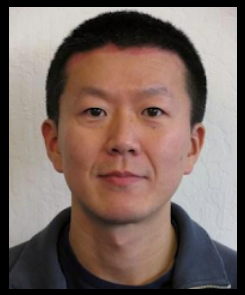
Tera-scale deep learning

Quoc V. Le
Stanford University and Google

Joint work with



Kai Chen

Greg Corrado

Jeff Dean

Matthieu Devin

Rajat Monga

Andrew Ng

Marc'Aurelio Ranzato

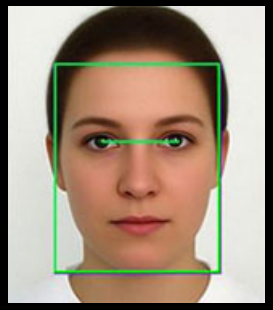
Paul Tucker

Ke Yang

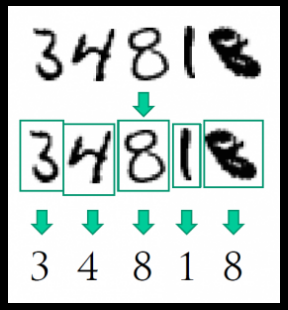
<u>Additional Thanks:</u>

Samy Bengio, Zhenghao Chen, Tom Dean, Pangwei Koh, Mark Mao, Jiquan Ngiam, Patrick Nguyen, Andrew Saxe, Mark Segal, Jon Shlens, Vincent Vanhouke, Xiaoyun Wu, Peng Xe, Serena Yeung, Will Zou

Machine Learning successes

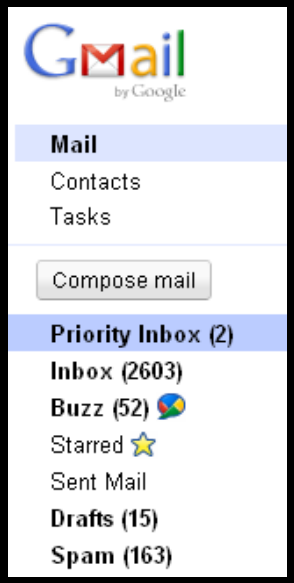


Face recognition



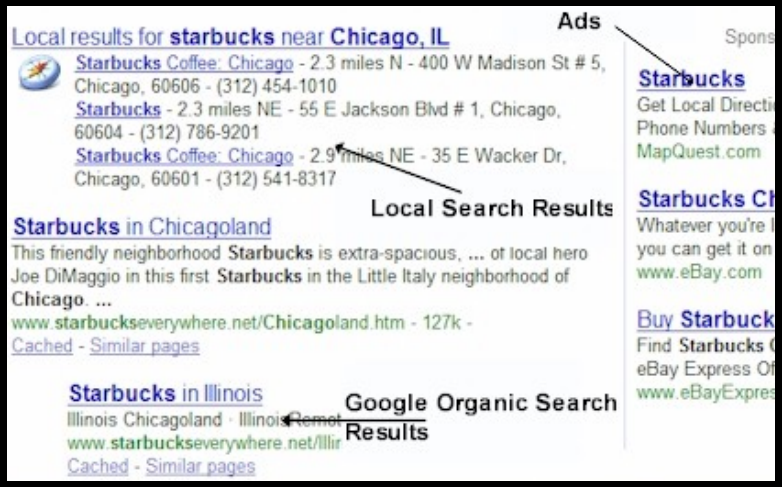
OCR

Autonomous car



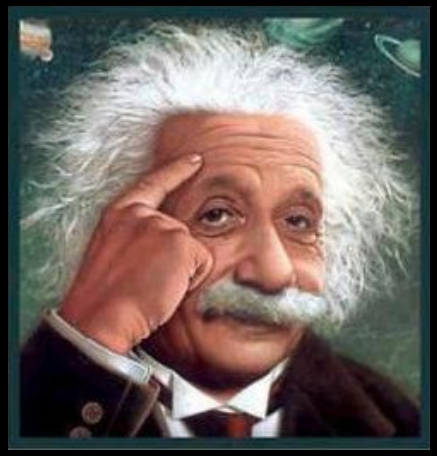
Email classification

Recommendation systems



Web page ranking

Feature Extraction

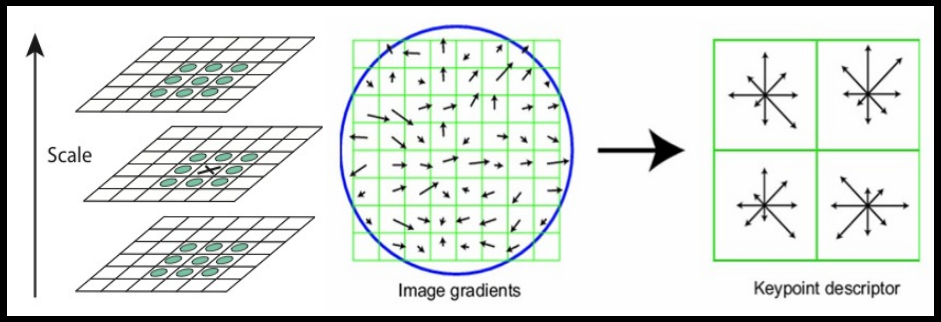


Classifier

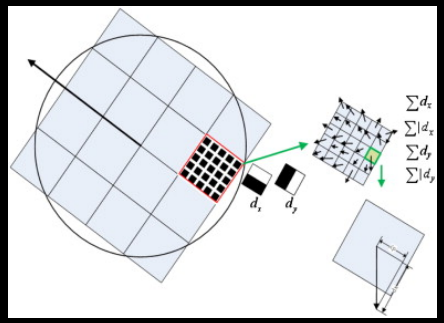
Feature extraction
(Mostly hand-crafted features)

Hand-Crafted Features

Computer vision:



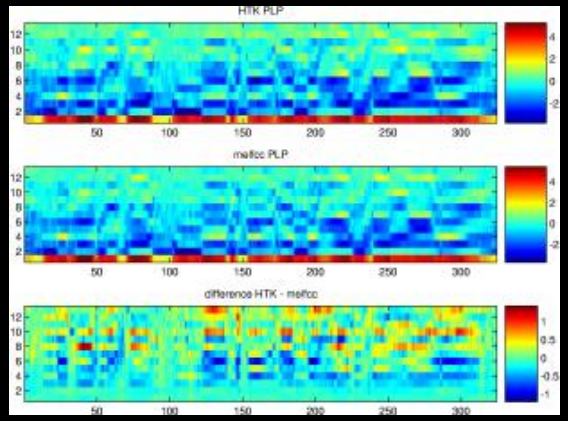
SIFT/HOG



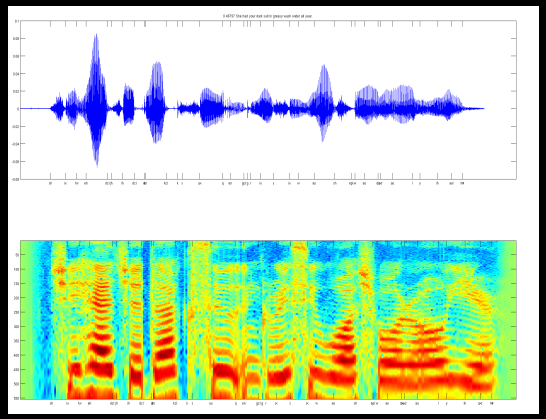
SURF

...

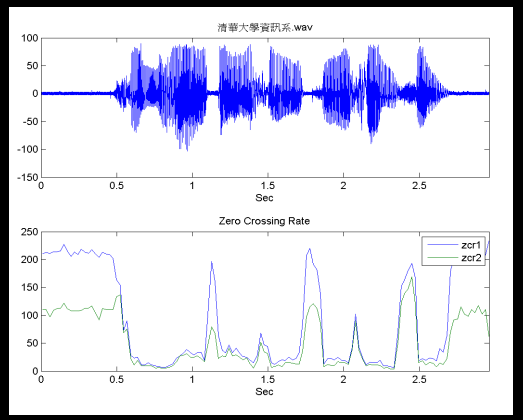
Speech Recognition:



MFCC



Spectrogram



ZCR

...

