition Challenge
- 1000 object classes, 1.4 million labelled training images



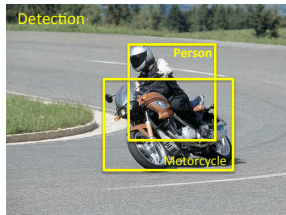
High performing systems on the ILSVRC datasets (2010-16).

Pink indicates a ConvNet based solution.

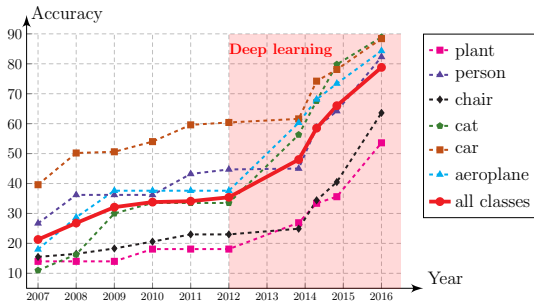
Deep ConvNets \implies great for image classification

What about Object Detection?

Deep ConvNets → much better object detection



Task of object detection



Progress on the Pascal VOC 2007 challenge.

Deep learning allows direct learning of better features

Key properties of deep learning

Provides a mechanism to:

- Learn a highly non-linear function.
(Efficiently encoded in a deep structure.)
- Learn it from data.
- Build feature hierarchies
 - Distributed representations
 - Compositionality
- Perform **end-to-end** learning (no more hand-crafted features)

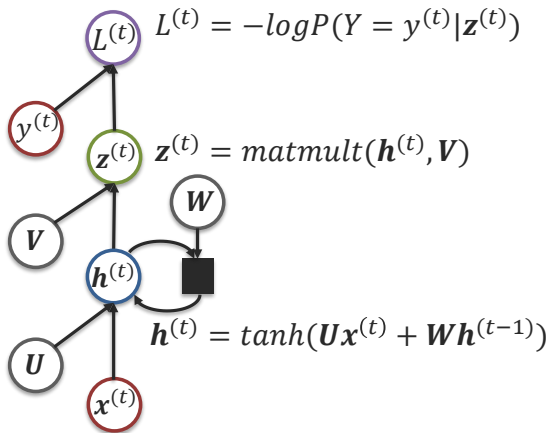
Current most exciting application domain of Deep Learning:
Natural Language Processing

Language Modelling - Sequence Modelling

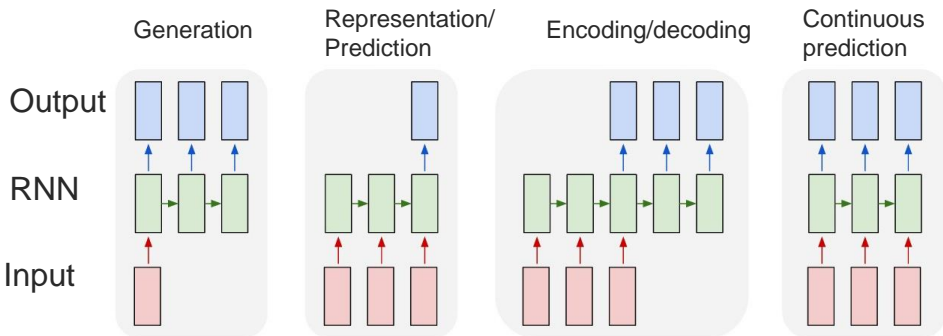
- Need a way to computationally model sequences of words - language.
- Need ways of representing language.

Recurrent Neural Networks (and variants) are the deep learning approach.

Recurrent Neural Networks are popular to model sequences



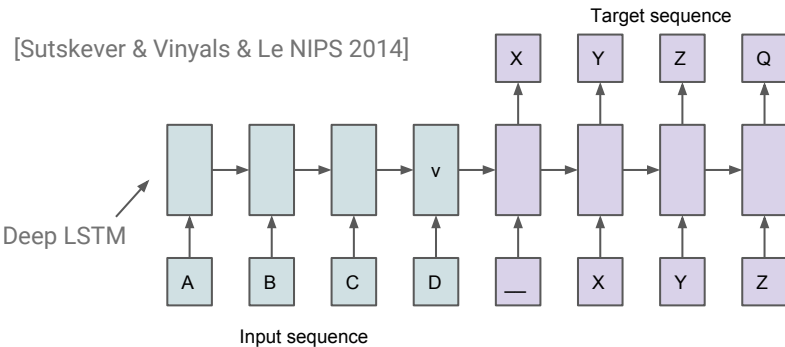
Use cases of RNNs



[\[http://karpathy.github.io/2015/05/21/rnn-effectiveness/\]](http://karpathy.github.io/2015/05/21/rnn-effectiveness/)

