DEEP LEARNING AND THE NEURAL NET RNNAISENSE

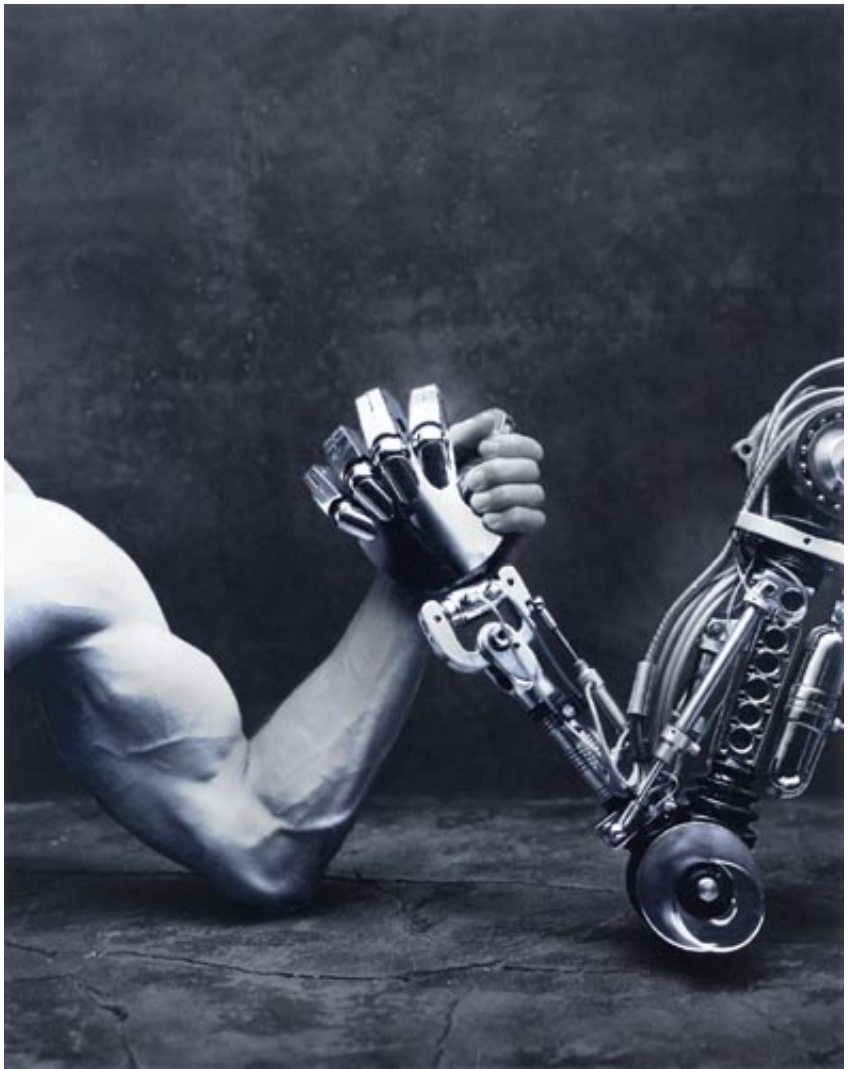
SGAICO – 2014
JONATHAN MASCI
UNIVERSITY OF LUGANO - IDSIA

ULTIMATE GOAL

Understand what are the principles that give rise to intelligence and develop models able to implement them

DEEP LEARNING

Provides a set of learning tools which are narrowing the gap between humans and machines



Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

MIT Techn Revie

Google to buy artificial intelligence company DeepMind

by Signe Brewster JAN. 26, 2014 - 7:39 PM PST

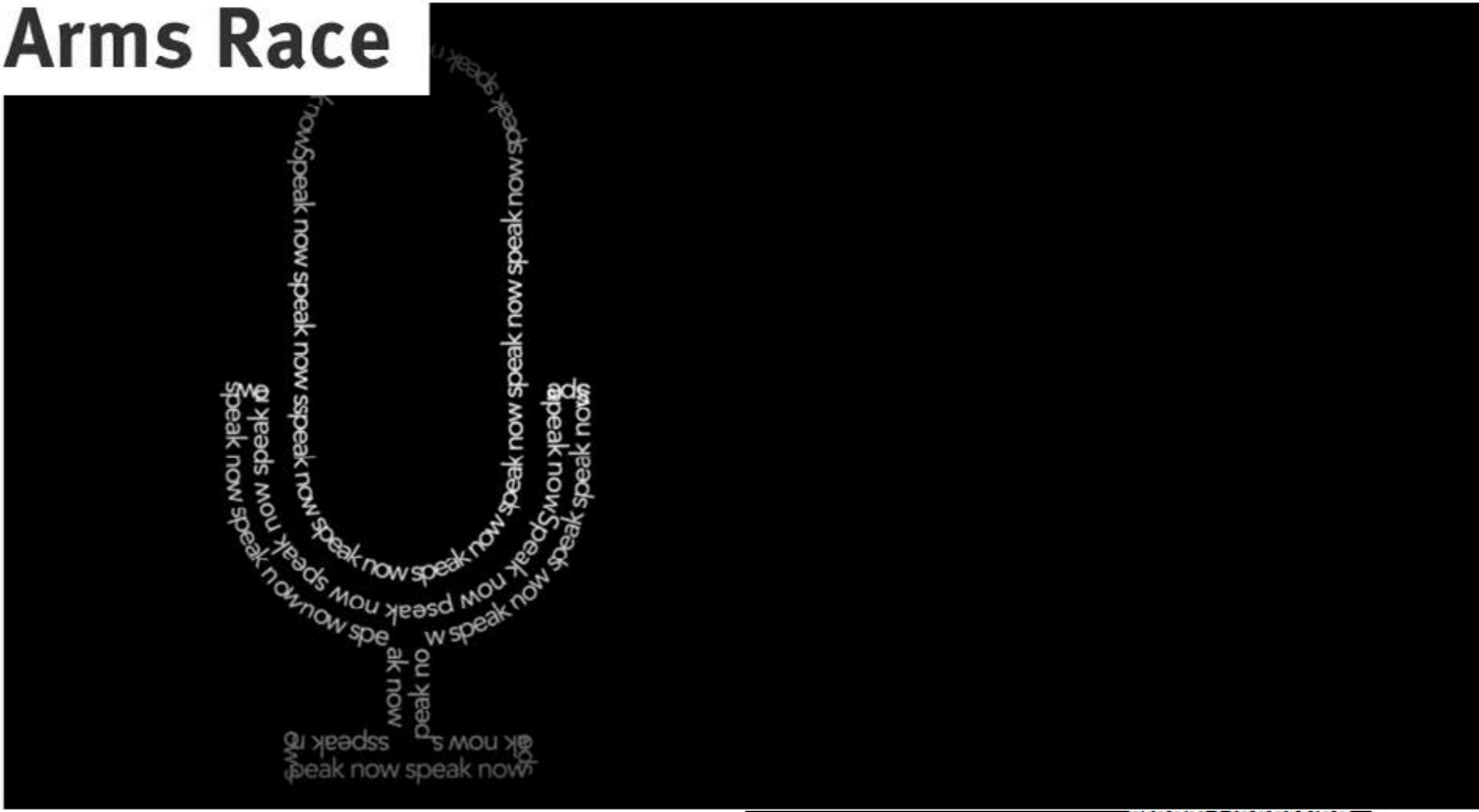
1 Comment +1

A▼ A▲

SUMMARY: *Google confirmed the acquisition to Re/code, which reports that it will pay $400 million for the London-based company.*

EXCLUSIVE

Facebook, Google in 'Deep Learning' Arms Race



credit: DeepMind

WHERE IS DL USED?