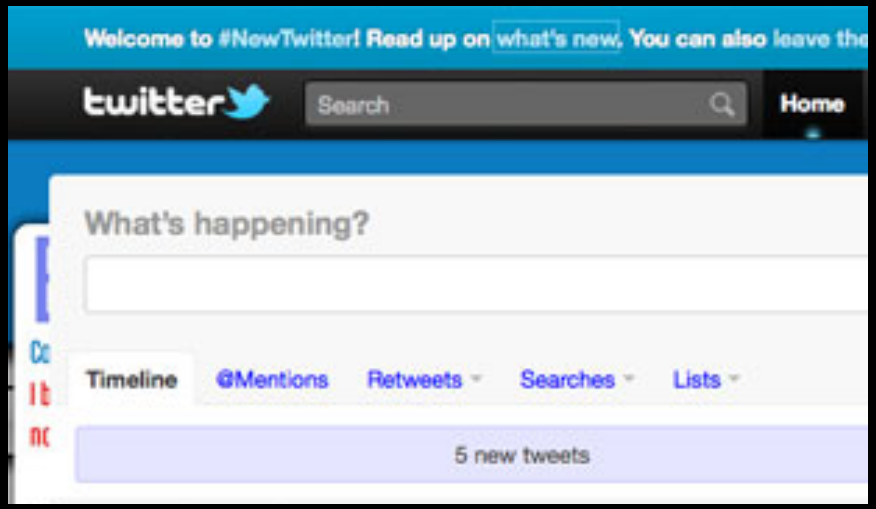
New feature-designing paradigm

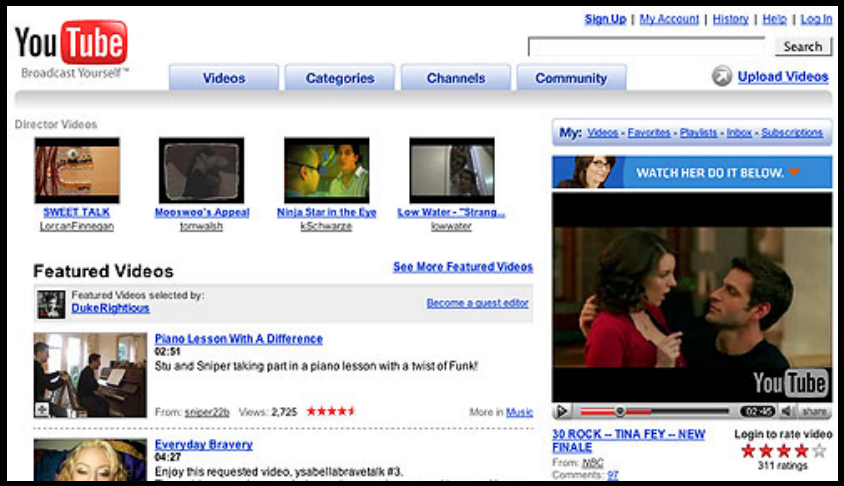
Unsupervised Feature Learning / Deep Learning

Reconstruction ICA

Expensive and typically applied to small problems

The Trend of BigData

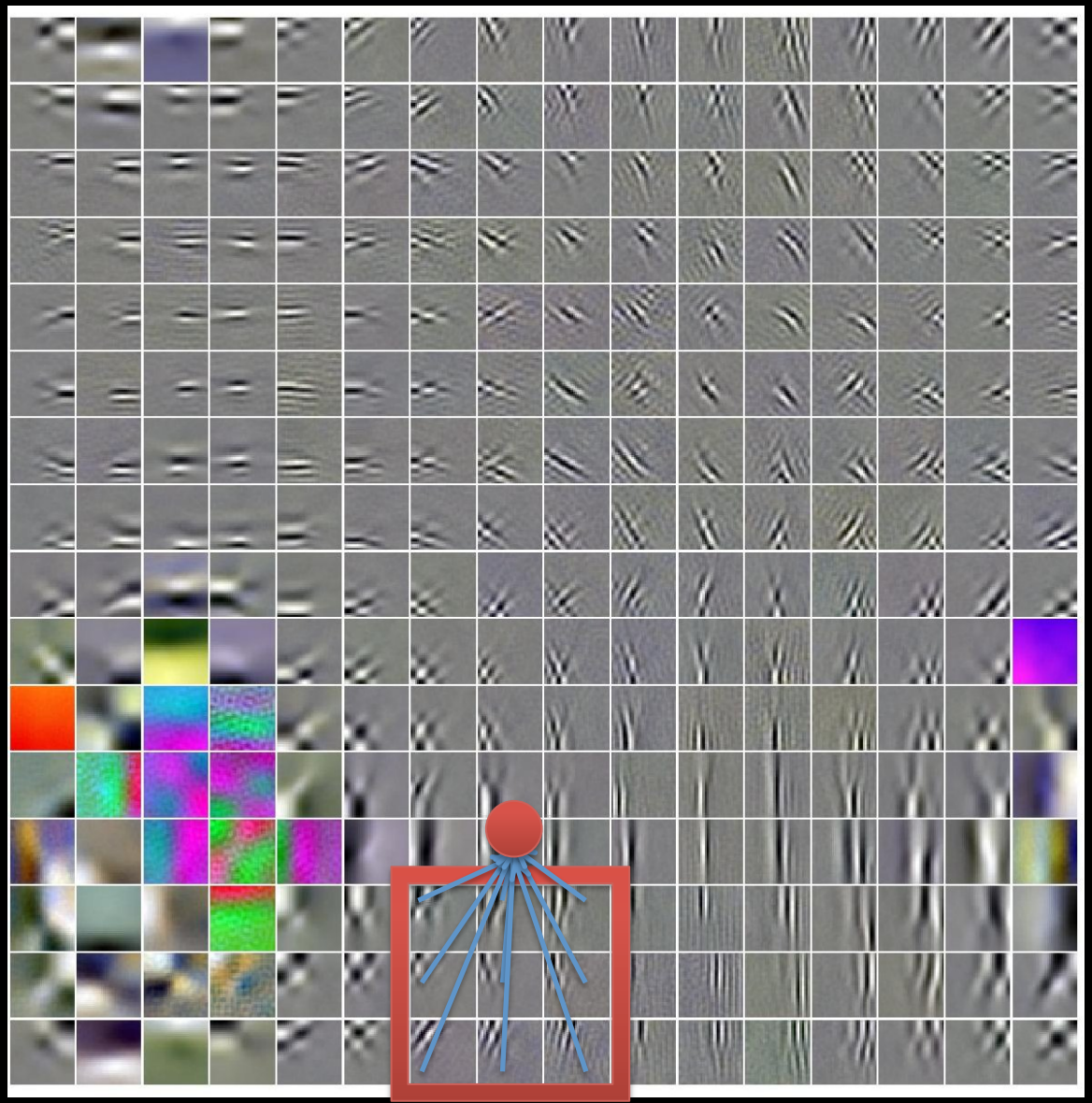




Outline

- Reconstruction ICA
- Applications to videos, cancer images
- Ideas for scaling up
- Scaling up Results

Topographic Independent Component Analysis (TICA)



1. Feature computation

$h_j(W; x^{(i)})$

$(\ w_1^T x^{(i)})^2$ Σ $(\ w_9^T x^{(i)}\ )^2$

$w_1^T x^{(i)}$ $w_9^T x^{(i)}$

w_1 w_9

$x^{(i)}$

Input data:

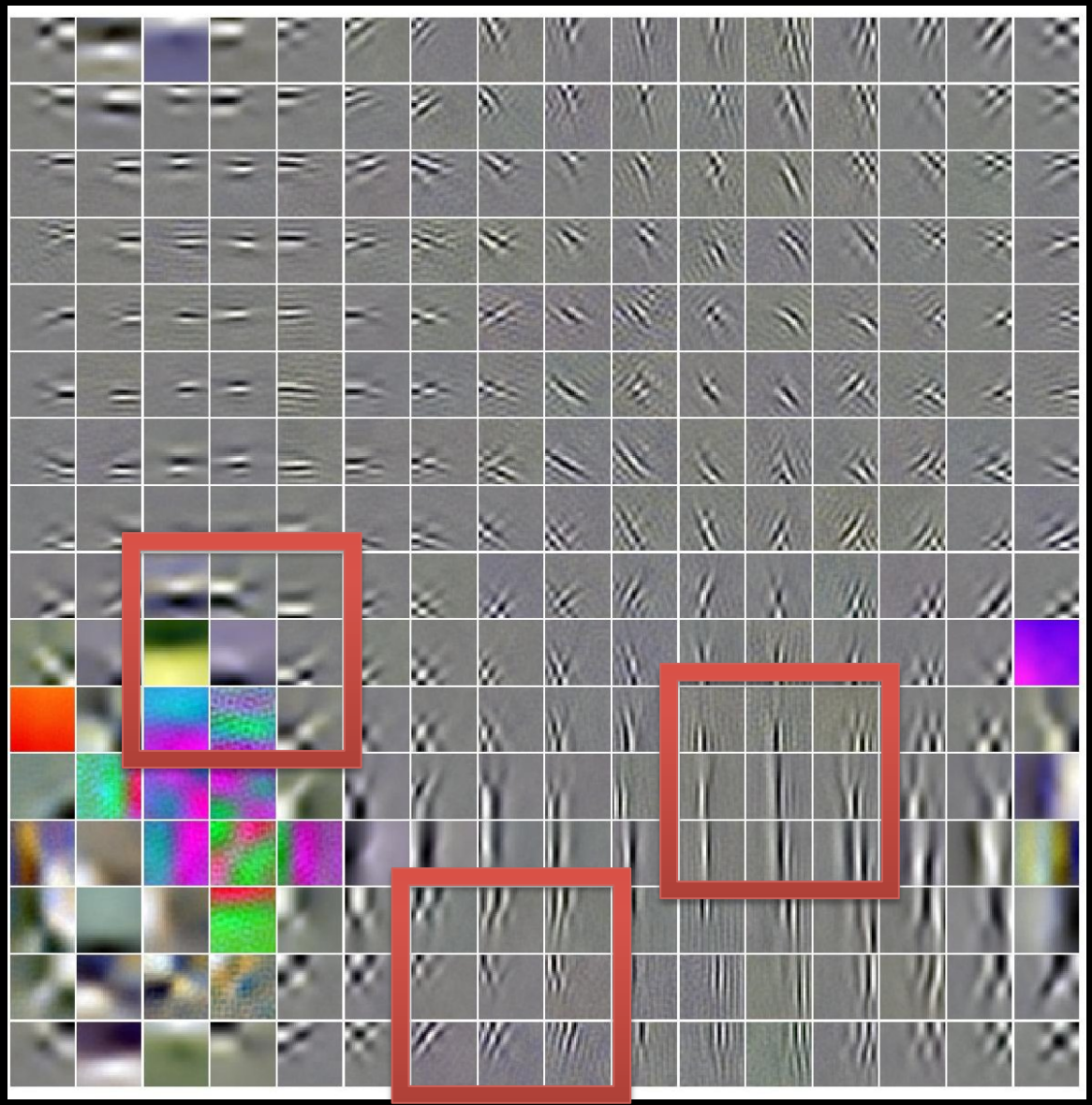
2. Learning

$$\min_W \sum_j \sum_i h_j(W; x^{(i)})$$

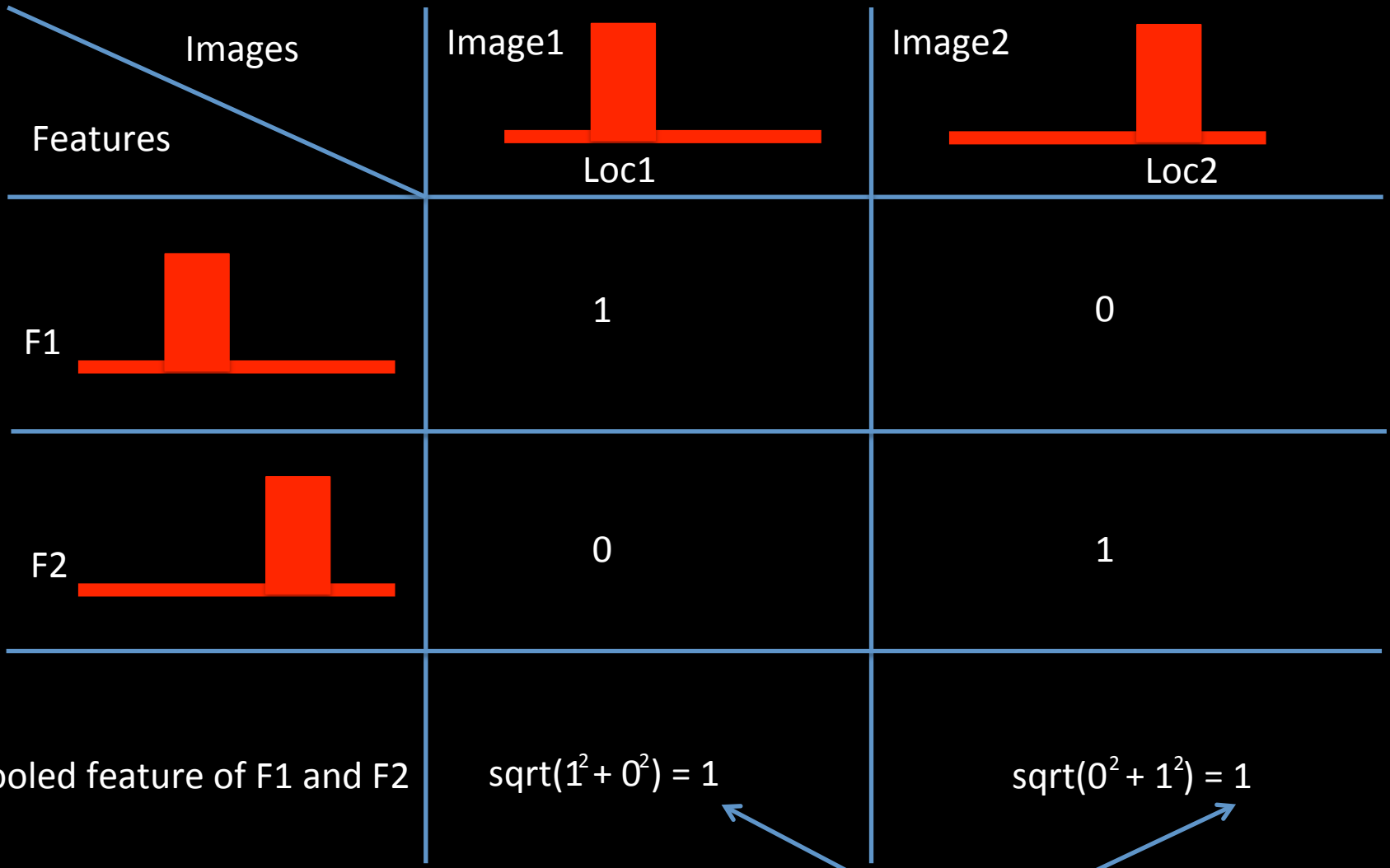
$$s.t. \quad WW^T = I$$

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{10000} \end{bmatrix}$$

Topographic Independent Component Analysis (TICA)



Invariance explained



TICA:

$$\min_W \sum_j \sum_i h_j(W; x^{(i)})$$

$$s.t. \quad WW^T = I$$

Reconstruction ICA:

$$\min_W \frac{\lambda}{m} \sum_{i=1}^{m} \left\| W^T W x^{(i)} - x^{(i)} \right\|_2^2 + \sum_j \sum_i h_j(W; x^{(i)})$$

Lemma 3.1 *When the input data* $\{x^{(i)}\}_{i=1}^m$ *is whitened, the reconstruction cost* $\frac{\lambda}{m} \sum_{i=1}^m \|W^T W x^{(i)} - x^{(i)}\|_2^2$ *is equivalent to the orthonormality cost* $\lambda \|W^T W - \mathbf{I}\|_{\mathcal{F}}^2$.