Sequence-to-Sequence Model

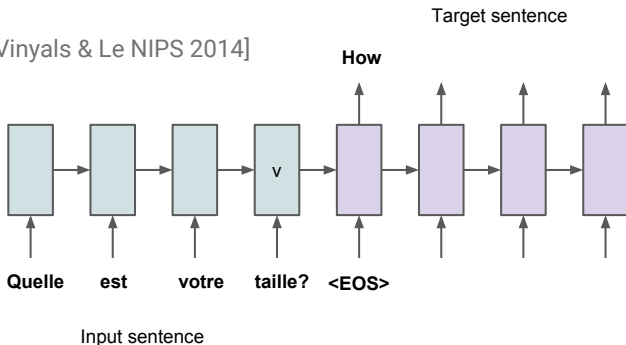
[Sutskever & Vinyals & Le NIPS 2014]



$$P(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots, y_{t-1})$$

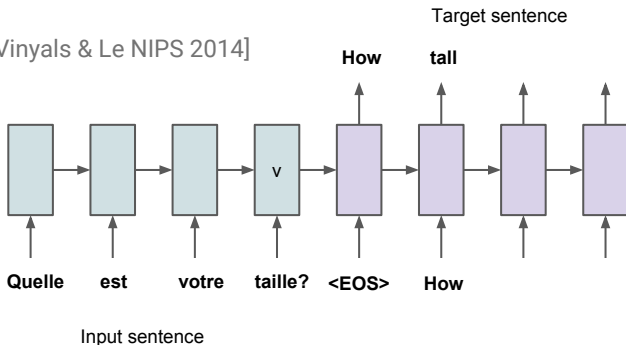
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



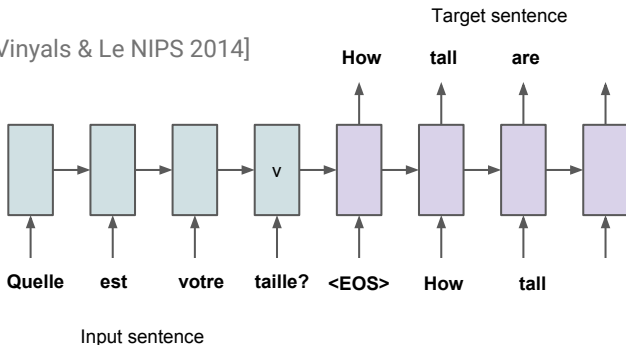
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



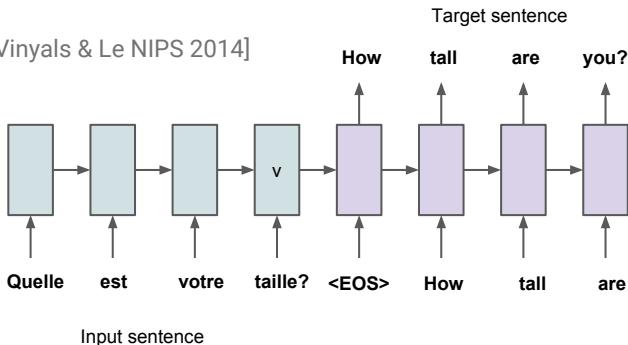
Sequence-to-Sequence Model: Machine Translation

[Sutskever & Vinyals & Le NIPS 2014]



Sequence-to-Sequence Model: Machine Translation

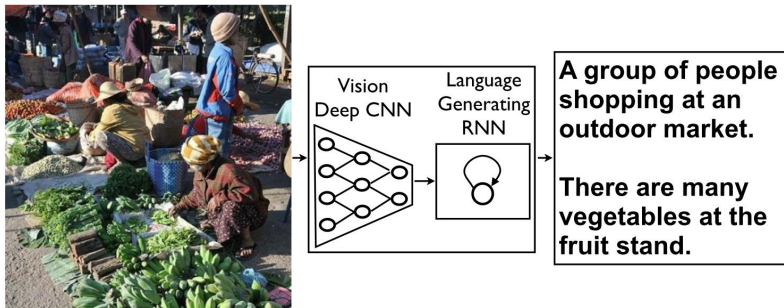
[Sutskever & Vinyals & Le NIPS 2014]



