Deep Learning for Visual and Multimodal Recognition
Séminaire TIM 2017
Traitement de l'Information Multimodale

Nicolas Thome
Conservatoire National des Arts et Métiers (Cnam)
Équipe MSDMA - Département Informatique
Prenom.Nom@cnam.fr
`http://cedric.cnam.fr/~thomen/`

6 Juillet 2017

Context

Big Data: Images & videos everywhere

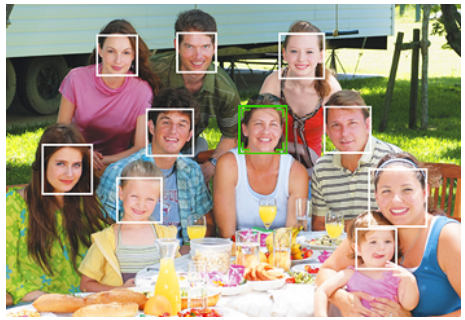
BBC: 2.4M videos

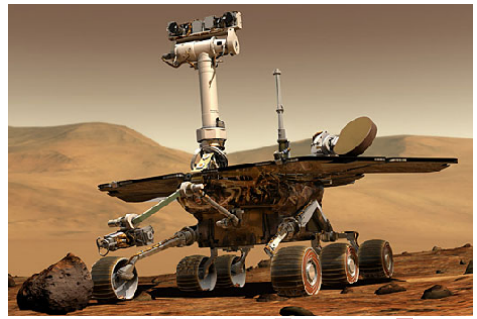
Social media,

***e.g.* Facebook: 1B each day**

100M monitoring cameras

- Obvious need to access, search, or classify these data: **Visual Recognition**
- Huge number of applications: mobile visual search, robotics, autonomous driving, augmented reality, medical imaging *etc*
- Leading track in major ML/CV conferences during the last decade

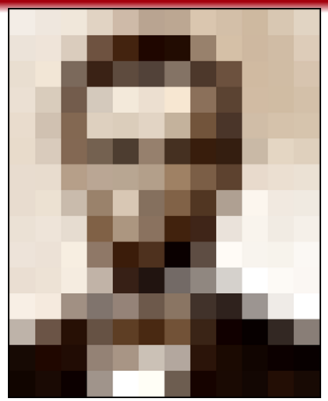




Outline

Recognition of low-level signals

Challenge: filling the semantic gap



What we perceive *vs*
What a computer sees

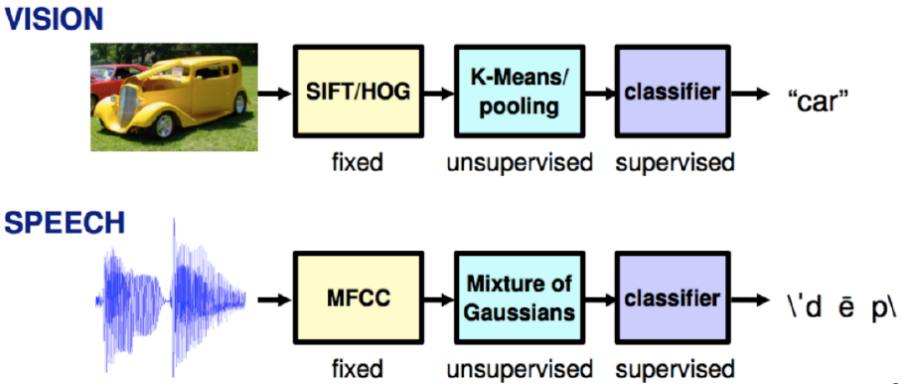
243	239	240	225	206	185	188	218	211	206	216	225
242	239	218	110	67	31	34	152	213	206	208	221
243	242	123	58	94	82	132	77	108	208	208	215
235	217	115	212	243	236	247	139	91	209	208	211
233	208	131	222	219	226	196	114	74	208	213	214
232	217	131	116	77	150	69	56	52	201	228	223
232	232	182	186	184	179	159	123	93	232	235	235
232	236	201	154	216	133	129	81	175	252	241	240
235	238	230	128	172	138	65	63	234	249	241	245
237	236	247	143	59	78	10	94	255	248	247	251
234	237	245	193	55	33	115	144	213	255	253	251
248	245	161	128	149	109	138	65	47	156	239	255
190	107	39	102	94	73	114	58	17	7	51	137
23	32	33	148	168	203	179	43	27	17	12	8
17	26	12	160	255	255	109	22	26	19	35	24

- Illumination variations
- View-point variations
- Deformable objects
- intra-class variance
- *etc*