Lemma 3.2 *The column orthonormality cost* $\lambda \|W^T W - \mathbf{I}_n\|_{\mathcal{F}}^2$ *is equivalent to the row orthonormality cost* $\lambda \|W W^T - \mathbf{I}_k\|_{\mathcal{F}}^2$ *up to an additive constant.*

- Equivalence between Sparse Coding, Autoencoders, RBMs and ICA
- Build deep architecture by treating the output of one layer as input to another layer

Le, et al., *ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning*. NIPS 2011

Reconstruction ICA:

$$\min_W \frac{\lambda}{m} \sum_{i=1}^{m} \left\| W^T W x^{(i)} - x^{(i)} \right\|_2^2 + \sum_j \sum_i h_j(W; x^{(i)})$$

Le, et al., *ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning*. NIPS 2011

Reconstruction ICA:

$$\frac{\lambda}{m}\sum_{i=1}^{m}\left\|W^T W x^{(i)} - x^{(i)}\right\|_2^2$$

Data whitening

$$\lambda\left\|W^T W - I\right\|_F^2 \quad \leftarrow \quad \lambda\left\|WW^T - I\right\|_F^2 \quad \leftarrow \quad WW^T = I$$

Le, et al., *ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning*. NIPS 2011

TICA:

$$\min_W \sum_{i=1}^{m} \sum_{j=1}^{N} p_j^i(W; x^{(i)})$$

$$s.t. \quad WW^T = I$$

Reconstruction ICA:

$$\min_W \frac{\lambda}{m} \sum_{i=1}^{m} \left\| W^T W x^{(i)} - x^{(i)} \right\|_2^2 + \sum_j \sum_i h_j(W; x^{(i)})$$

Data whitening

$$\lambda \left\| W^T W - I \right\|_F^2 \quad \Leftarrow \quad \lambda \left\| WW^T - I \right\|_F^2$$

Le, et al., *ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning*. NIPS 2011

Why RICA?

Algorithms	Speed	Ease of training	Invariant Features
Sparse Coding	✗	✓	✗
RBMs/Autoencoders	✓	✗	✗
TICA	✗	✓	✓
Reconstruction ICA	✓	✓	✓

Le, et al., *ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning*. NIPS 2011

Summary of RICA

- Two-layered network
- Reconstruction cost instead of orthogonality constraints
- Learns invariant features