Vision

- Recognition
- Segmentation
- Parsing
- Detection

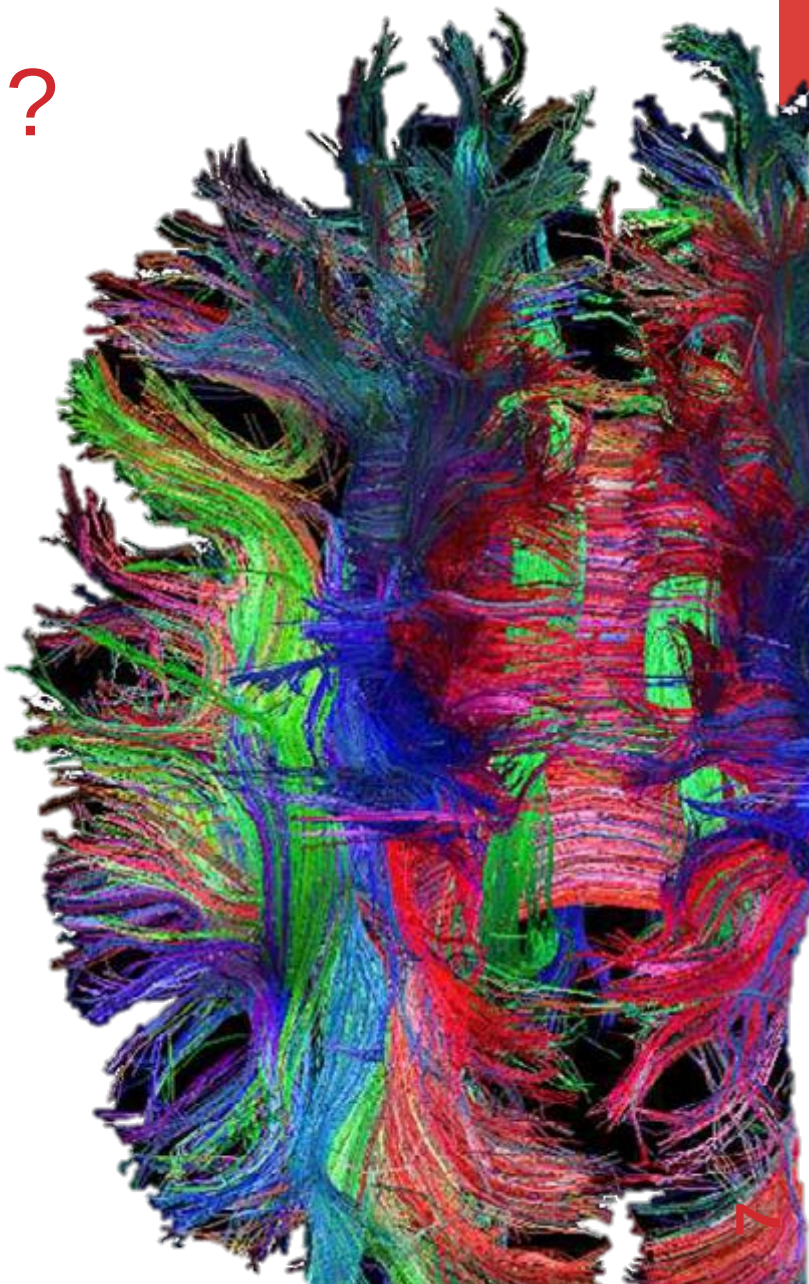
Metric learning

- Learning invariance
- Hashing

Speech

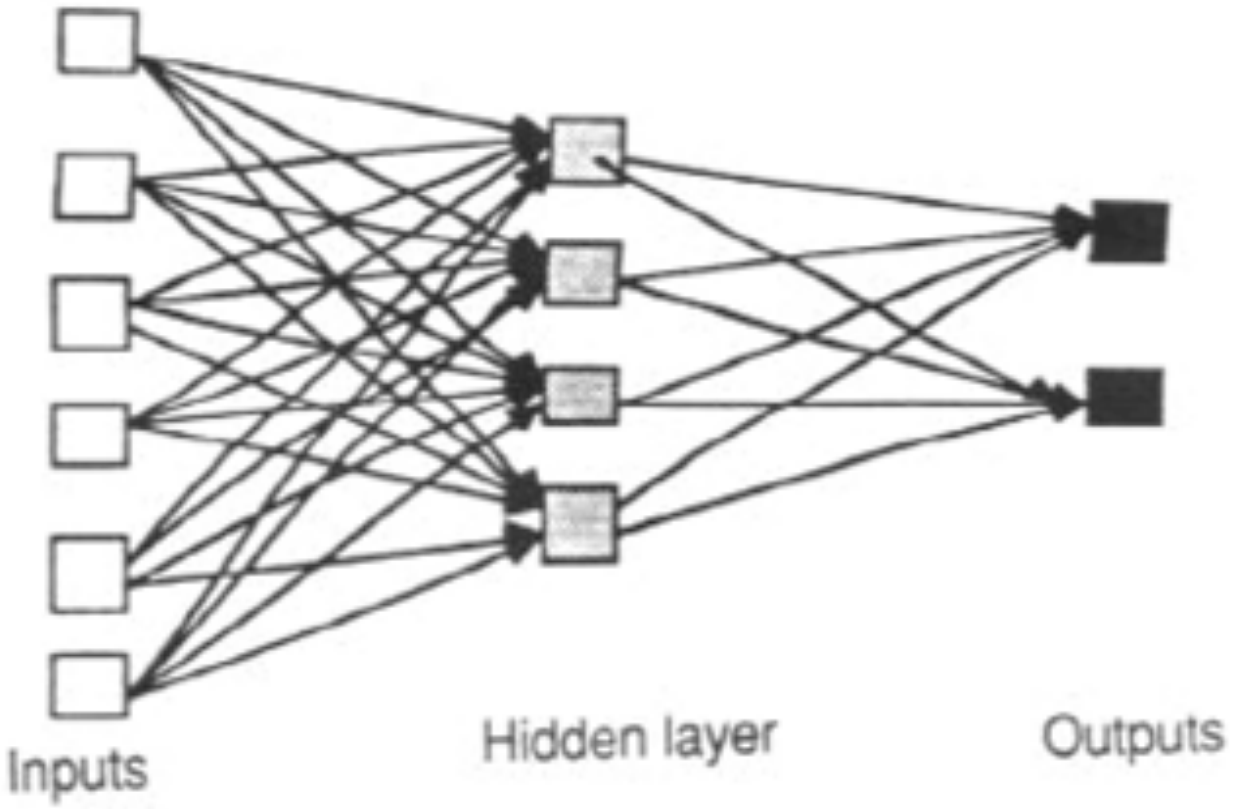
NLP

- Next breakthrough: it’s happening!
- Neural Machine Translation



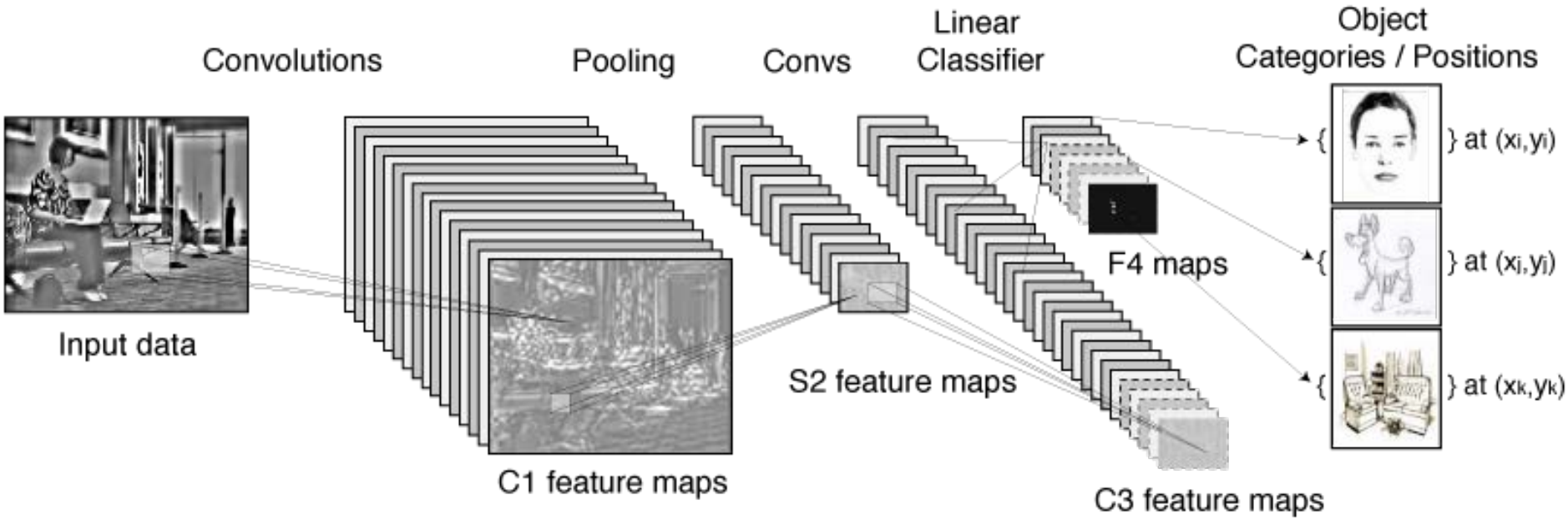
WHAT PEOPLE THINK IT IS?

WHAT REALLY IS...



Perceptron: Rosenblatt 1957 – MLP since 80s' Rumelhart, Hinton and Williams

WELL, ALMOST...

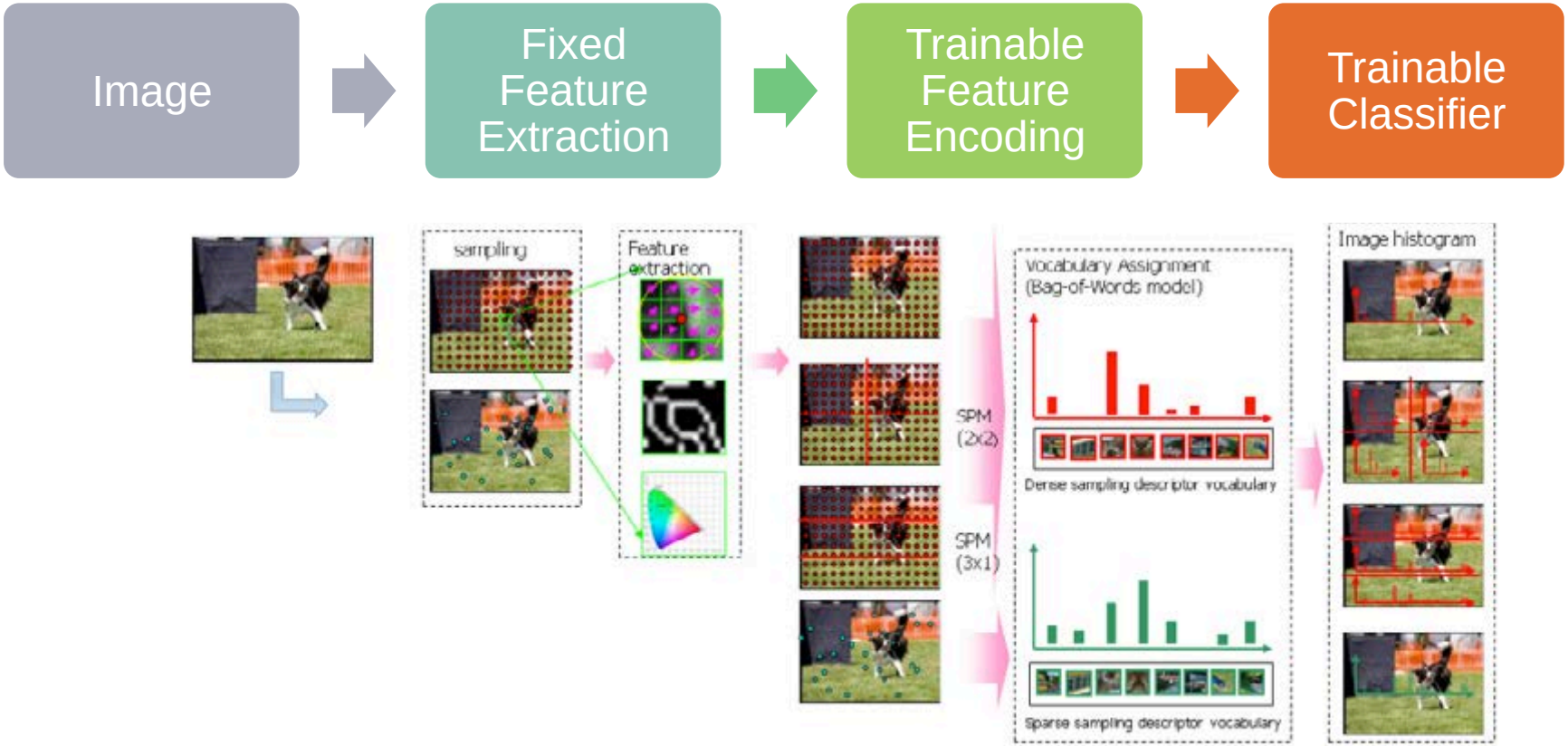


- **Quite few things have changed in the architectures**
- **Novel techniques allow for faster and better training**
- **But the underlying principles remain: learn highly non-linear functions from the data**

DL FOR VISION: THE REAL DEAL

- **Brought these techniques to popularity and to widespread usage in start-up and large companies**
  - Google, Facebook, Yahoo!, MSR, Amazon, you name it!
- **How to select the right representation?**
- **What are good criterion to craft a function to work with images? We all know local features work the best (i.e. SIFT, HOG, LBP, etc.)**
- **DL learns the features through a direct mapping from pixels to labels**

CLASSICAL CV RECOGNITION SYSTEM



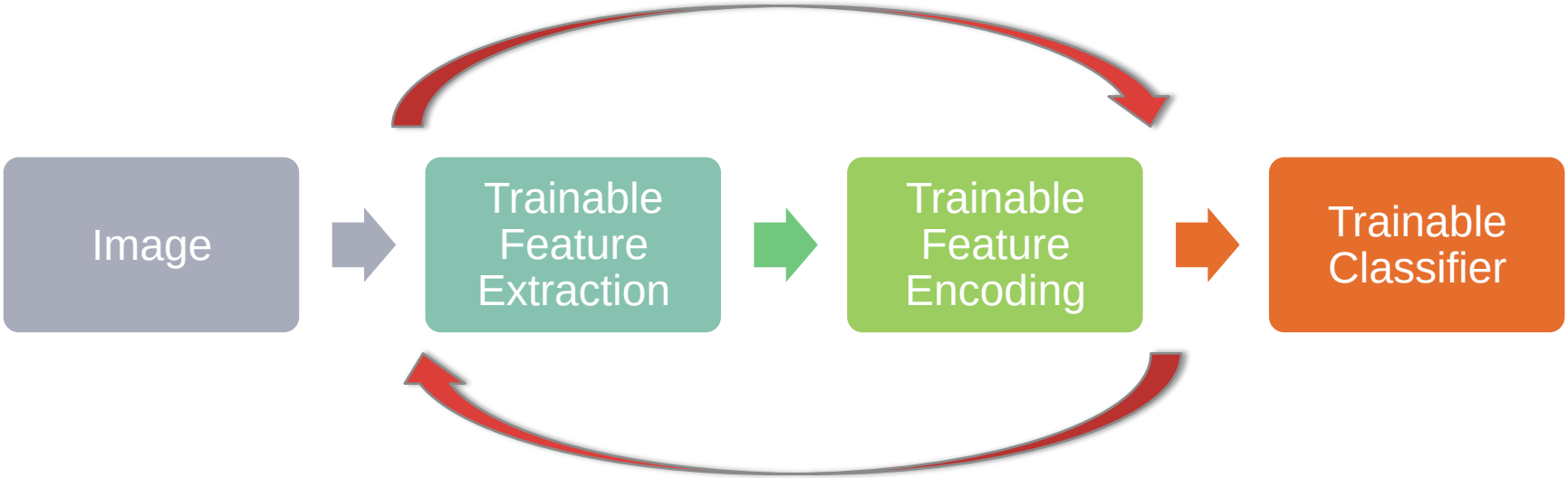
Yan & Huang (Winner of Pascal 2010 classification competition)

HAND CRAFTED FEATURES

- **It is not easy to select the best feature for the given task, and also features complement each-others:**
  - Use them all: MKL
- **MKL does give only little improvement though**
- **Better feature encoding algorithms (i.e. LLC)?**
  - Nice gain in performance
  - Hit a plateau, no much to improve
- **Where can we get a substantial gain?**
- **FEATURE EXTRACTION**
  - But how to design better features?

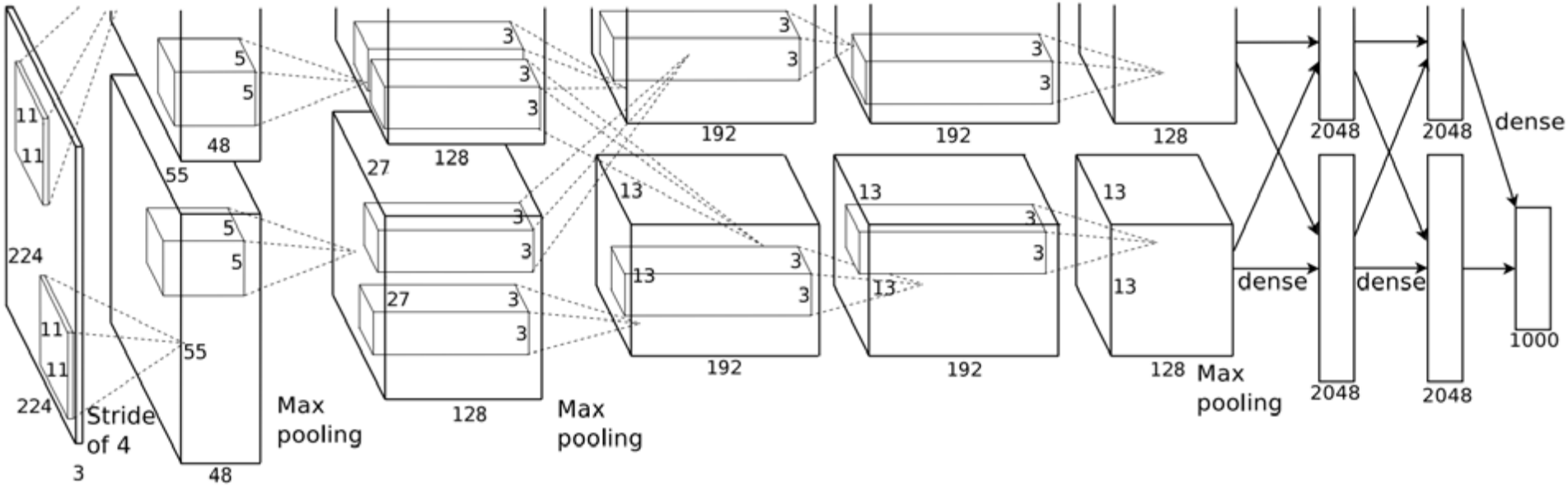
Wang, Yang, Yu, Lv, Huang, Gong [CVPR 2010]

DL APPROACH



- **Multiple stages jointly trained for the same objective**
- **The objective is not directly linked with the model:**
  - Classification, regression, detection, reconstruction etc.