⇒ **How to design "good" intermediate representation ?**

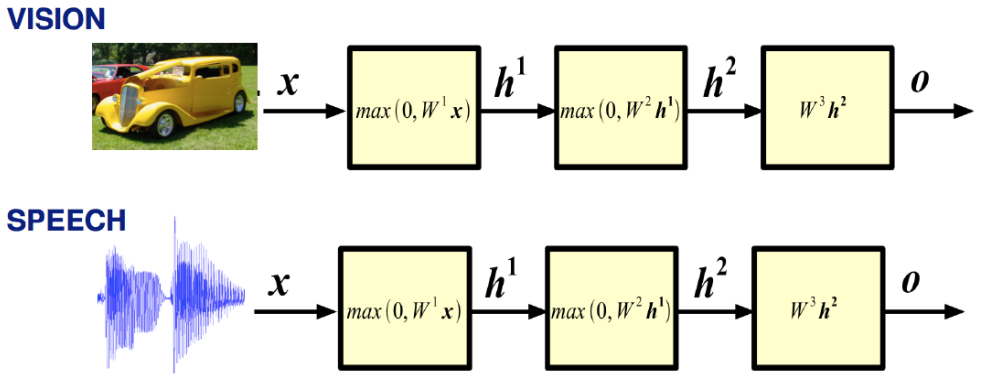
Deep Learning (DL) & Recognition of low-level signals

- Before DL: **handcrafted intermediate representations** for each task
  - ⊖ Needs expertise in each field
  - ⊖ Shallow: low-level features



Credit: I. Kokkinos

- Since DL: **Representation Learning**
  - ⊕ Deep: hierarchy, gradually learning higher-level representations
  - ⊕ Common learning methodology ⇒ field independent, no expertise

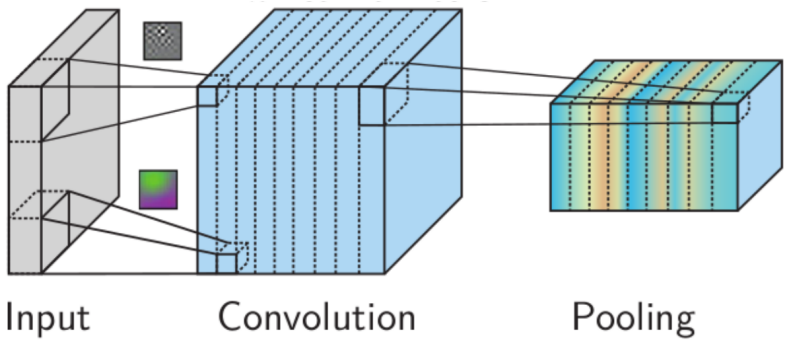


Credit: I. Kokkinos

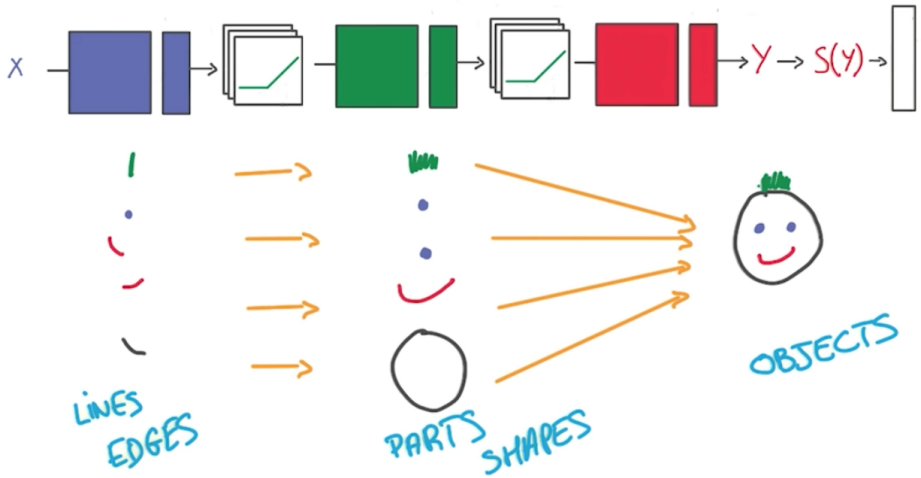
- **All parameters trained with backpropagation with class labels**

Convolutional Neural Networks (ConvNets)