Applications of RICA

Action recognition

Sit up

Drive Car

Get Out of Car

Eat

Answer phone

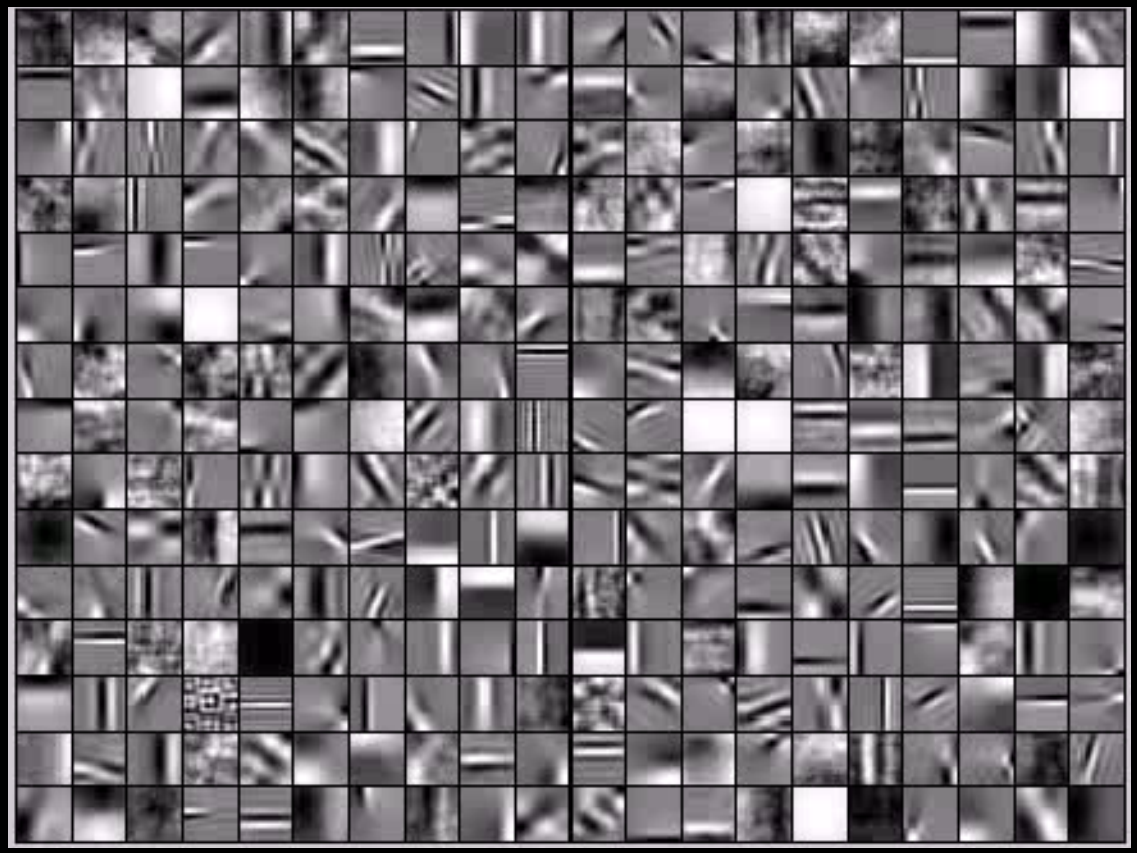
Kiss

Run

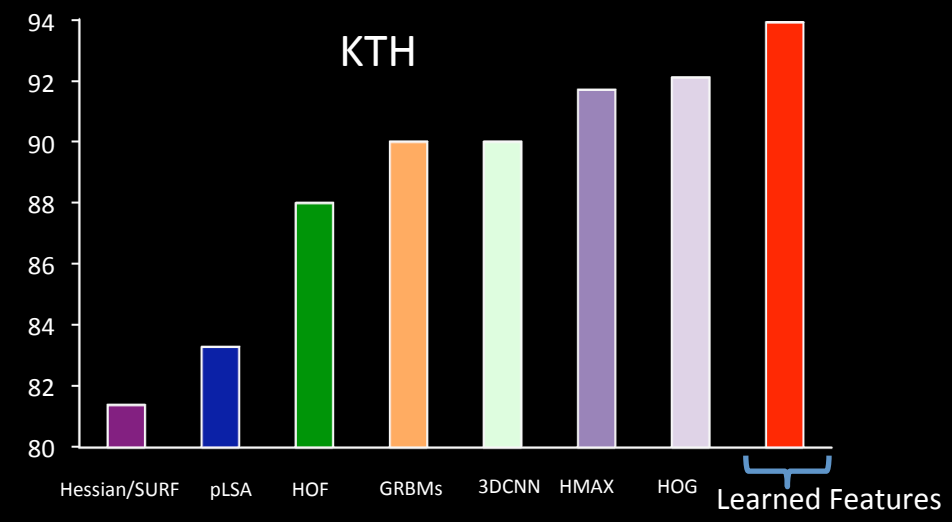
Stand up

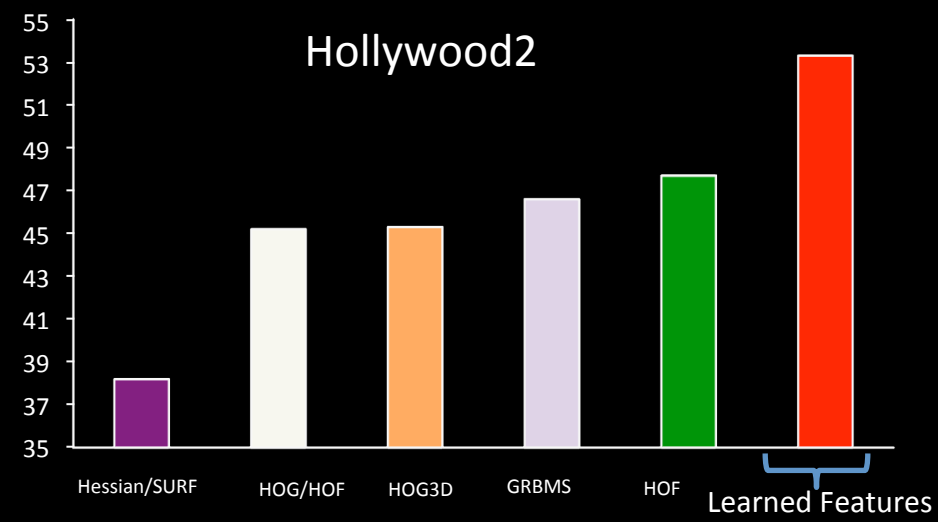
Shake hands

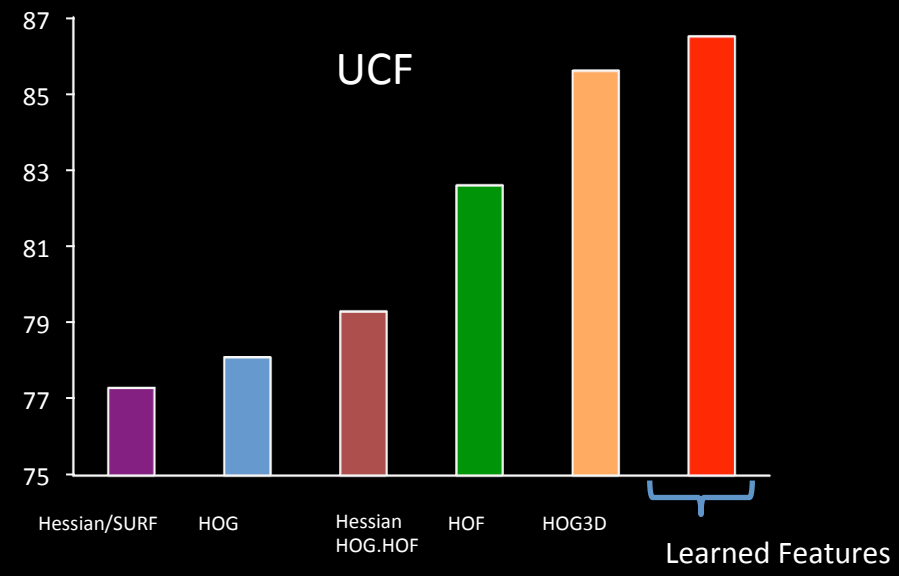
Le, et al., *Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis*. CVPR 2011

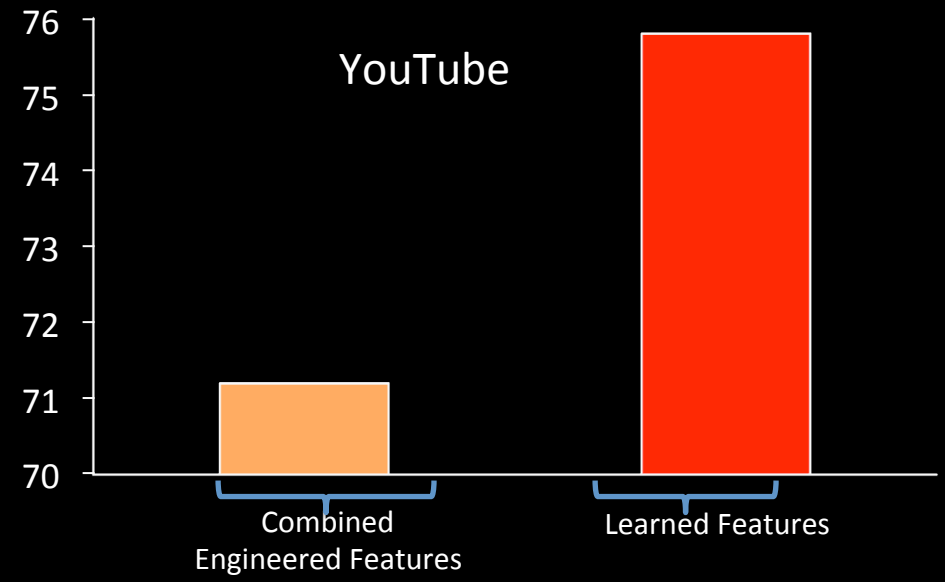


Le, et al., *Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis*. CVPR 2011









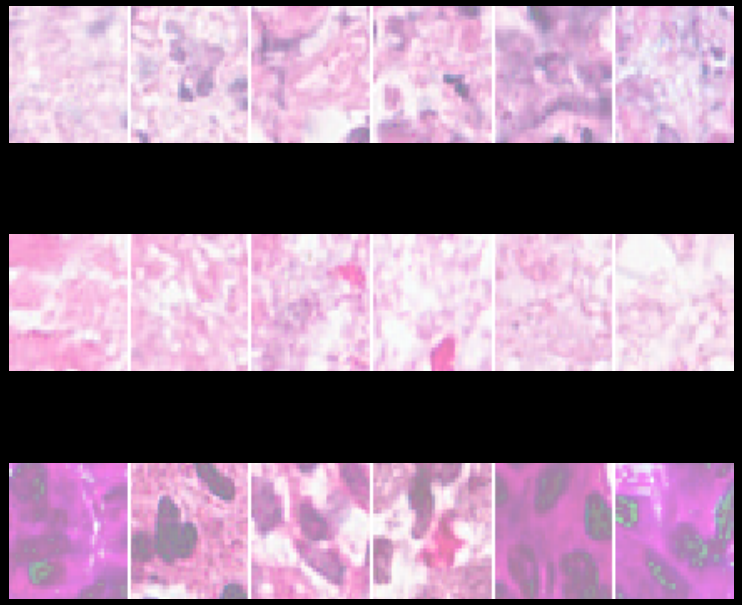
Le, et al., *Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis*. CVPR 2011