Can also have

Multi-modal translation

Image captioning with RNNs



[Vinyals et al., "Show and Tell: A Neural Image Caption Generator", CVPR 2015]

Image captioning with RNNs

- Same training as before, but now the encoder is a CNN
- We can train a system end to end using cross entropy loss
- Often use already pre-trained CNN and RNN models
 - CNN on visual object classification
 - RNN on language modeling
 - Can also train a coordinated multimodal representation space as well
- Training is done on pairs of images and captions
- Datasets
 - MS COCO
 - Flickr8k
 - Flickr30k



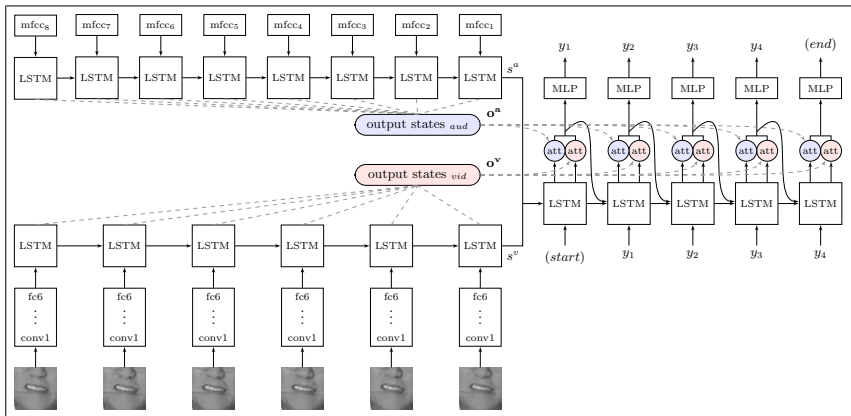
The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.



Translation from Speech & Video → Text



To review...

The rise of end-to-end learning

- Learning with integer or real-valued outputs

Problem	Input	Output
Spam classification	Email	Spam/Not spam (0/1)
Image recognition	Input	Integer label
House pricing prediction	Features of house	Price in dollars
Product recommendation	Product & user features	Chance of purchase

- Learning with complex outputs

Problem	Input	Output
Image captioning	Image	Text
Machine translation	English text	French text
Question answering	(Text, Question) pair	Answer text
Speech recognition	Audio	Transcription
TTS	Text features	Audio

Major Categories of Deep Learning Models

1. General neural networks
2. Sequence models (1D sequences)
 - Recurrent Neural Network (RNN)
 - Gated Recurrent Unit (GRU)
 - Long short-term memory (LSTM)
 - Connectionist Temporal Classification (CTC)
 - attention models
3. Image models
 - 2D and 3D convolutional networks
4. Advanced /future technology
 - Unsupervised learning (sparse coding, ICA, autoencoders)
 - Reinforcement learning
 - Self-supervised learning (GANS)

The ones we'll cover in the course

1. General neural networks
2. Sequence models (1D sequences)
 - Recurrent Neural Network (RNN)
 - Gated Recurrent Unit (GRU)
 - Long short-term memory (LSTM)
 - Connectionist Temporal Classification (CTC)
 - Attention models
3. Image models
 - 2D and 3D convolutional networks
4. Advanced /future technology
 - Unsupervised learning (sparse coding, ICA, autoencoders)
 - Reinforcement learning
 - Self-supervised learning (GANS)

**Deep neural networks are making significant strides in
understanding for
speech, vision, language, ...
and translating between the different modalities.**

- It's a technology that is (and will be) at the forefront of automating the interpretation (and manipulation) of input signals at the capacity of a human (expert & beyond).
- It's going to be fun to teach you all about the techniques, maths, learning algorithms, and networks underpinning these developments....