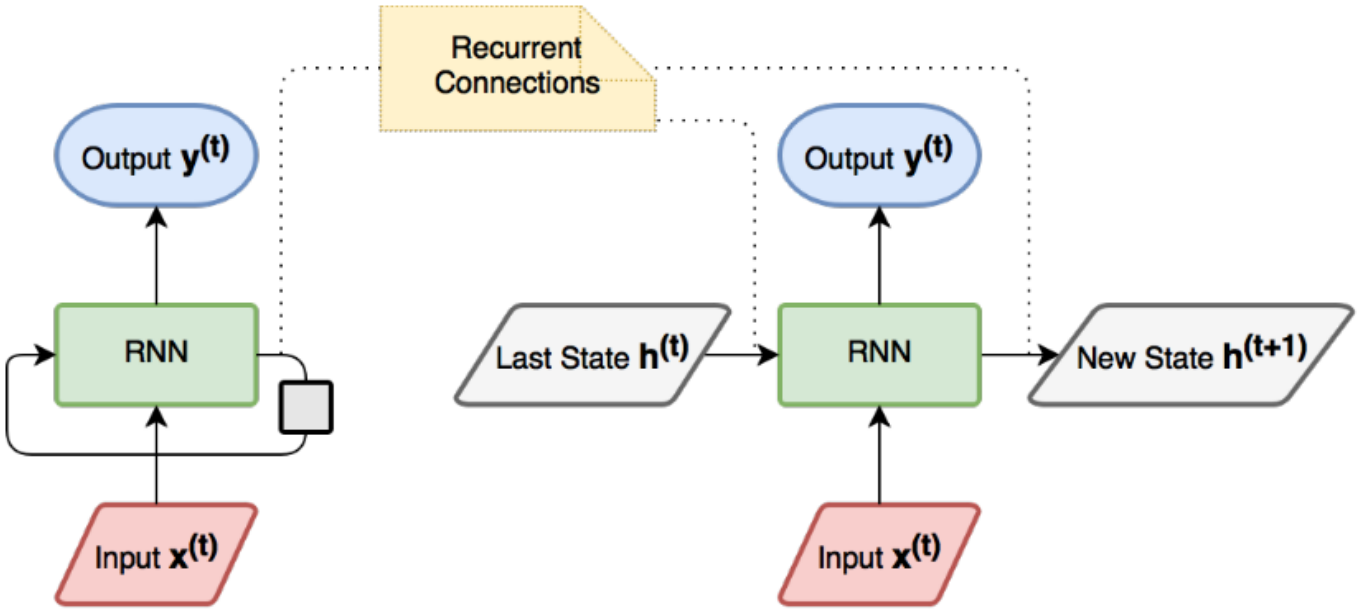
- Convolution on tensors, *i.e.* multidimensional arrays: T of size $W \times H \times D$
  - Convolution: $C[T] = T'$, T' tensor of size $W' \times H' \times K$
  - Each filter locally connected with shared weights (K number of filters)



- **An elementary block: Convolution + Non linearity (*e.g.* ReLU) + pooling**
- **Convolution:** structure (local processing), **Pooling:** invariance
- **Stacking several Blocks:** intuitive hierarchical information extraction



Recurrent Neural Networks (RNNs) for Sequence Modeling

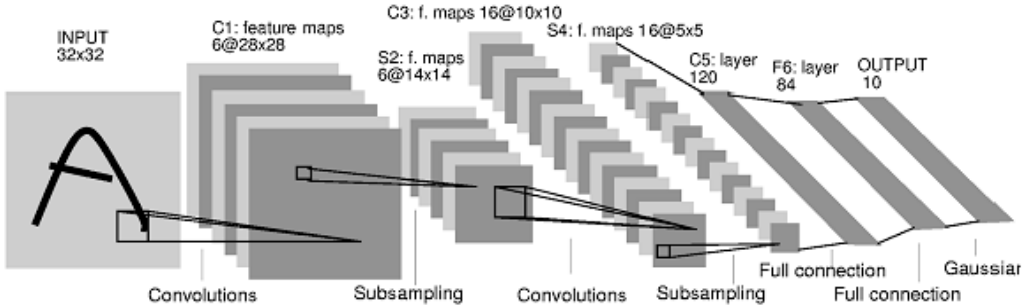


- **Sequences:** 1d/2d signals (*e.g.* audio, videos), molecules, text,*etc*
- Input vector $x(t)$, *e.g.* word (text) or image representation (CNN)
- Input/Output $h(t)$: vector representing model "short-term memory"
- Output vector $y(t)$: task dependent
- All parameters trained with backpropagation through time