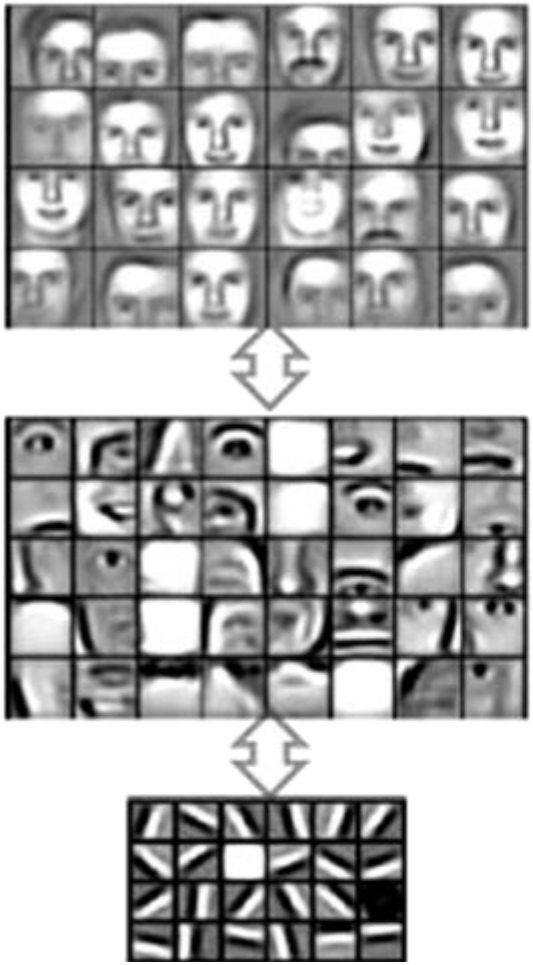
THE HAMMER OF IMAGE RECOGNITION



- **CNN: Hubel & Wiesel 1962, Fukushima 1979, LeCun et al. 1989, Riesenhuber & Poggio 1999, Ciresan et al. 2011**
- **Alternate convolution and max-pooling layers until the fully-connected classifier**
- **Hierarchical feature extraction: low-, mid- and high level features**

WHAT IT LEARNS

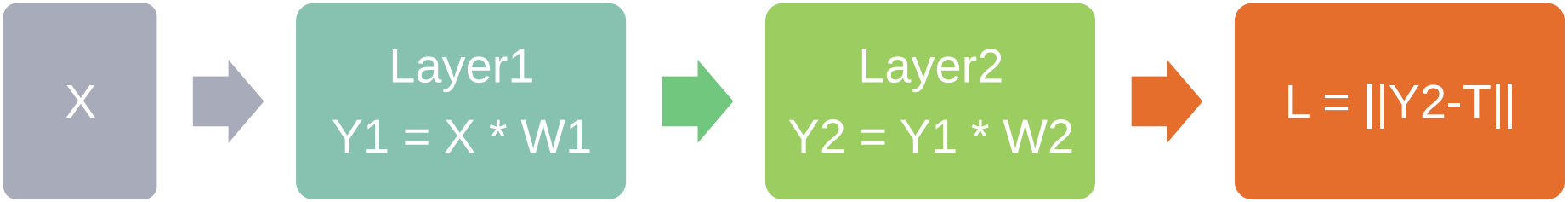
- **Features at higher stages encode abstract representations combining lower layers encodings:**
  - Low features combined into mid- and finally into high level semantic parts



Right-image: Lee, Grosse, Ranganath, Ng. [ICML 2009]

BACK-PROPAGATION

- **Algorithm to compute the gradient of deep models, and any model actually**
- **It is just the application of the chain rule of derivatives**



- **The gradient of each layer's parameters is easily computed:**
  - Define a function to compute the derivative of the output w.r.t. its input (back-propagation step)
  - Define a function to compute the derivative of the weights w.r.t. the loss (the gradient). Easily written in terms of result of back-propagation.

Linnainmaa 1970, Werbos 1974, LeCun 1985, Rumelhart 1986, Werbos 1989
Comprehensive list of contributions for back-propagation: Juergen 2014 Survey on DL

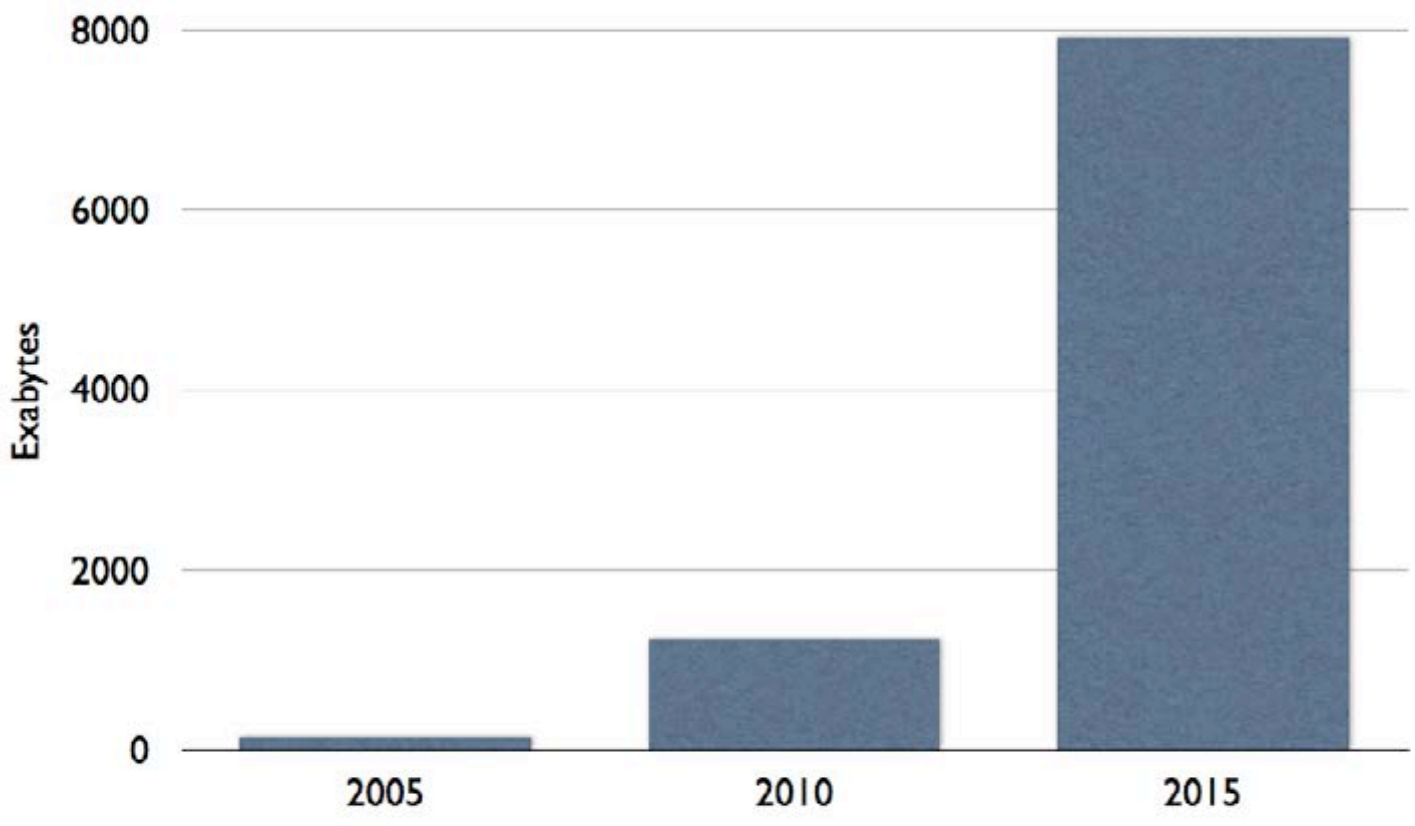
WHAT IS MAKING THE DIFFERENCE?

- **CNNs are more than two decades old, so why are they now working so well? Any substantial change in the architecture?**
  - Big Data
  - Dropout
  - Competition-based activation functions: LWTA, Maxout
  - GPUs
  - The deeper the better

- **Rule of thumb? Add as many parameters as possible for your hardware and train the hell out of it with proper regularization! [cit. Yann]**

DATA

Worldwide Data Growth

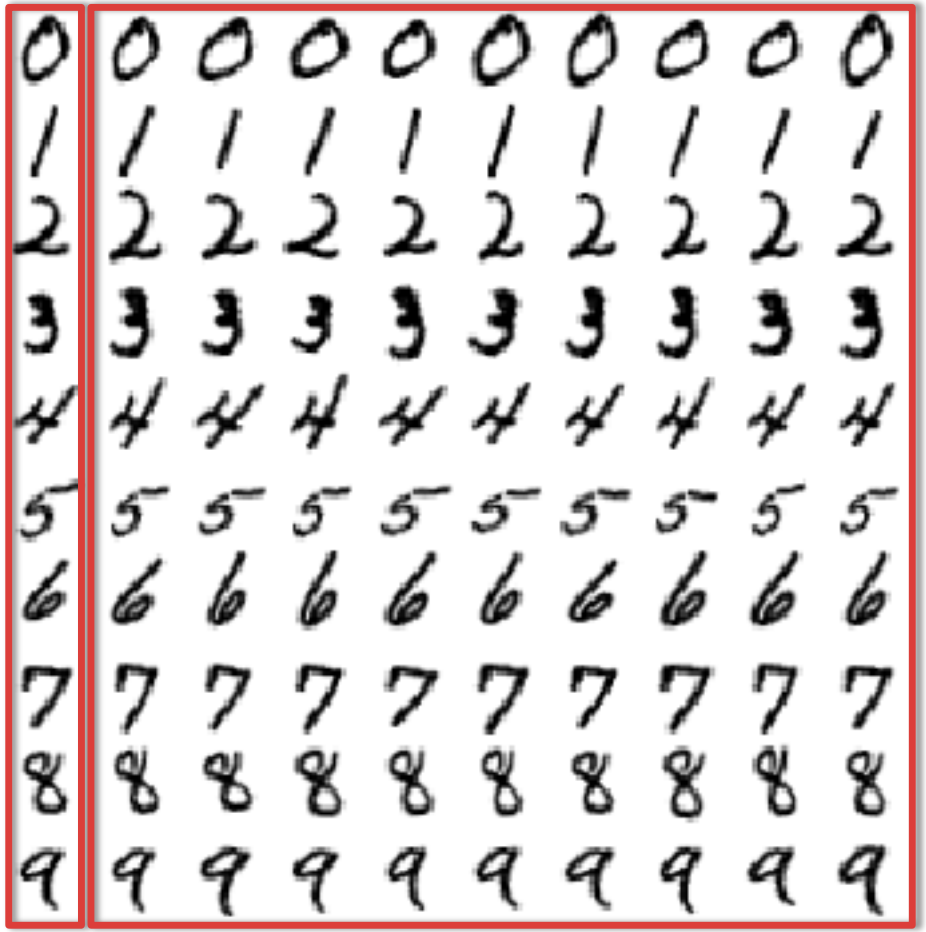


Source: IDC, EMC. 1EB = 1 Billion GB.

BLOGS.FORBES.COM/DAVEFEINLEIB

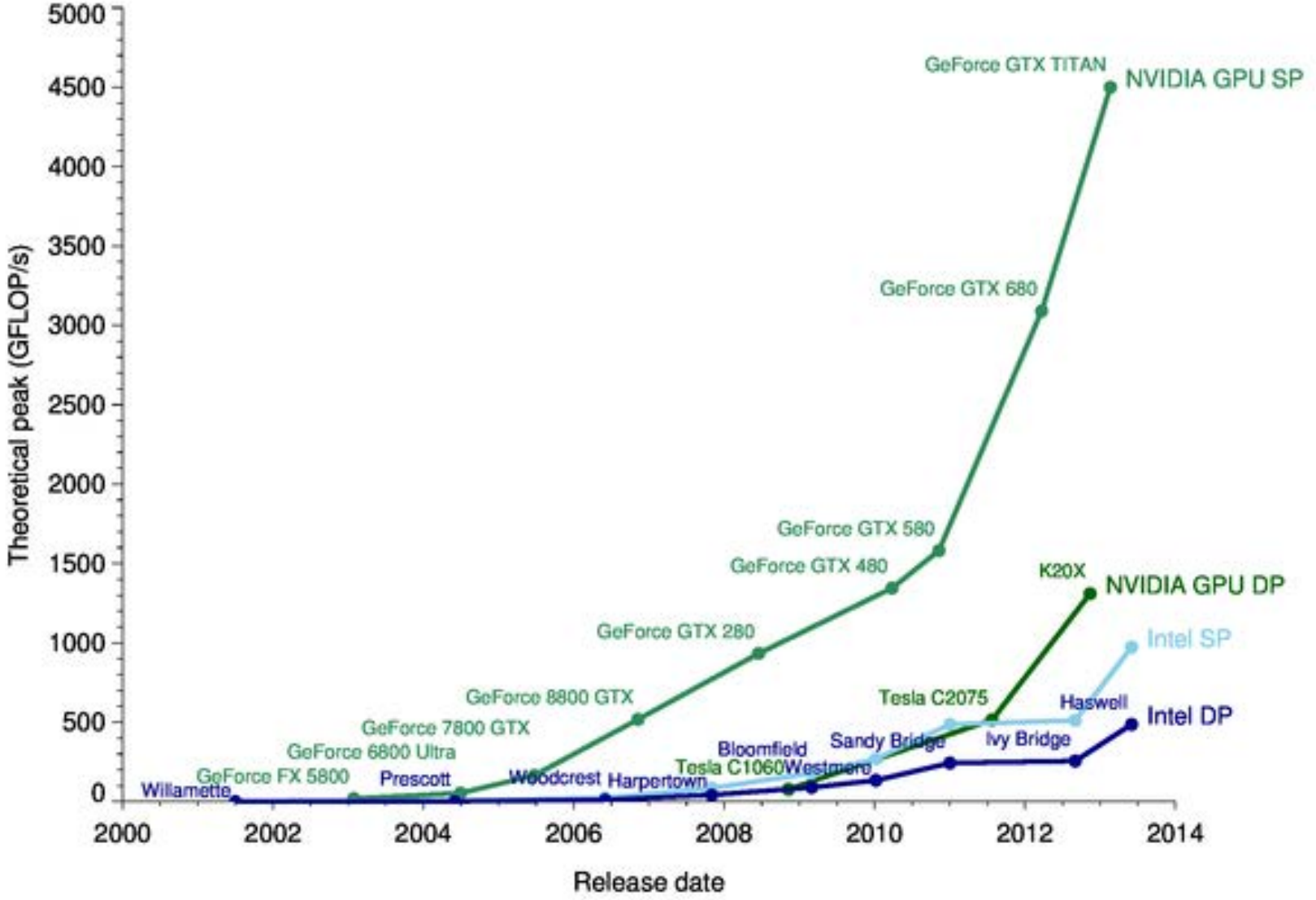
DATA AUGMENTATION

- **When data is scarce we can synthetically generate training samples. SOA for handwriting digit recognition**
- **Invariance by construction**
- **Used by all modern deep learners**