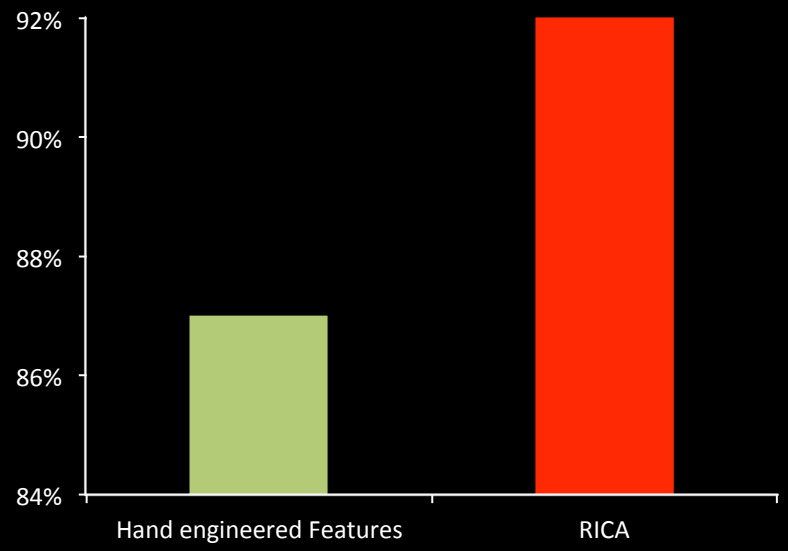
Cancer classification

Apoptotic

Viable tumor region

Necrosis



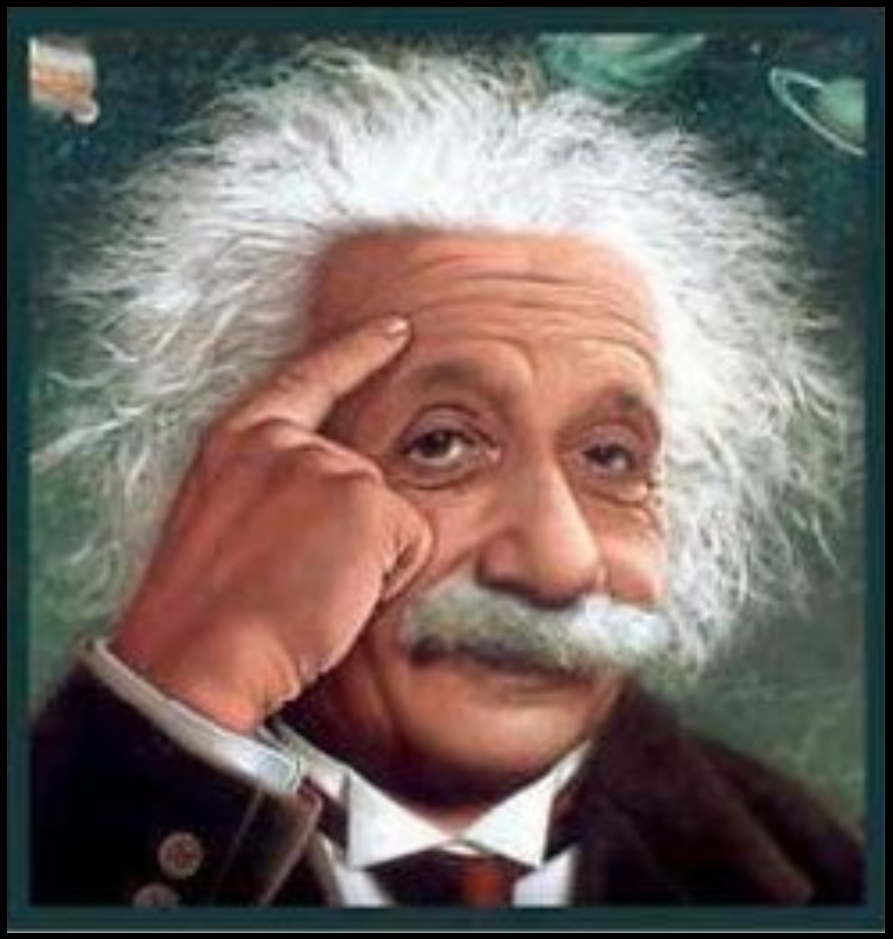


Le, et al., *Learning Invariant Features of Tumor Signatures*. ISBI 2012

Scaling up