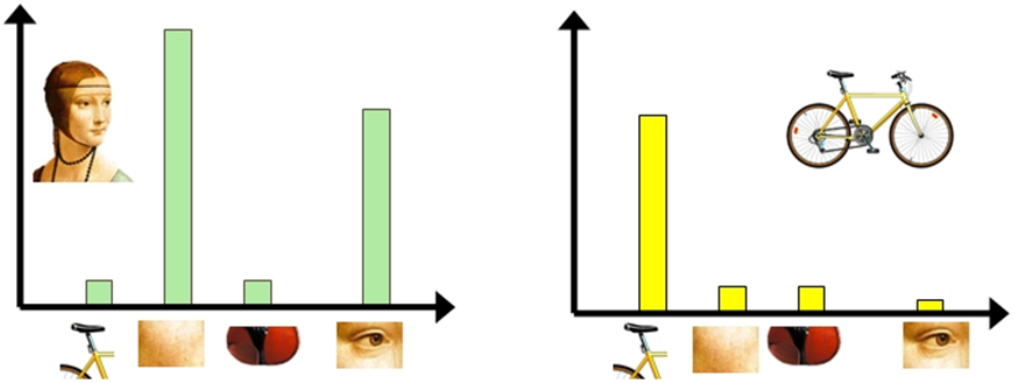
Outline

Visual Recognition History: Trends and methods in the last four decades

- 80's: training Convolutionnal Neural Networks (CNN) with back-propagation ⇒ postal code reading [LBD+89]

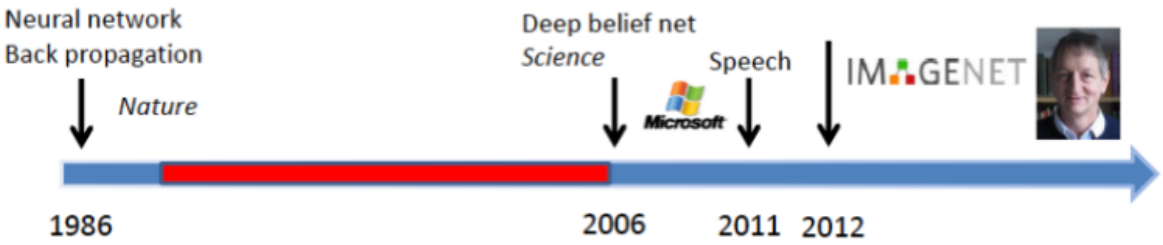


- 90's: golden age of kernel methods, NN = black box
- 2000's: BoW + SVM : state-of-the-art CV



Visual Recognition History: Trends and methods in the last four decades

- Deep learning revival in 2012: CNN success of ConvNets in ImageNet [KSH12]



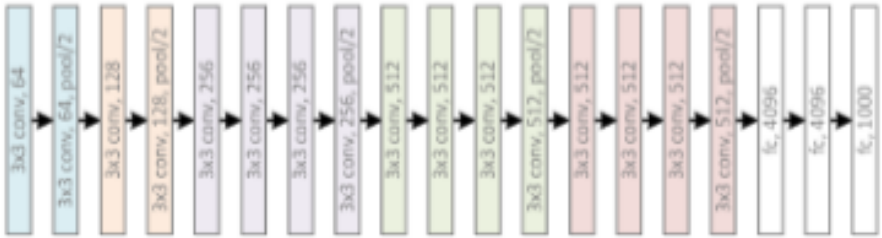
Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

- **Two main practical reasons:**
  1. Huge number of labeled images (10^6 images)
  2. GPU implementation for training

IMAGENET

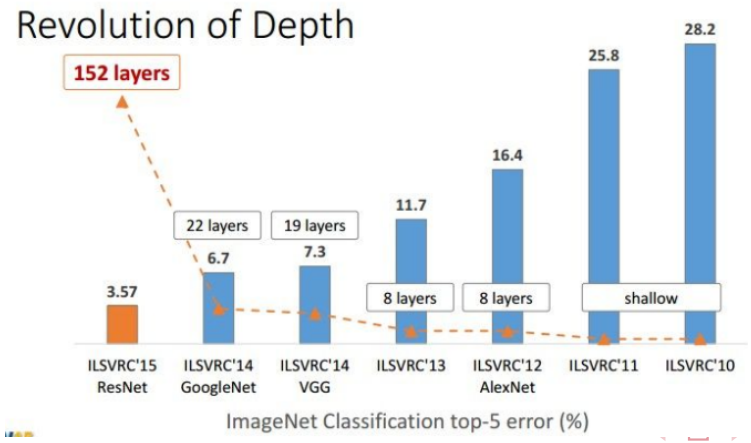
Deep Learning since 2012: Larger & larger networks