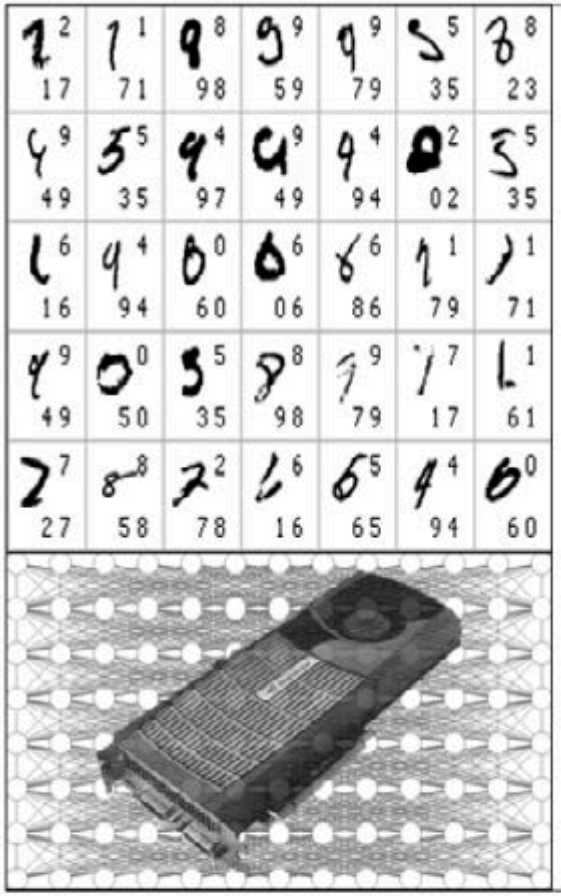


Seed Synthetic samples

COMPUTATIONAL POWER: GPU!



MAX POOLING: MPCNN DAS HAMMER!

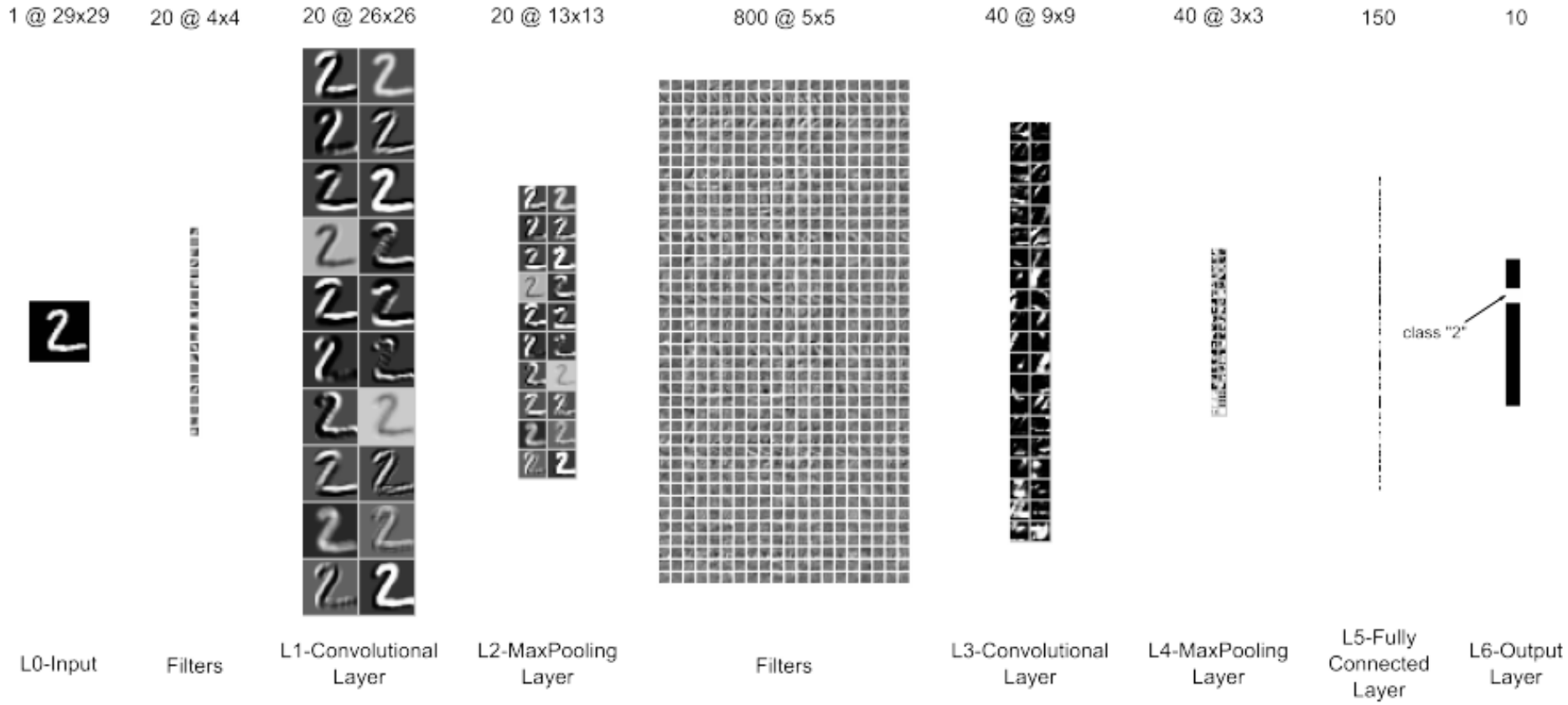


DEEP SPARSE CNN
+ MAX-POOLING
+ MLP ON TOP: 1
YEAR ON CPU = 1
WEEK ON GPU; $5*10^9$
WEIGHT UPDATES/s
2011-2012: FIRST
HUMAN-COMPETITIVE
MNIST RESULT: 0.2%
(after almost a decade of
roughly 0.4%)

CIRESAN, MEIER, MASCI, GAMBARDELLA, SCHMIDHUBER
(IJCAI 2011) - CIRESAN, MEIER, SCHMIDHUBER (CVPR 2012)

Slide: Juergen

FIRST DEEP MPCNN ON GPU



Same architecture used in almost all recognition competitions

Ciresan, Meier, Masci, Schmidhuber [IJCAI 2011]

MPCNN WINS CHINESE RECOGNITION CONTEST

ICDAR 2011 OFFLINE CHINESE HANDWRITING RECOGNITION CONTEST (4000 CLASSES):
1ST & 2ND RANK
OCT 2013: AGAIN BEST RESULTS,
NEAR-HUMAN PERFORMANCE

Slide: Juergen

TRAFFIC SIGNS COMPETITION

- **IJCNN 2011 on-site traffic sign recognition competition (Aug. 2nd 2011):**
  - 1st (0.56% error)
  - 2nd humans (1.16%)
  - 3rd (1.69%)
  - 4th (3.86%)

- **First super-human visual pattern recognizer**

Ciresan, Meier, Masci, Schmidhuber [IJCNN 2011, NN 2012]

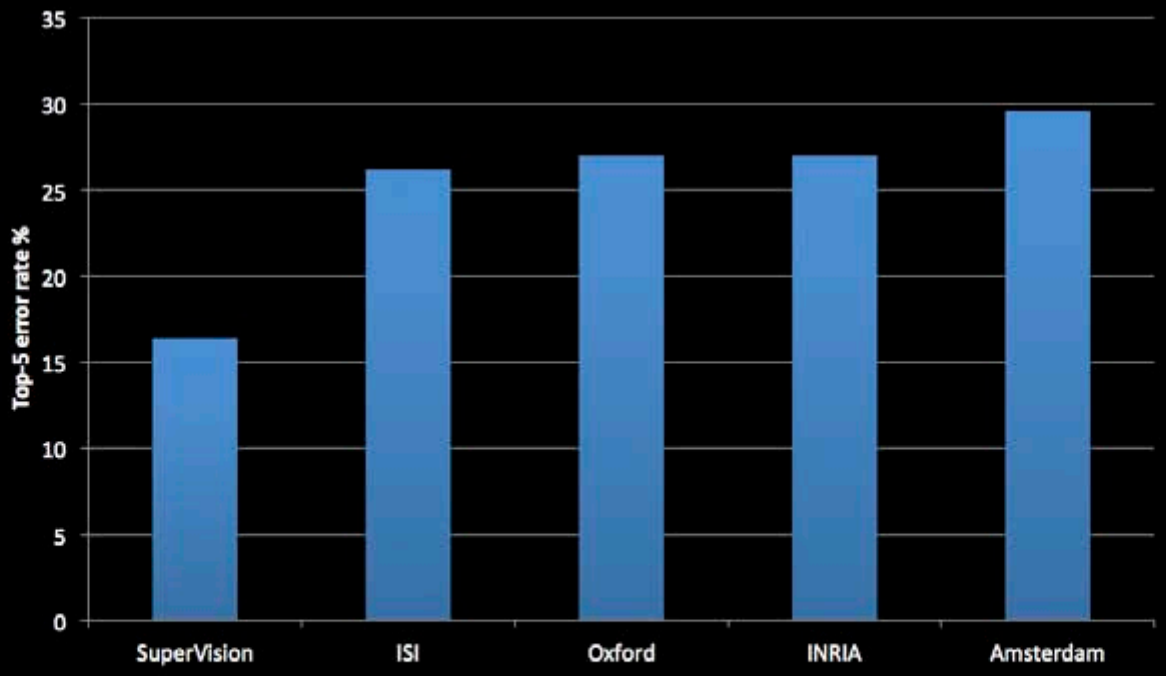
DROPOUT

- **Implicitly trains an ensemble of models**
- **Given h (input activations to the layer) and a dropping probability p there will be $2^{(n*(1-p))}$ possible sub-networks which can be randomly selected for training**
- **Train:**
  - Sample a binomial mask M and modify the usual forward pass of a layer with: $fwd_{dropout\_train}(x) = M \odot fwd(x)$
- **Test:**
  - In theory we should sample an exponential number of models, |M|, and average their predictions. In practice this can be approximated by: $fwd_{dropout\_test}(x) = fwd(x * (1-p))$

Krizhevsky, Sutskever, Hinton [NIPS 2012]
Hinton, Srivastava, Krizhevsky, Sutskever, Salakhutdinov [arXiv 2012]

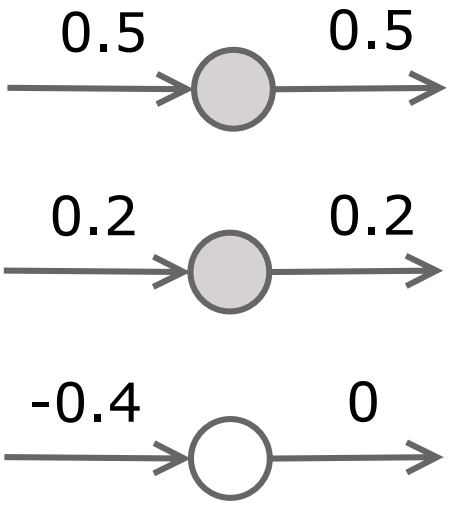
IMAGE-NET CLASSIFICATION: 2012

- Krizhevsky et al. -- 16.4% error (top-5)
- Next best (non-convnet) – 26.2% error

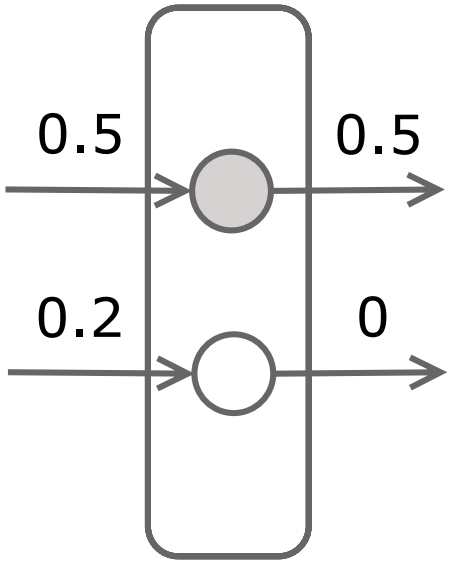


Slide: Rob Fergus

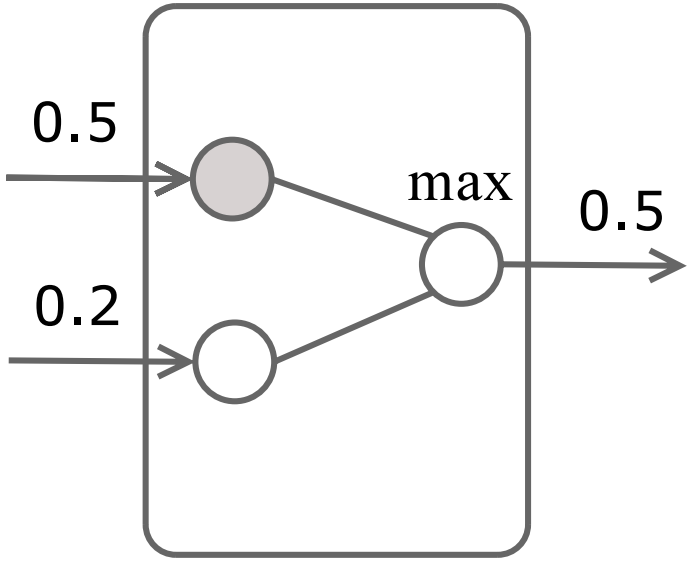
WINNER TAKE ALL UNITS



3 ReLUs



An LWTA block



A Maxout unit

Srivastava, Masci, Kazerounian, Gomez, Schmidhuber [NIPS2013]