deep RICA networks

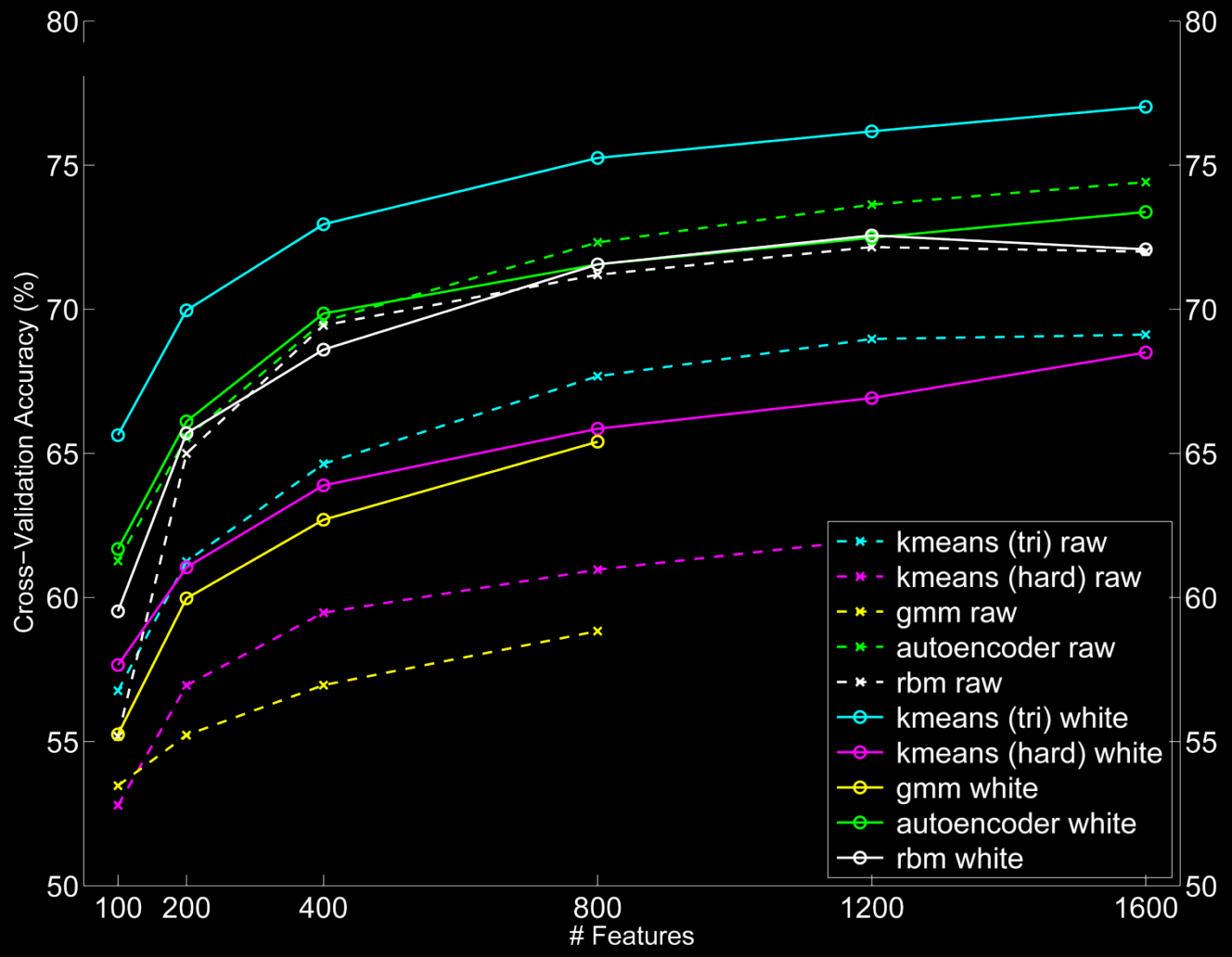
Scaling up Deep Learning

Deep learning data

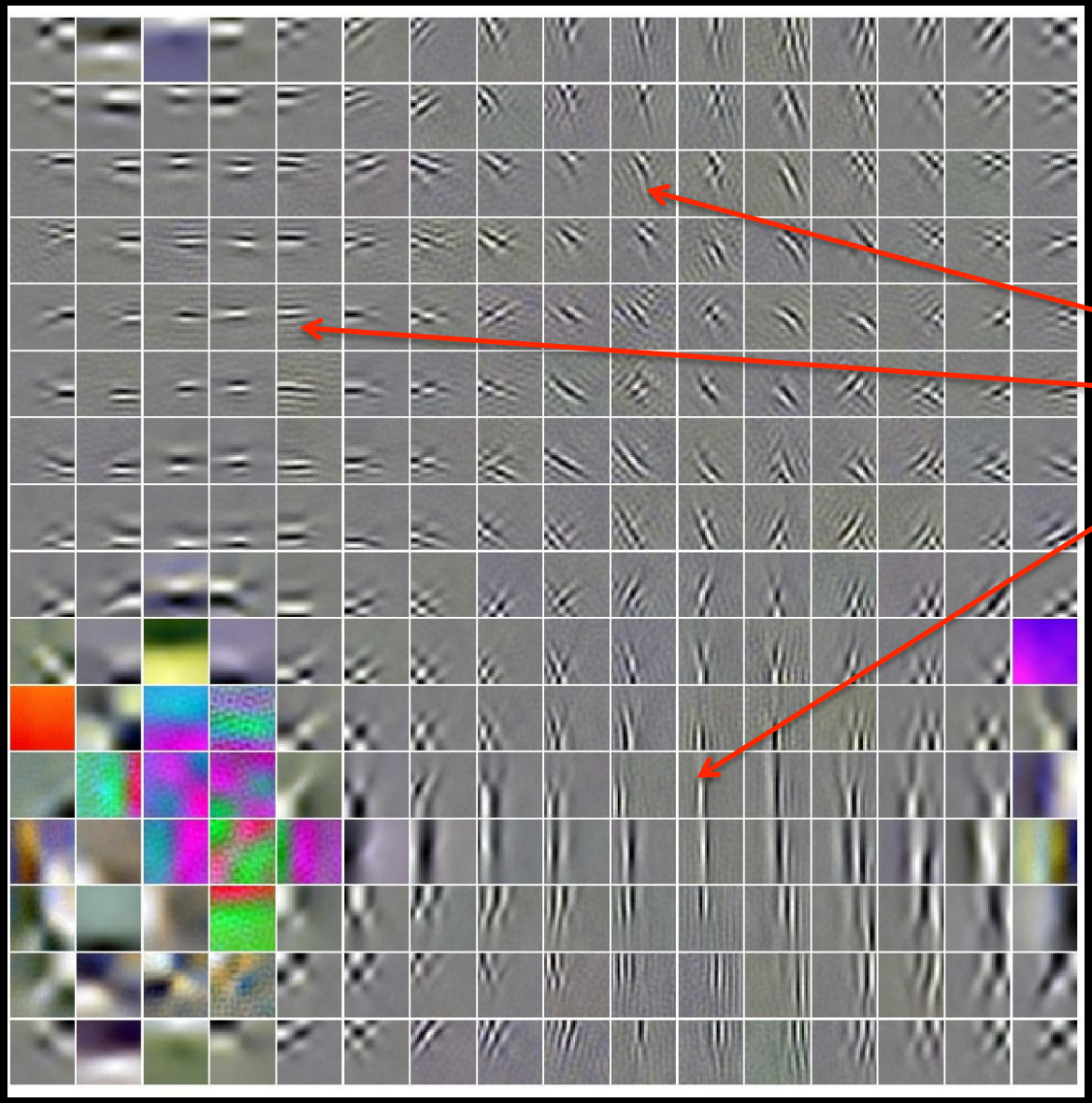


Real data

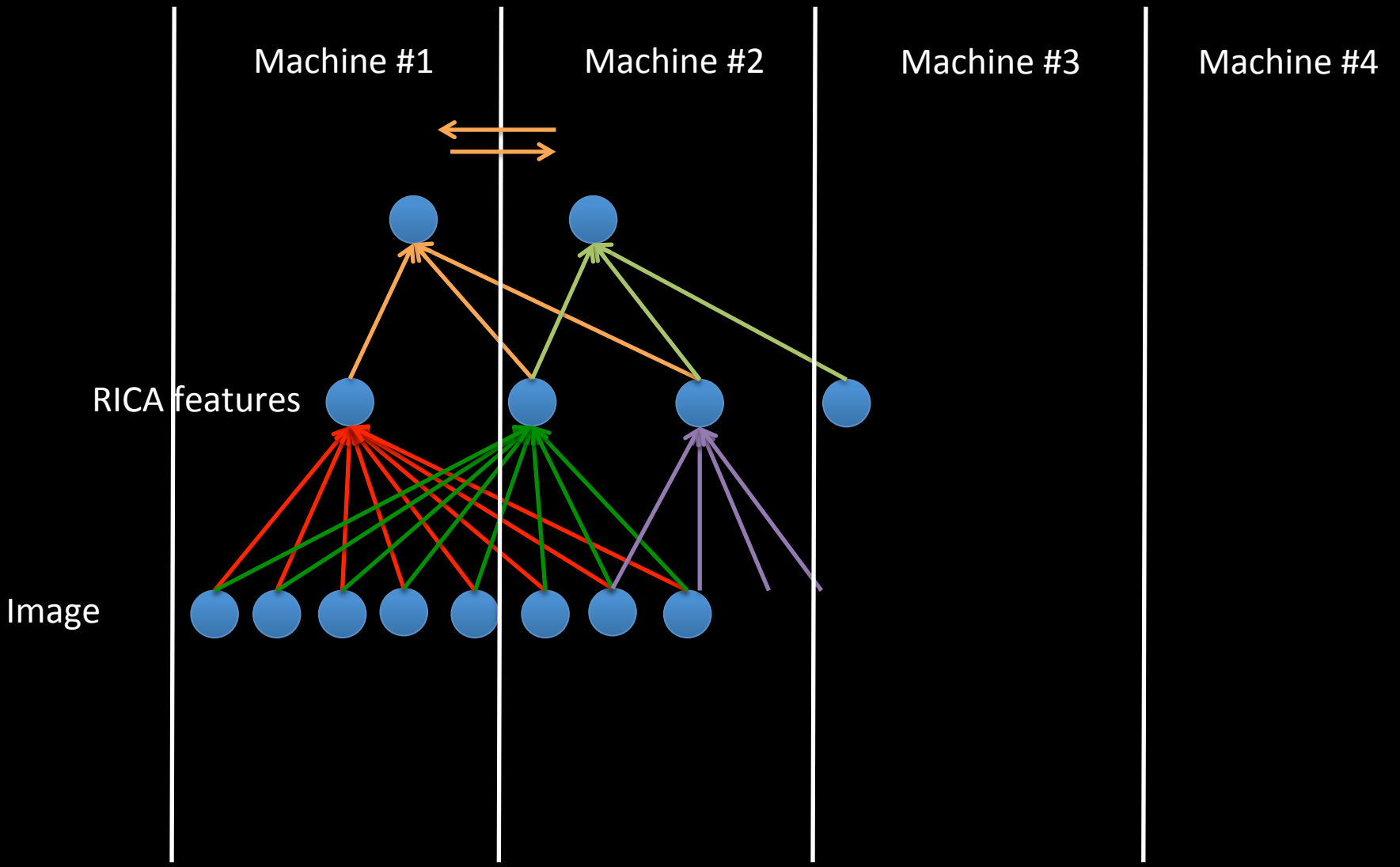
It's better to have more features!