VGG, 16/19 layers, 2014



GoogleNet, 22 layers, 2014

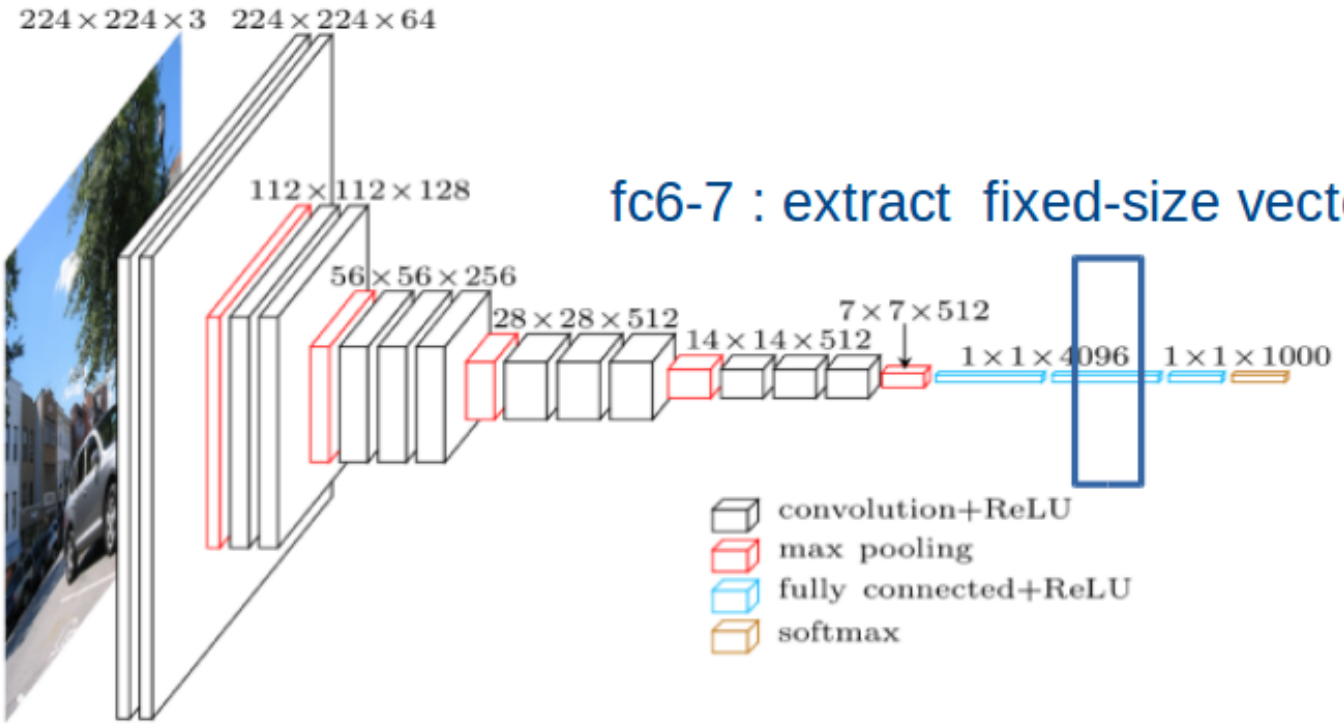
ResNet, 152 layers, 2015



Deep Learning since 2012

Transferring Representations learned from ImageNet

- Deep ConvNets require large-scale annotated datasets
- **BUT:** Extract layer ⇒ fixed-size vector: **"Deep Features" (DF)**



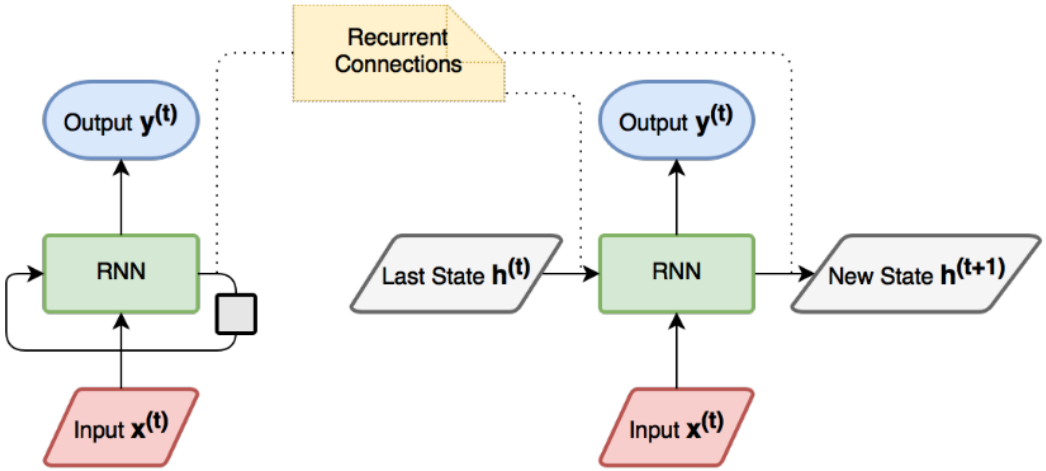
- **Now state-of-the-art for any visual recognition task** [ARS+16]
- **Ex: Domain adaptation for image classif:**
  - **DF very robust to data variations, *e.g.* medical / astronomy images**