NEURONAL MEMBRANE SEGMENTATION

First feedforward DL to win Image Segmentation Competition [ISBI 2012]

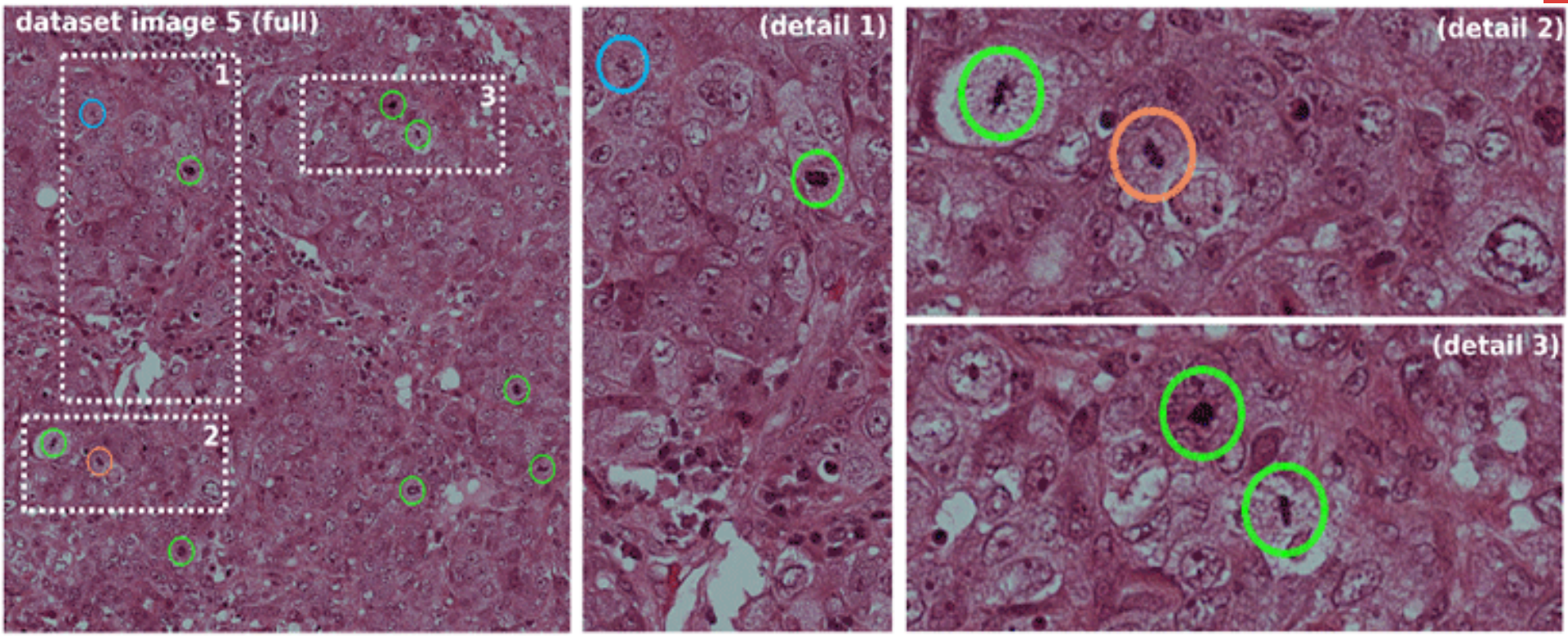
Masci, Giusti, Ciresan, Fricout, Schmidhuber [ICIP 2013]
Ciresan, Giusti, Gambardella, Schmidhuber [NIPS 2012]

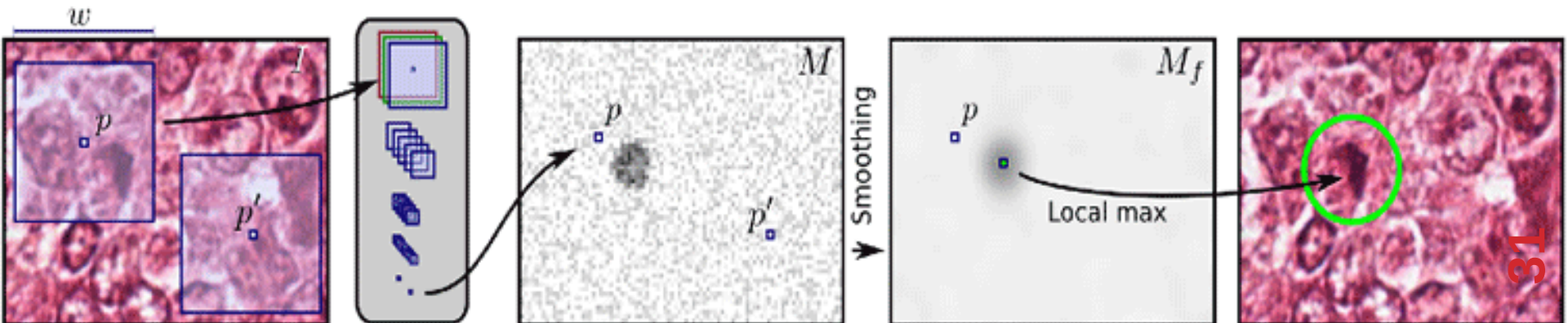
MEDICAL IMAGE DETECTION

First feedforward DL to win a contest in Object Detection (in large images) [ICPR2012]

Giusti, Caccia, Ciresan, Schmdhuber, Gambardella [ISBI 2014]
Ciresan, Giusti, Gambardella, Schmidhuber [MICCAI 2013]

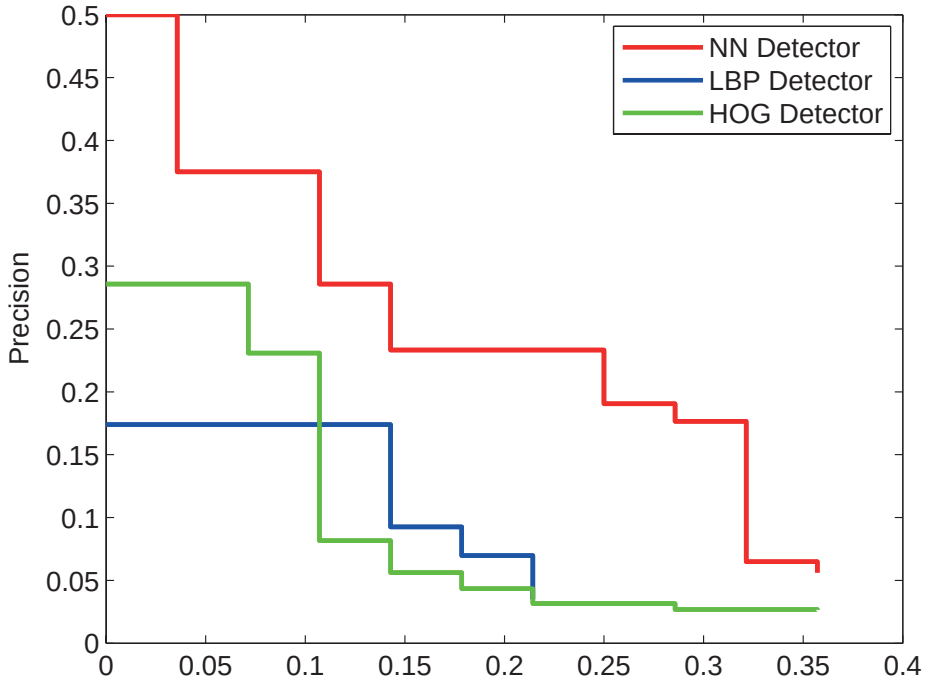
MODEL: YET ANOTHER MPCNN





STEEL DEFECTS

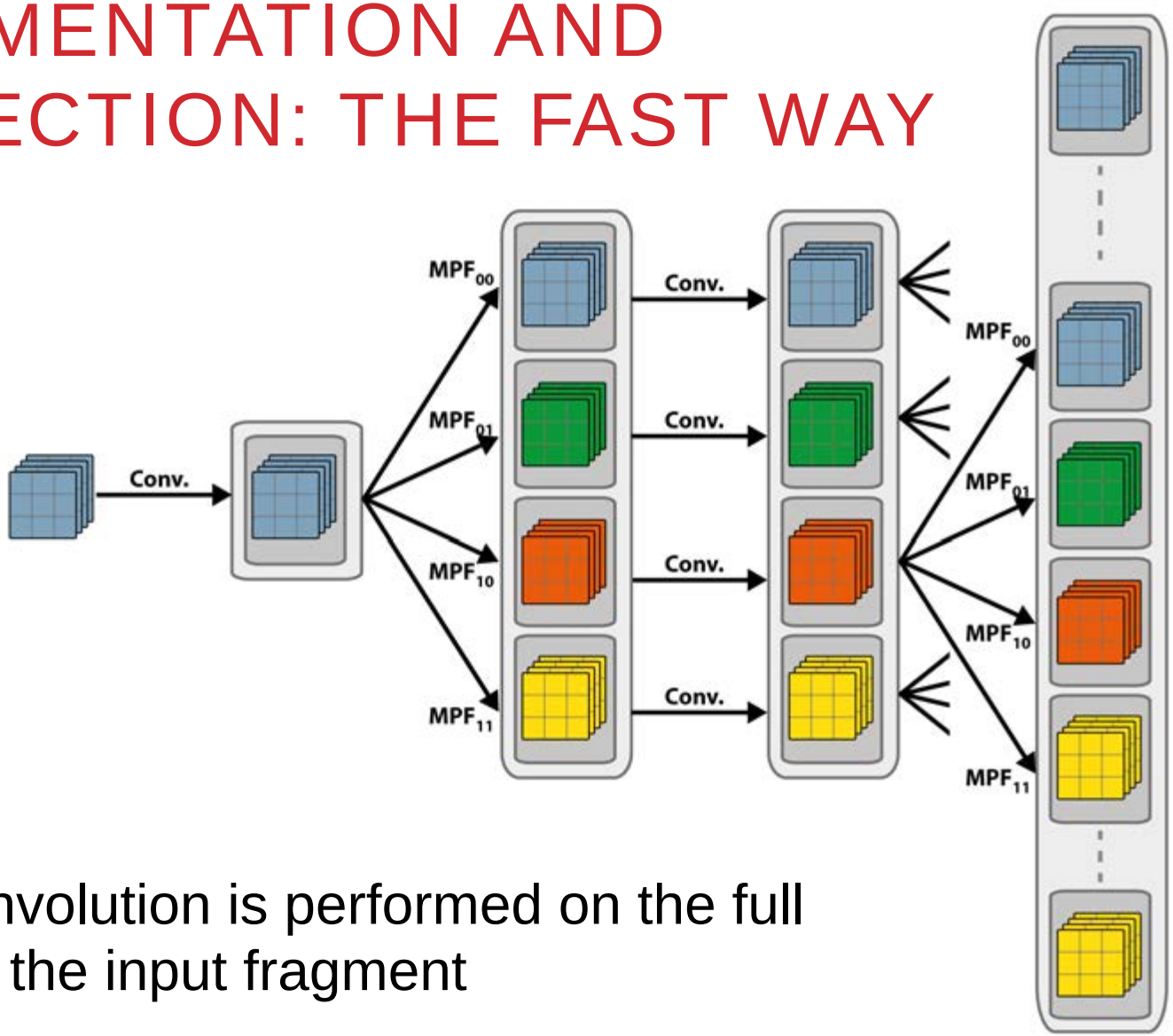
In collaboration with
ArcelorMittal Research



Input Target MPCNN

Masci, Giusti, Ciresan, Schmidhuber [ICIP 2013]
Giusti, Ciresan, Masci, Gambardella, Schmidhuber [ICIP 2013]