**Deep neural networks are making significant strides in
understanding for
speech, vision, language, ...
and translating between the different modalities.**

- It's a technology that is (and will be) at the forefront of automating the interpretation (and manipulation) of input signals at the capacity of a human (expert & beyond).
- It's going to be fun to teach you all about the techniques, maths, learning algorithms, and networks underpinning these developments....

Course Admin & Your Workload

- **Assignments (4.5 hp)**
 - Four programming assignments
 - Pen and Paper exercises may be interspersed within these assignments
- Either **Project (3.0 hp)** or **Written Exam (3.0 hp)**
 - **Project**
 - * Freely chosen project in DL completed in groups of 3.
 - **Take Home Written Exam**
 - * Answer theory and problem questions about the material covered in the course.

Assignments: Pass / Fail + Possibility for Bonus Points

There will be 4 programming assignments

- To pass an assignment task you must
 - Make good attempt / finish the coding task.
 - Write a short report containing result and intermediary result figures and description on the structure and debugging of your code etc. More instructions will be given in the assignment specification.
 - Upload your code and pdf report to Bilda.
- To get **bonus points** you can complete optional extensions to the basic assignment. Here the bonus points will be towards your final grade in the course.

More details about the assignments

Important Dates:

- When the assignments will be released:
 - Assignment 1: 27th of March (after lecture)
 - Assignment 2: 30th of March (after lecture)
 - Assignment 3: 6th of April (after lecture)
 - Assignment 4: 18th of April (after lecture)
- Deadline for the submission of assignments:
 - Assignment 1: 11th of April
 - Assignment 2: 18th of April
 - Assignment 3: 25th of April
 - Assignment 4: 9th of May

Note The submission deadlines are not hard as regard passing the course etc. They are hard in the sense that if you miss one then I cannot guarantee that I will review/correct the assignment in a timely fashion.

Project: Grade A-F

- In groups of 3 you will complete project in deep learning.
- Your project will be assessed based on:
 - **Written Report**
 - **Oral Examination**
- The grade will be mainly be based on the **Written Report** but the **Oral Examination** will confirm to me whether all members of the group were involved and knowledgeable about the technical/experimental detail submitted in the project.

Project: Grade A-F

- You will be free to use whatever software package you like - TensorFlow, Torch, ...
- You will be free to choose the topic and scope.
- Before starting with your project you will need to get a (short) proposal approved.
- Make sure attend you some “Project help” session to get some advice about the feasibility of your project idea.

Important Dates:

- Your deadline for submission of project proposal: **18th April**
- My deadline for approving your project proposal: **21st April**
- Deadline for submission of final project report: **19th May**
- Project Oral Examination Period: **22nd May – 2nd June**

Sign-up for an account at PDC:

Once you have formed your project group of three:

- One member of the group should apply for an account at PDC and when you have done so fill this [google form](#) to let me know.
- Please do this as soon as possible because
 - PDC needs some time to create and confirm these accounts.
 - Training and running multiple experiments on deep networks can be very time-consuming.
 - The more computational resources you have access to, the more interesting things you can do.
 - I can then arrange PDC to give an instructional lecture about using their machines.

Written Exam: A-F

- Written take home exam.
- The exam will have a compulsory part - **Part I** - covering the core topics of the course. This part will have to be answered at a high level to get a grade E.
- To obtain higher grades other questions, other topics and **Part I** material at a higher level of understanding, will have to be answered in **Part II**.
- Nearer the exam, I will be very explicit about the content and the level of difficulty of **Part I** and **Part II** of the exam.

Final Grade: A-F

If you pass the **Written Exam** or the **Project** then your final grade will either be

- Written Exam Grade + Bonus Points or
- Project Grade + Bonus Points

Bonus Points will potentially be able to bump you up one grade (E → D, D → C, etc)

You cannot use the Bonus Points to lift your failing grade to a passing one.

More details about the Project Help sessions

- You will book a 15 minute time slot with a TA. (Doodle will be made available soon on the KTH Social webpage.)
- Please book a slot with your 2 other project partners. **We do not have the resources to meet students individually.**
- Attendance is not mandatory.
- But you should attend some to get advice on your project.
- You can also get some advice/guidance with the assignment.

- I **cannot** register you for the course!
- You have to ask your student adviser or fill in the appropriate forms online.
- If you cannot see the DD2424 bilda page by the end of the week contact me and I will get you added.