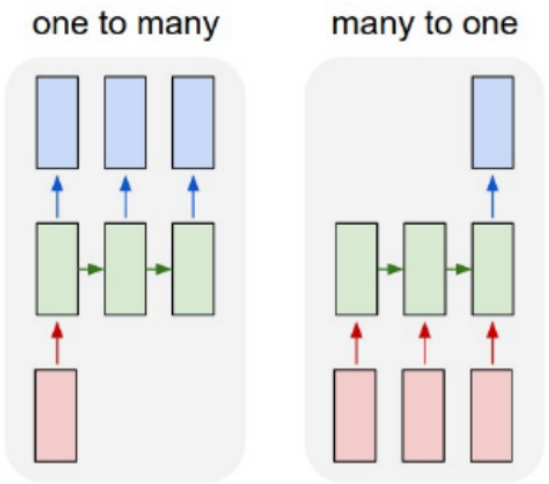
Outline

Ongoing Issues in Deep Learning

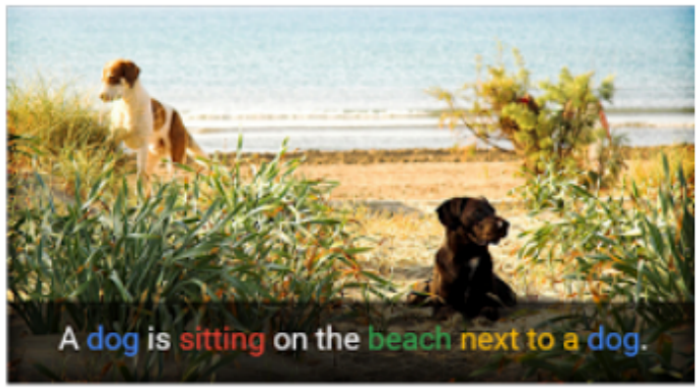
New Tasks in Artificial Intelligence

- Intersection of vision and language research:
  - Improvements in Vision Understanding with ConvNets
  - Language (text) Modeling with RNNs





- One to Many: Image captioning
- Many to One: Visual Question Answering



Many to One - Visual Question Answering (VQA)

- Goal: build a system that can answer questions about images

How many slices of pizza are there?
Is this a vegetarian pizza?

Does it appear to be rainy?
Does this person have 20/20 vision?

What color are her eyes?
What is the mustache made of?

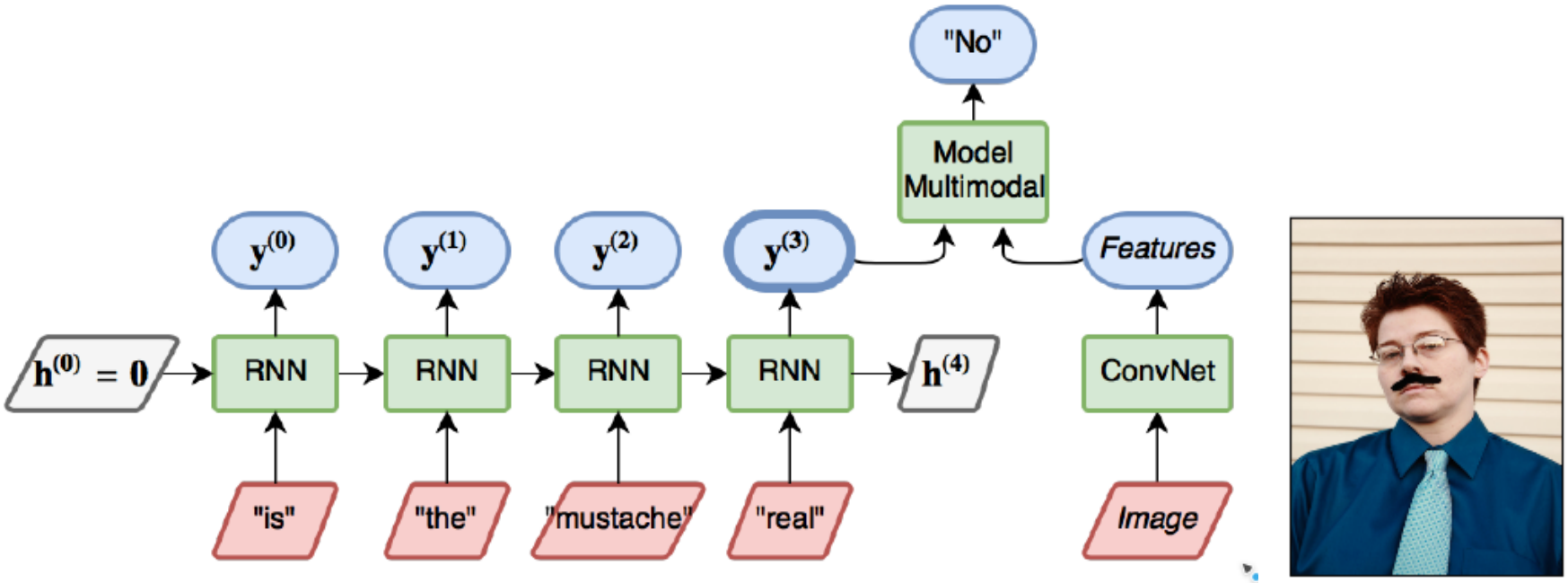
- Very complex task, that requires :
  - Precise image and text models
  - High level interaction modeling
  - Full scene understanding
  - Reasoning (*e.g.* spatial ...)