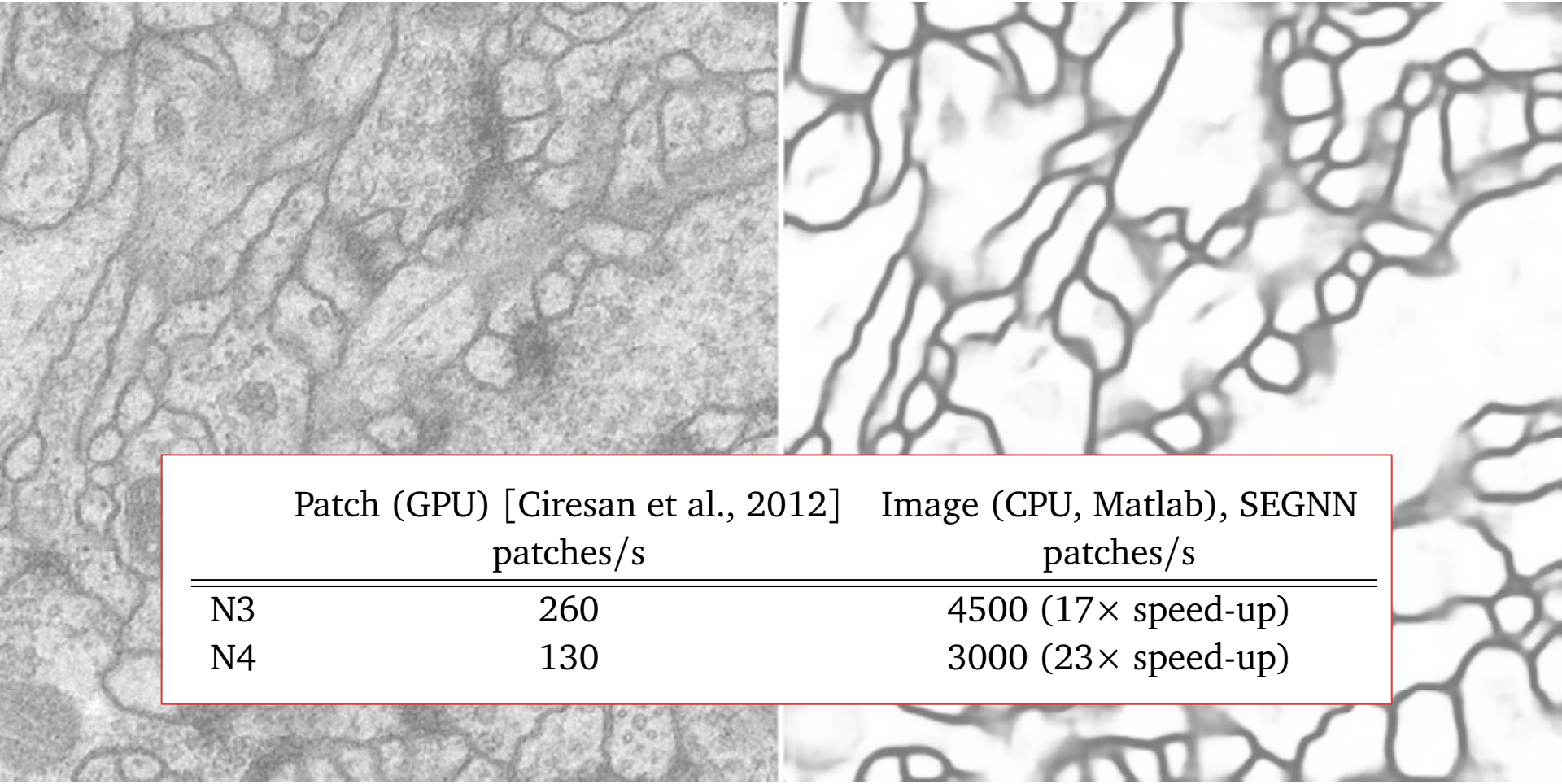
SEGMENTATION AND DETECTION: THE FAST WAY



Each convolution is performed on the full image in the input fragment

Masci, Giusti, Ciresan, Schmidhuber [ICIP 2013]
Giusti, Ciresan, Masci, Gambardella, Schmidhuber [ICIP 2013]

SEGMENTATION AND DETECTION: THE FAST WAY



	Patch (GPU) [Ciresan et al., 2012] patches/s	Image (CPU, Matlab), SEGNN patches/s
N3	260	4500 (17× speed-up)
N4	130	3000 (23× speed-up)

Masci, Giusti, Ciresan, Schmidhuber [ICIP 2013]
Giusti, Ciresan, Masci, Gambardella, Schmidhuber [ICIP 2013]

IMAGE-NET 2014 CLASSIFICATION AND LOCALIZATION

- **Almost all entries use a CNN as feature extractor**
- **As the net needs to be evaluated at different scales and at all position (of a given search window) variants of our approach are used**
- **What this year's competition taught us?**
  - Deep nets work better and use less parameters [Google]
  - Epitomes are a nice way to reduce the redundancy in large nets [Papandreou, Kokkinos]

MISERABLE FAILURES

(a) (b)

Figure 5: Adversarial examples generated for AlexNet [9].(Left) is a correctly predicted sample, (center) difference between correct image, and image predicted incorrectly magnified by 10x (values shifted by 128 and clamped), (right) adversarial example. All images in the right column are predicted to be an *"ostrich, Struthio camelus"*. Average distortion based on 64 examples is 0.006508. Plase refer to `http://goo.gl/huaGPb` for full resolution images. The examples are strictly randomly chosen. There is not any postselection involved.

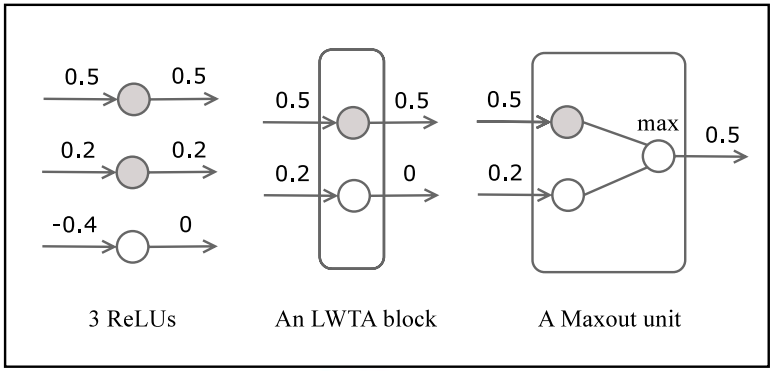
Christian Szegedy et al. [2014]

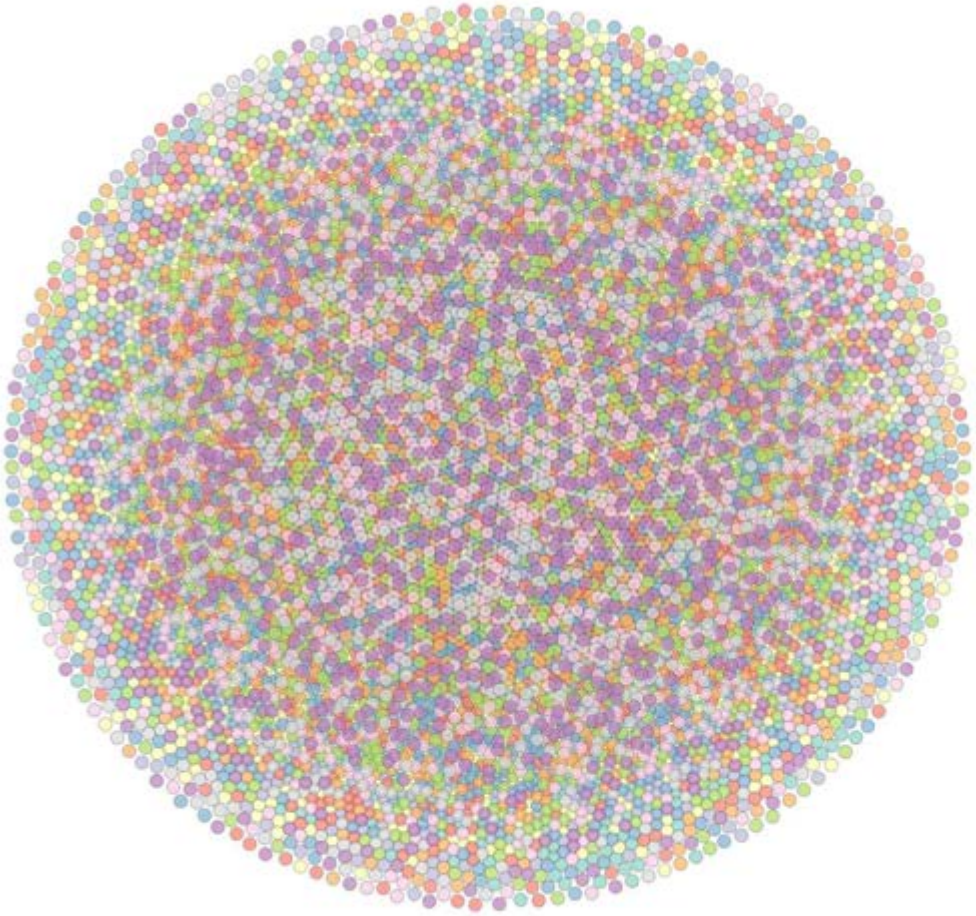
FUTURE DIRECTIONS

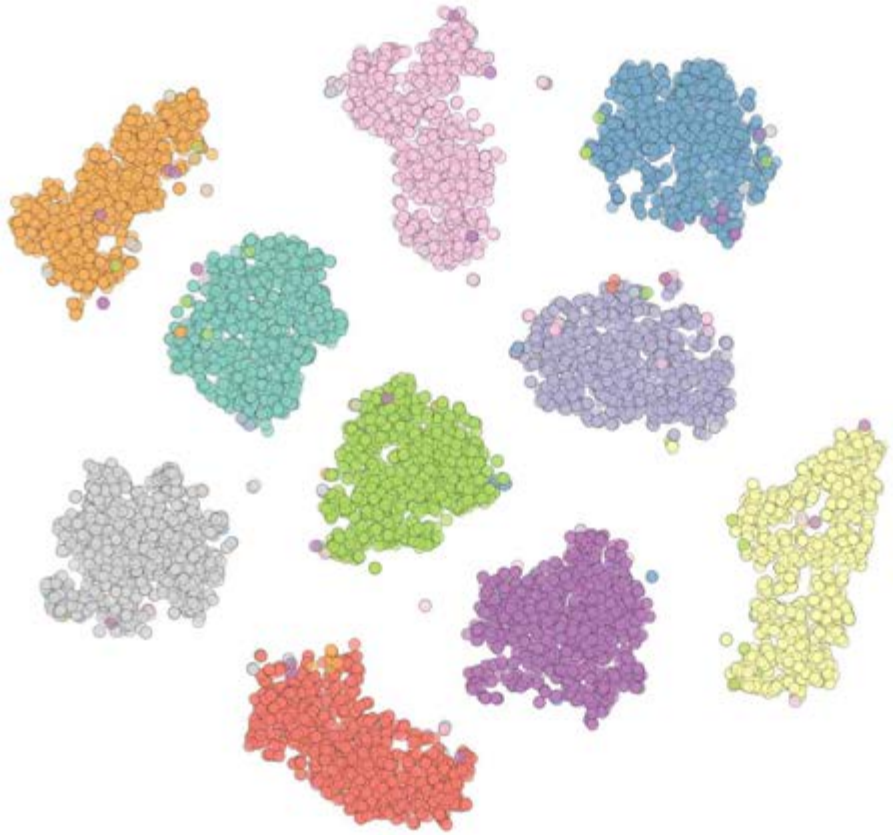
- **Video is relatively under-explored:**
  - Lack of large annotated datasets
  - Different tasks, it is not just classification or detection
  - Recurrent nets are paramount in these applications
- **Non-Euclidean**
  - How to work on non-Euclidean domains such as graphs or shapes? Spectral networks in other words.
- **Multimodal:**
  - Combine several representations in a joint space
  - Generate text from images: Few applications are coming out, stay tuned for next CVPR

NIPS PREVIEW

UNDERSTANDING LOCAL COMPETITION





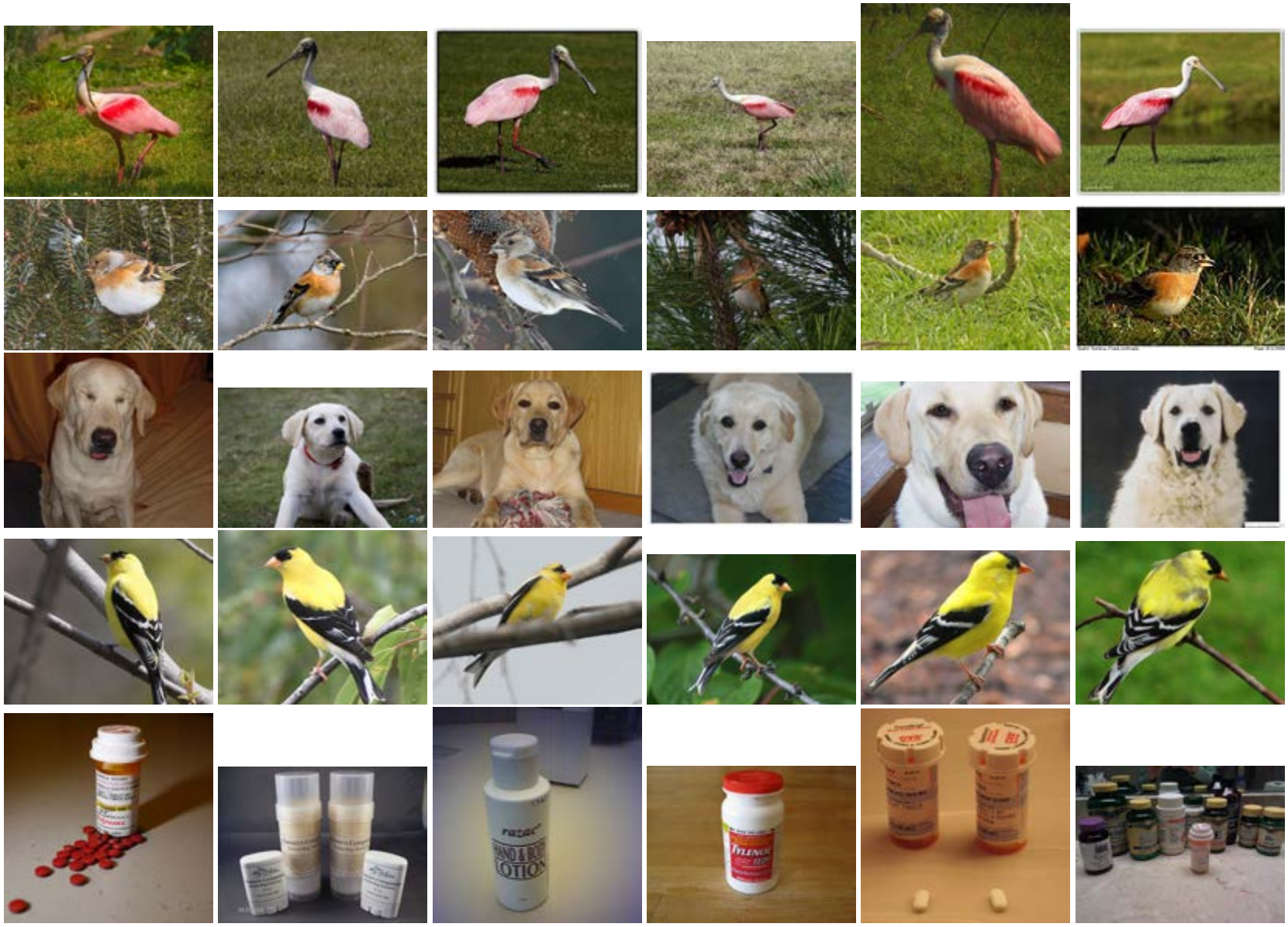


CIFAR10 LWTA UNTRAINED

CIFAR10 LWTA TRAINED

Srivastava R. K., Masci J., Gomez F., Schmidhuber J. [2014]
Srivastava R. K., Masci J., Kazerounian S., Gomez F., Schmidhuber J, [2013]