What color is the fire hydrant on the right ? yellow

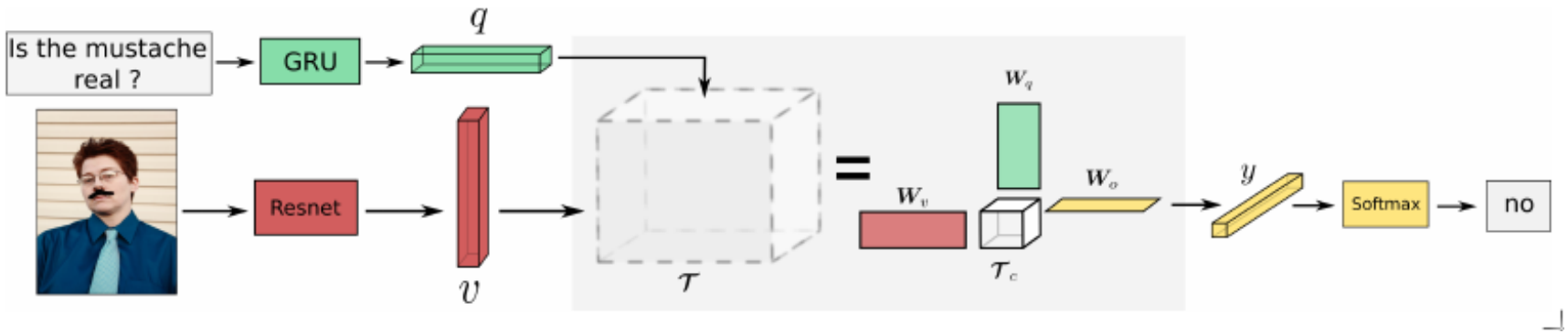
What color is the fire hydrant on the left ? green

Many to One - Visual Question Answering (VQA)

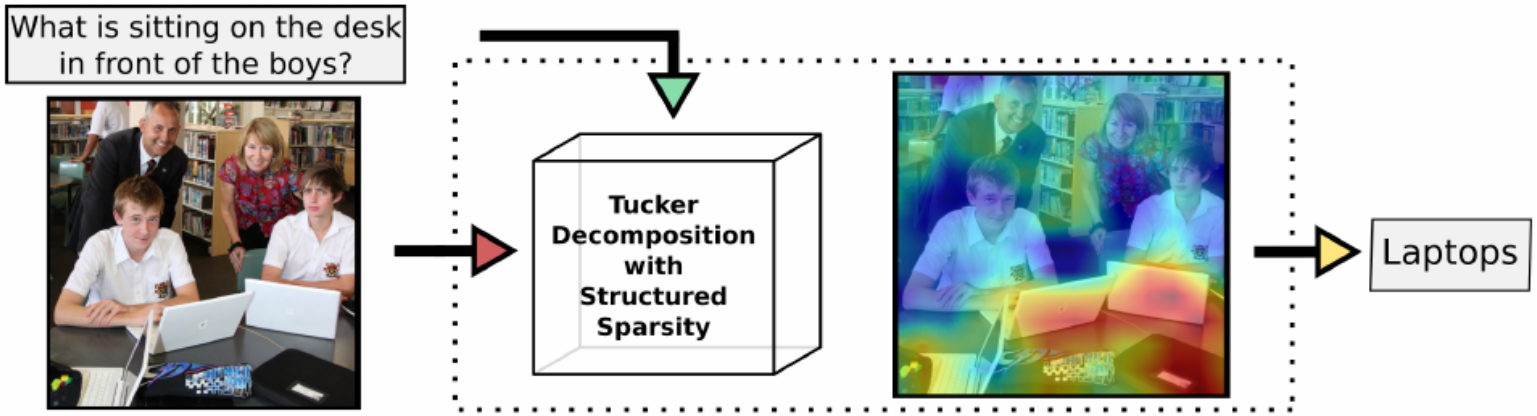
- Input: question & image
- Output: answer



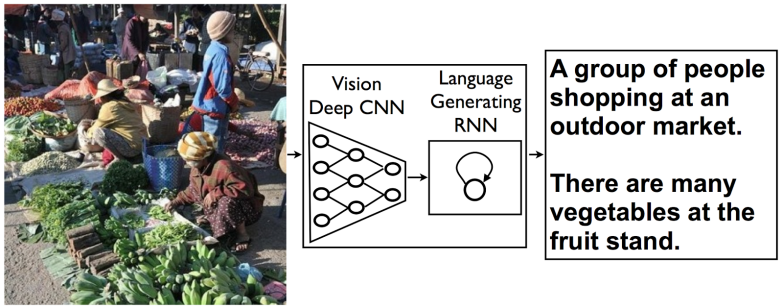
Multimodal Tucker Fusion for Visual Question Answering (MUTAN)

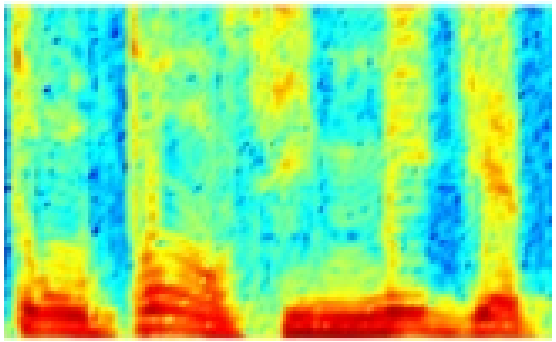


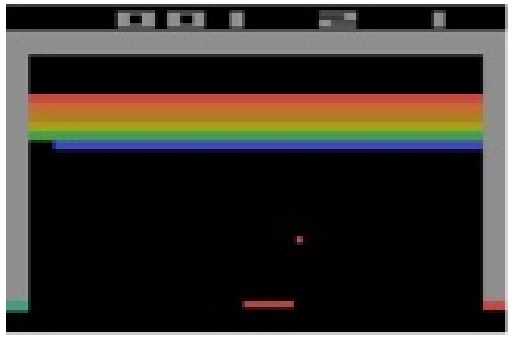
- State-of-the-art mono-modal representations:
  - Visual representation: ResNet-152
  - Question representation: pre-trained GRU (Gated Recurrent Units)
- How to perform multi-modal fusion ?
  - State-of-the-art: bilinear models [FPY+16, KOK+17] ⇒ accurate interactions
  - BUT full bilinear models intractable: factorization based on Tucker decomposition [BCCT17]



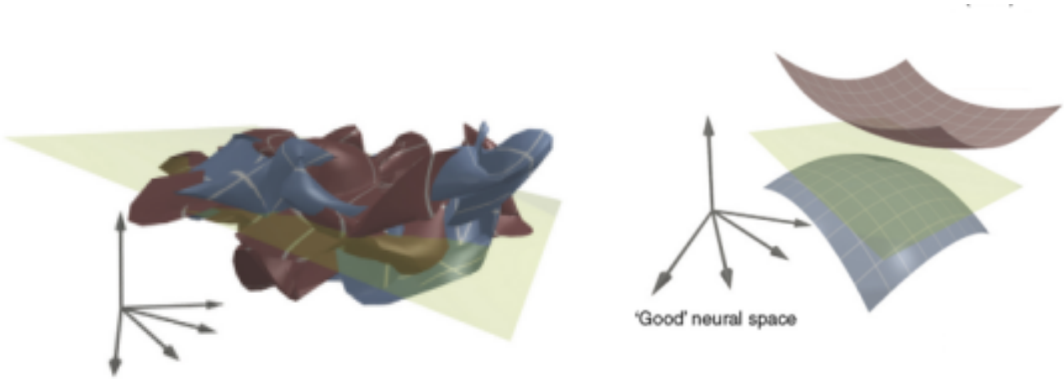
Conclusion







- Deep Learning: huge impact in terms of experimental results
- Still a long way to go toward real AI: reasoning, memory, predictive models, common knowledge, *etc*
- Formal understanding still limited:
  - Optimization: non-convex problem
  - Model: ability to untangle manifold
  - Robustness to over-fitting & generalization



References I

Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson, *Factors of transferability for a generic convnet representation*, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016), no. 9, 1790–1802.

Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome, *MUTAN: multimodal tucker fusion for visual question answering*, CoRR abs/1705.06676 (2017).

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, *Multimodal compact bilinear pooling for visual question answering and visual grounding*, arXiv:1606.01847 (2016).

Jin-Hwa Kim, Kyoung-Woon On, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang, *Hadamard Product for Low-rank Bilinear Pooling*, 5th International Conference on Learning Representations, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012, pp. 1097–1105.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, *Backpropagation applied to handwritten zip code recognition*, Neural computation 1 (1989), no. 4, 541–551.