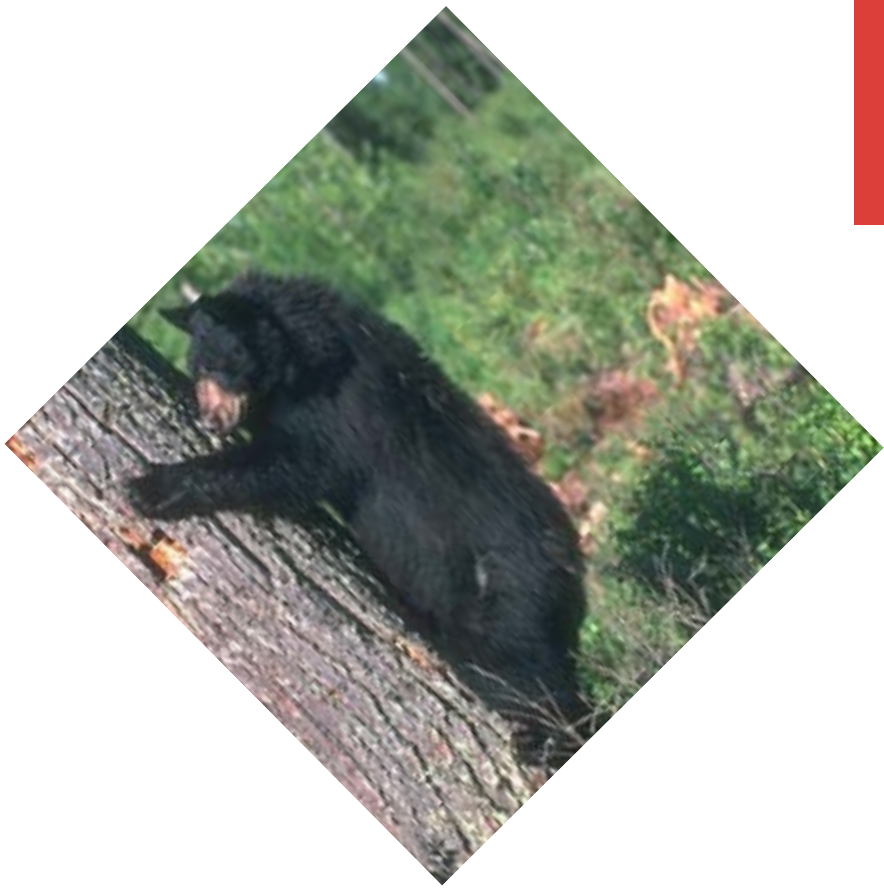
RETRIEVAL WITH SUB-NETWORK IDENTIFIER



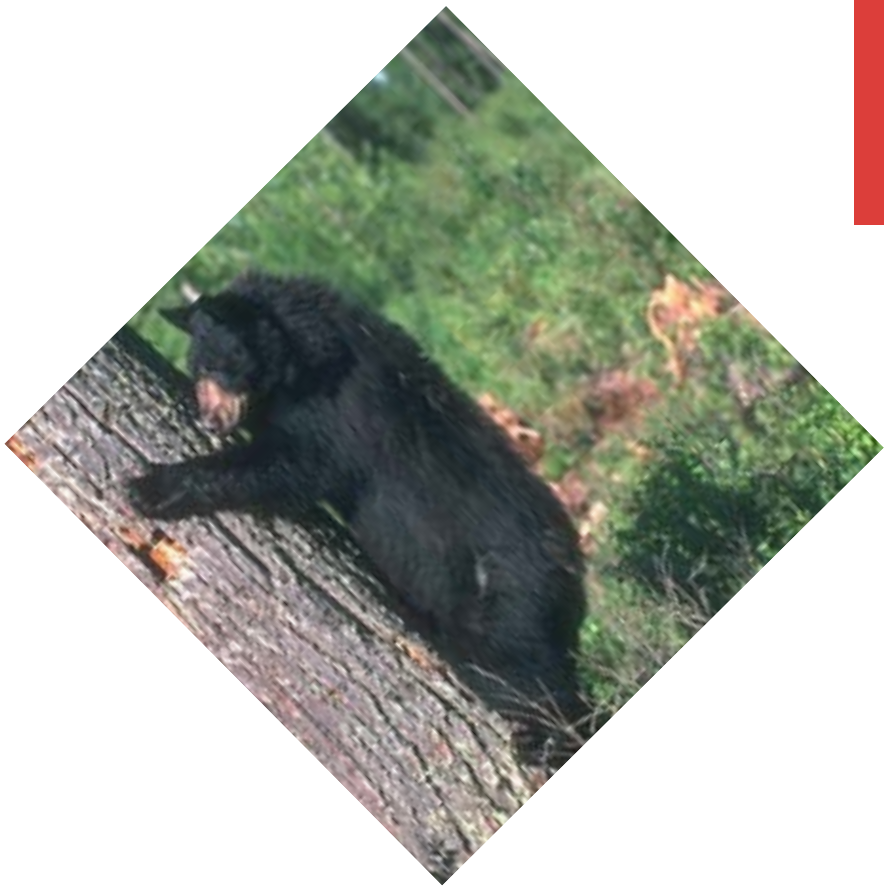
EXPERIMENT

EXPERIMENT

EXPERIMENT



EXPERIMENT



EXPERIMENT

EXPERIMENT

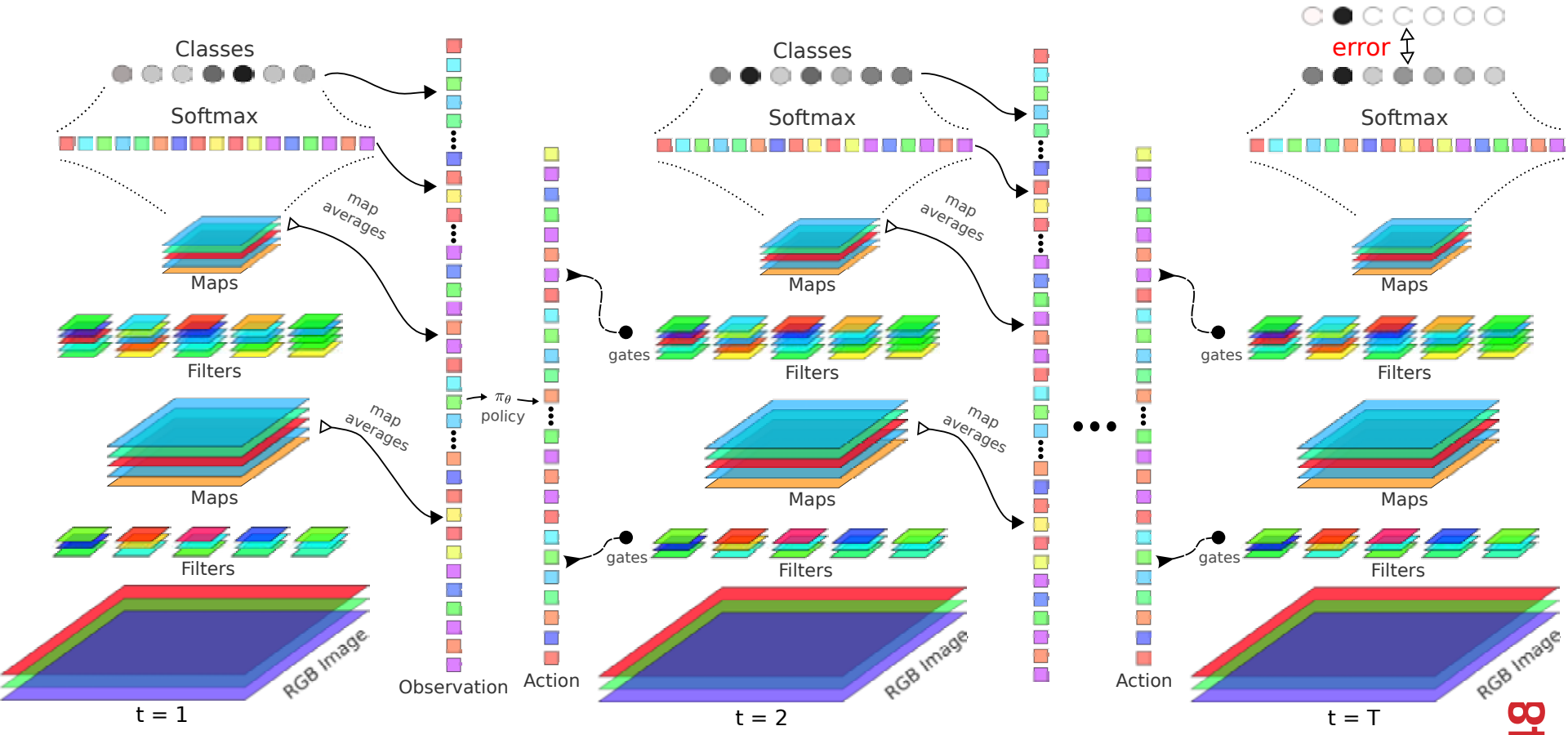
HOW DO WE DO RECOGNITION?

- **Response time is proportional to the number of objects in the visual scene**
- **We have to search in the image where to look before deciding what it is**
- **Sometimes to assess what it is we need more samples:**
  - Fine grain classification is an example
- **We need feedback, top-down reasoning, to drive this “internal search light”!**

DASNET

Policy: $\pi_{\theta_i}(\mathbf{o}) = dim(A)\sigma(\boldsymbol{\theta}_i \mathbf{o_t}) = \mathbf{a_t},$

Gated fwd: $y_j^\ell = a_j^\ell \sum_{i=0}^{i=c} \phi(x_i * F_{i,j}^\ell)$

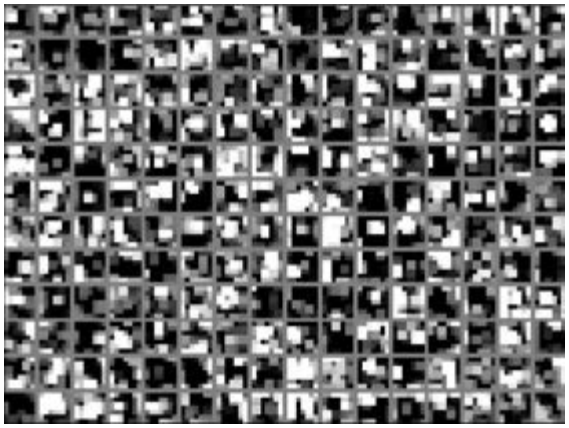


Stollenga M., Masci J., Gomez F., Schmidhuber J. [2014]

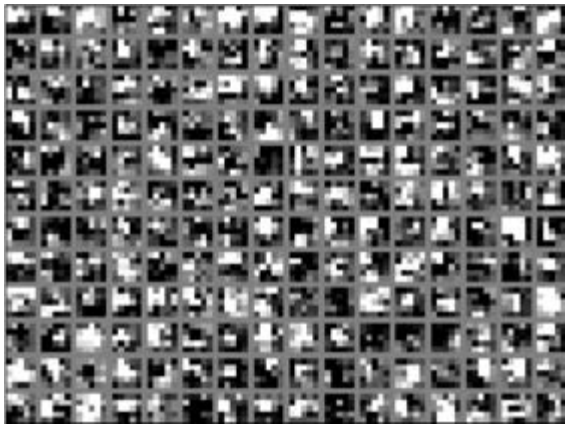
DASNET

Method	CIFAR-10	CIFAR-100
Dropconnect [3] (12 CNNs)	9.32%	-
Stochastic Pooling [46]	15.13%	-
Multi-column CNN [2]	11.21%	-
Maxout [4]	9.38%	38.57%
Maxout (our model)	9.61%	34.54%
dasNet	**9.22%**	**33.78%**

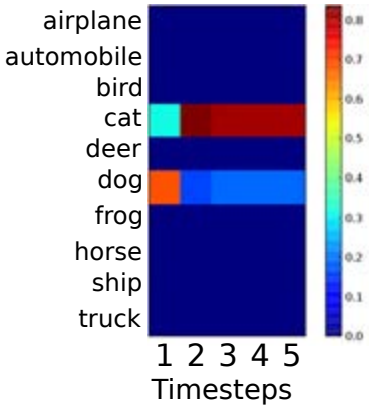
layer 0



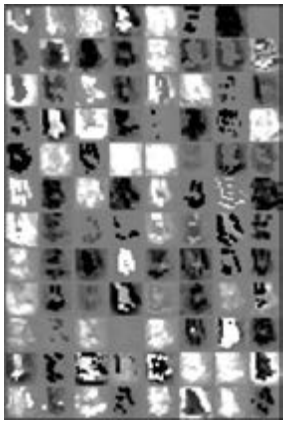
layer 1



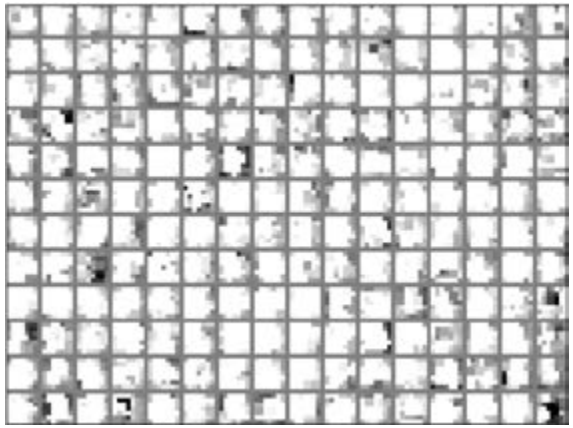
layer 2



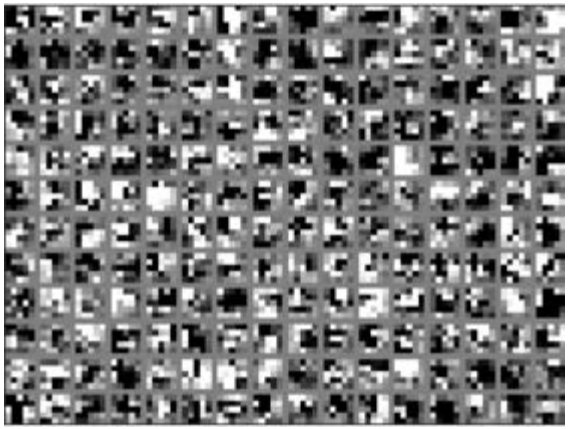
class probabilities



change of layer 0



change of layer 1



change of layer 2

Stollenga M., Masci J., Gomez F., Schmidhuber J. [NIPS 2014]

RNN

- **Recurrent Nets are in principle Turing complete**
  - Program = Weights
- **Hard to train because of the vanishing gradient problem**
  - Long Short Term Memory RNNs (LSTMs) do not suffer such problems and are now paramount for challenging applications
- **But this model is indeed as old as CNNs:**
  - Again, thanks to few additional tricks, lots of data and computational power everybody is now using them
- **Juergen was right!**

BREAKTHROUGH

- **From Juergen Schmidhuber's G+ post:**

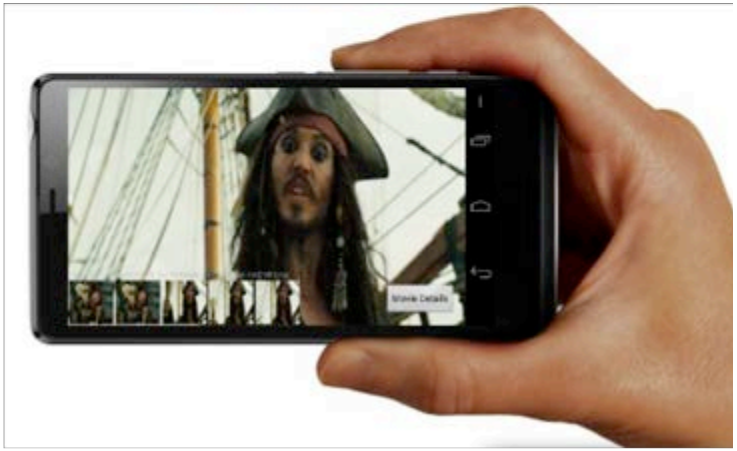
Recent (2014) benchmark records in speech recognition and machine translation achieved with the help of deep Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs), often at major IT companies:

- *Large vocabulary speech recognition (Sak et al., Google, Interspeech 2014)*
- *English to French translation (Sutskever et al., Google, NIPS 2014)*
- *Text-to-speech synthesis (Fan et al., Microsoft, Interspeech 2014)*
- *Prosody contour prediction (Fernandez et al., IBM, Interspeech 2014)*
- *Language identification (Gonzalez-Dominguez et al., Google, Interspeech 2014)*
- *Medium vocabulary speech recognition (Geiger et al., Interspeech 2014)*
- *Audio onset detection (Marchi et al., ICASSP 2014)*
- *Social signal classification (Brueckner & Schulter, ICASSP 2014)*
- *Arabic handwriting recognition (Bluche et al., DAS 2014)*
- *TIMIT phoneme recognition (Graves et al., ICASSP 2013)*
- *Optical character recognition (Breuel et al., ICDAR 2013)*

Deep Learning in Neural Networks: An Overview [http://www.idsia.ch/~juergen/deep-learning-overview.html]
Original papers on LSTM and its various topologies and learning algorithms since 1995: http://www.idsia.ch/~juergen/rnn.html

VNOME

- **Very large-scale video identification and retrieval from a short sequence**
- **Motivation:**
  - Watching a movie in another language we would like to add subtitle in our language
  - Independent video producers want to keep track of their videos
  - Directors need short clip with a given semantic while editing a longer sequence; i.e. advertising
- **Android application**

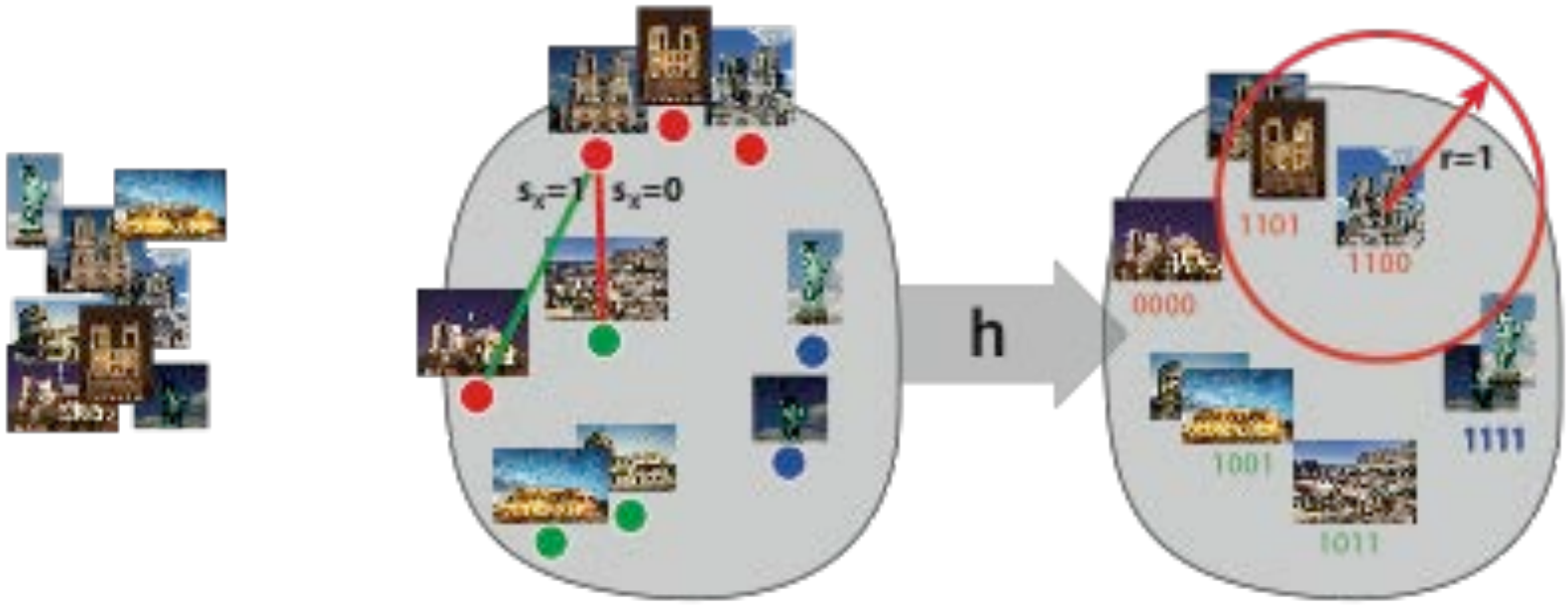


CHALLENGE

- **Index a large collection of videos N, each with approximately 50K frames on average**
  - O(N * d * 50K) where d is the frame representation in bits
  - Trillion frame scale 10^{12}
- **Representation for each frame?**
  - Pixels: 1280*1024*32 bits
  - BoW: minimum representations take 1024*32 bits and can go up to more than 100K*32 bits
  - …well you get it, we need something smarter
- **Search complexity:**
  - Even logarithmic search is not fast enough
  - With LUT search time is constant
- **Invariance to frame transformations:**
  - Affine, color, etc.

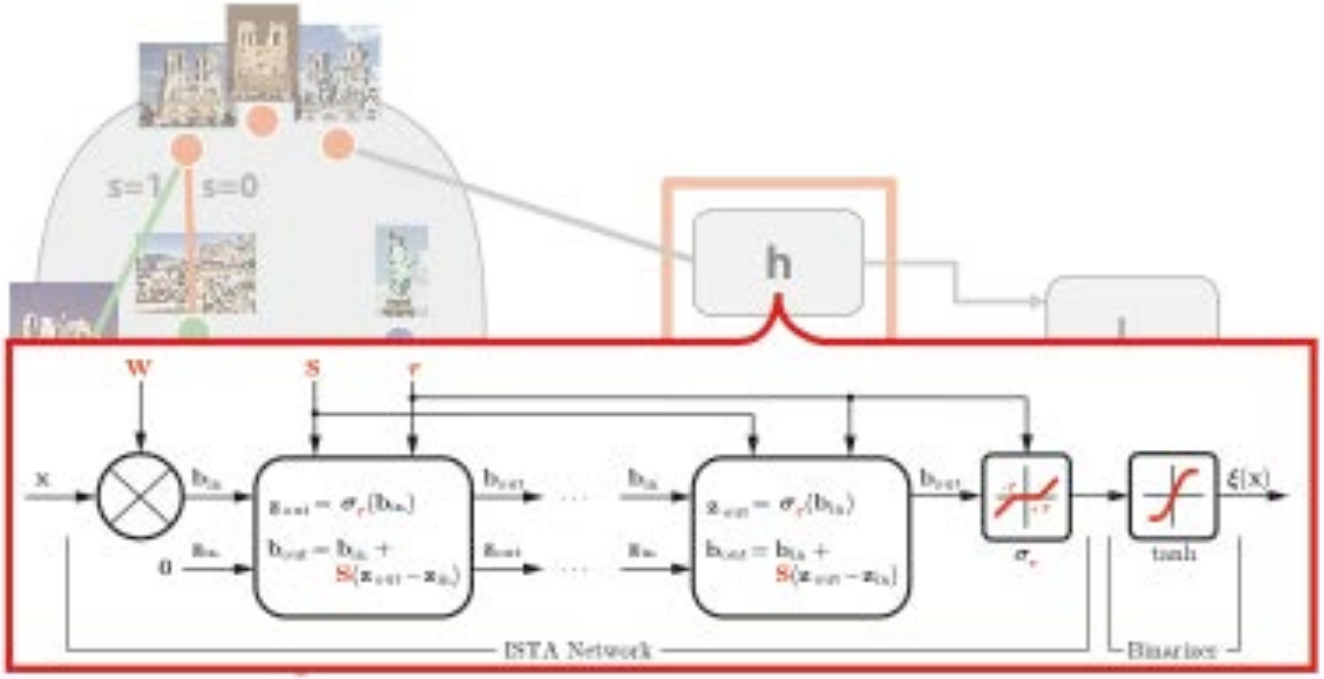
SPARSE-HASH

- **How do we store and handle such massive amount of information?**
  - Dimensionality reduction: Hashing (binary representation)
- **If we assume some known similarities we can train a discriminative deep learner h to map “images to bits”**



Masci, Bronstein, Bronstein, Schmidhuber [PAMI 2014]
Masci, Bronstein, Sprechmann, Bronstein, Sapiro [ICLR 2014]

SPARSE-HASH MODEL



$$\begin{aligned}\mathcal{L} &= s(\mathbf{x},\mathbf{x}')\|\mathbf{h}(\mathbf{x})-\mathbf{h}(\mathbf{x}')\|_1 \\ &+ \lambda(1-s(\mathbf{x},\mathbf{x}'))\max\{0, M-\|\mathbf{h}(\mathbf{x})-\mathbf{h}(\mathbf{x}')\|_1\} \\ &+ \alpha(\|\mathbf{h}(\mathbf{x})\|_1+\|\mathbf{h}(\mathbf{x}')\|_1)\end{aligned}$$

Masci, Bronstein, Bronstein, Schmidhuber [PAMI 2014]
Masci, Bronstein, Sprechmann, Bronstein, Sapiro [ICLR 2014]

BACK TO VNOME

- **Each frame is represented by 32 bit codes obtained with Sparse-Hash**
- **Data is indexed using Redis, a no-sql database and retrieval is done in a smart and classified way ☺**
- **Working with the camera is the big challenge…but we are about to solve it, hopefully before our demo at**
  - International BASP Frontiers in Villars-sur-Ollon
- **Stay tuned for future developments!**

CONCLUSION

- **DL is not the answer to all problems but a powerful tool to be used to tackle real challenging AI problems**
  - Industrial applications (Google, Facebook, Samsung…)
- **Eventually these modules tied together will allow an artificial agent to operate your kitchen devices and to learn from experience how to solve problems**
- **Data alone is not enough**
- **We need better models for general AI, make a smarter use of the data**
- **Understand the visual scene: a cup is not just a collection of edges. We need more context aware models that understand objects and their relations.**

ACKNOWLEDGEMENTS


Juergen Schmidhuber


Michael Bronstein


Guillermo Sapiro


Alex Bronstein


Faustino Gomez


Ueli Meier


Dan Ciresan


Alessandro Giusti


Pablo Sprechmann


Davide Eynard


Marijn Stollenga


Rupesh K. Srivastava

ICLR 2015

- **Submission deadline: December 19, 2014**
- **Location: Hilton San Diego Resort & Spa, May 7-9, 2015**
- **http://www.iclr.cc**
- **Reference conference for deep learning and representation learning**
- **Still new so still manageable to get to know everyone and discuss possible collaborations!**
- **OpenReview publishing process**