Coates, et al., *An Analysis of Single-Layer Networks in Unsupervised Feature Learning*. AISTATS'11


Most are
local features

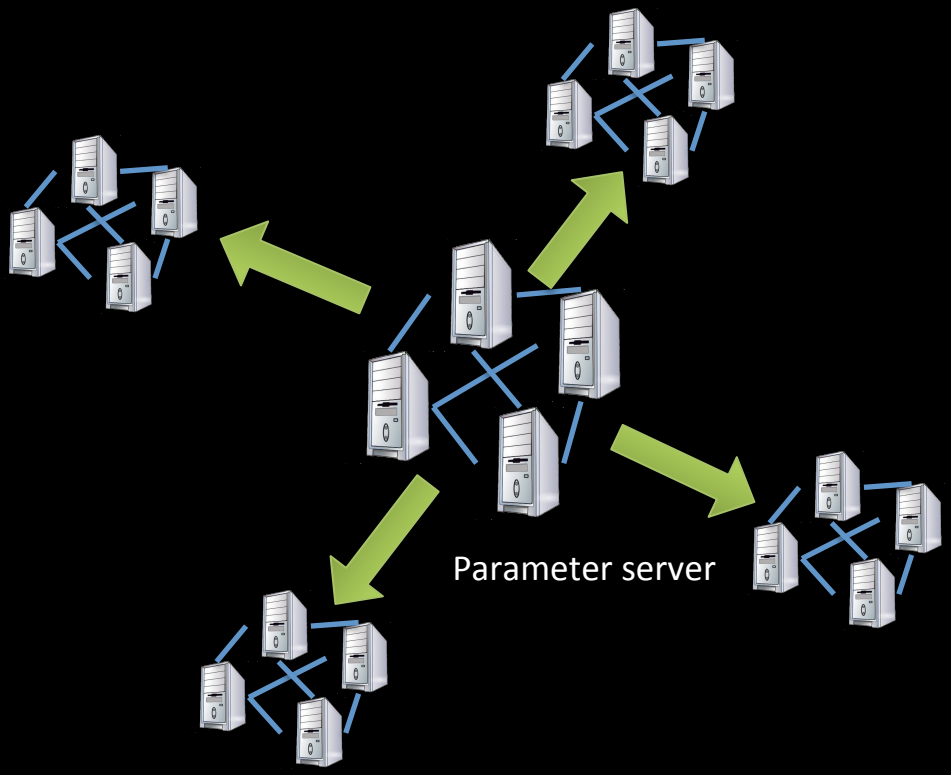
Local receptive field networks



Le, et al., *Tiled Convolutional Neural Networks*. NIPS 2010

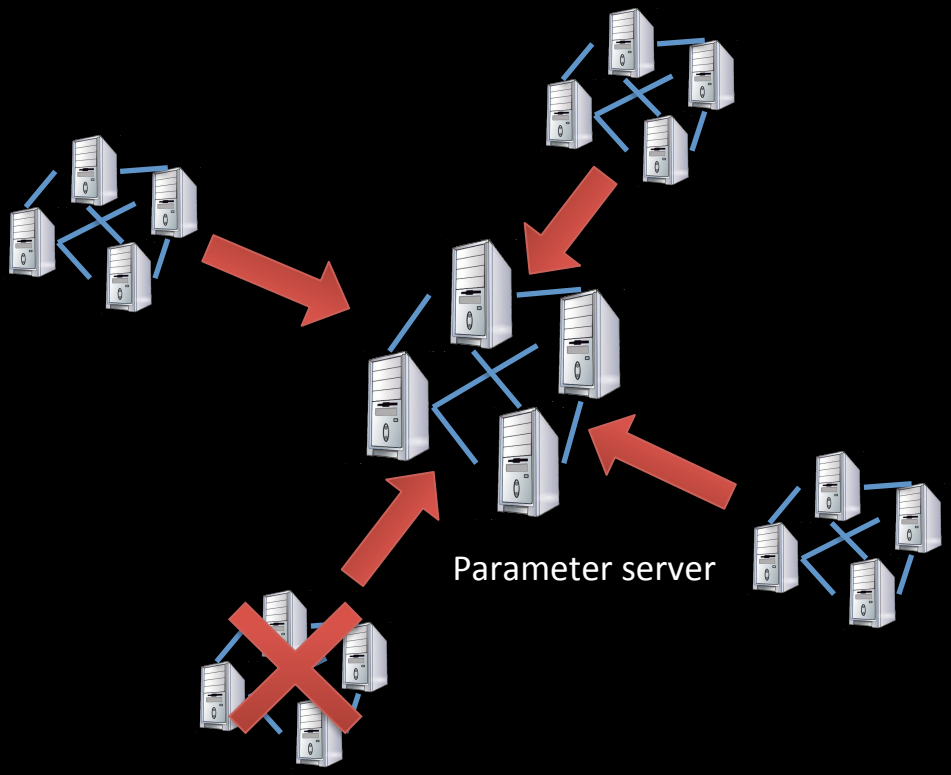
Challenges with 1000s of machines

Asynchronous Parallel SGDs



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Asynchronous Parallel SGDs



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Summary of Scaling up

- Local connectivity
- Asynchronous SGDs

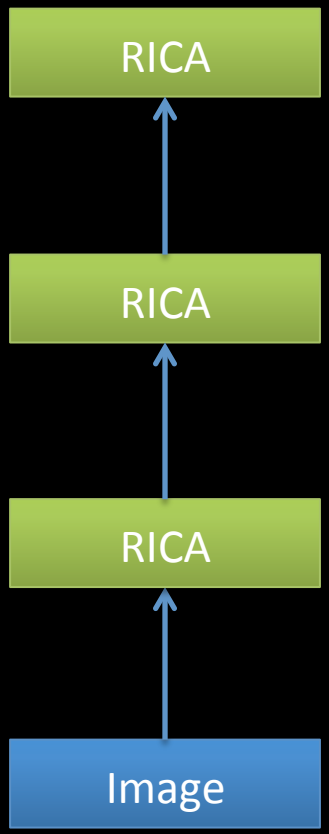
… And more

- RPC vs MapReduce
- Prefetching
- Single vs Double
- Removing slow machines
- Optimized Softmax
- …

10 million 200x200 images

1 billion parameters

Training



Dataset: 10 million 200x200 unlabeled images from YouTube/Web

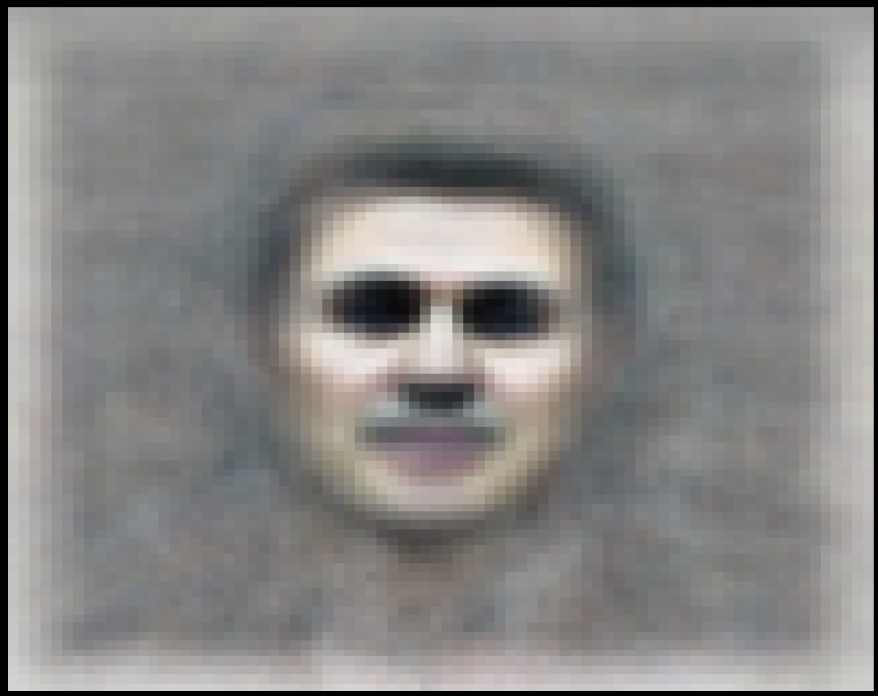
Train on 2000 machines (16000 cores) for 1 week

1.15 billion parameters
- 100x larger than previously reported
- Small compared to visual cortex

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

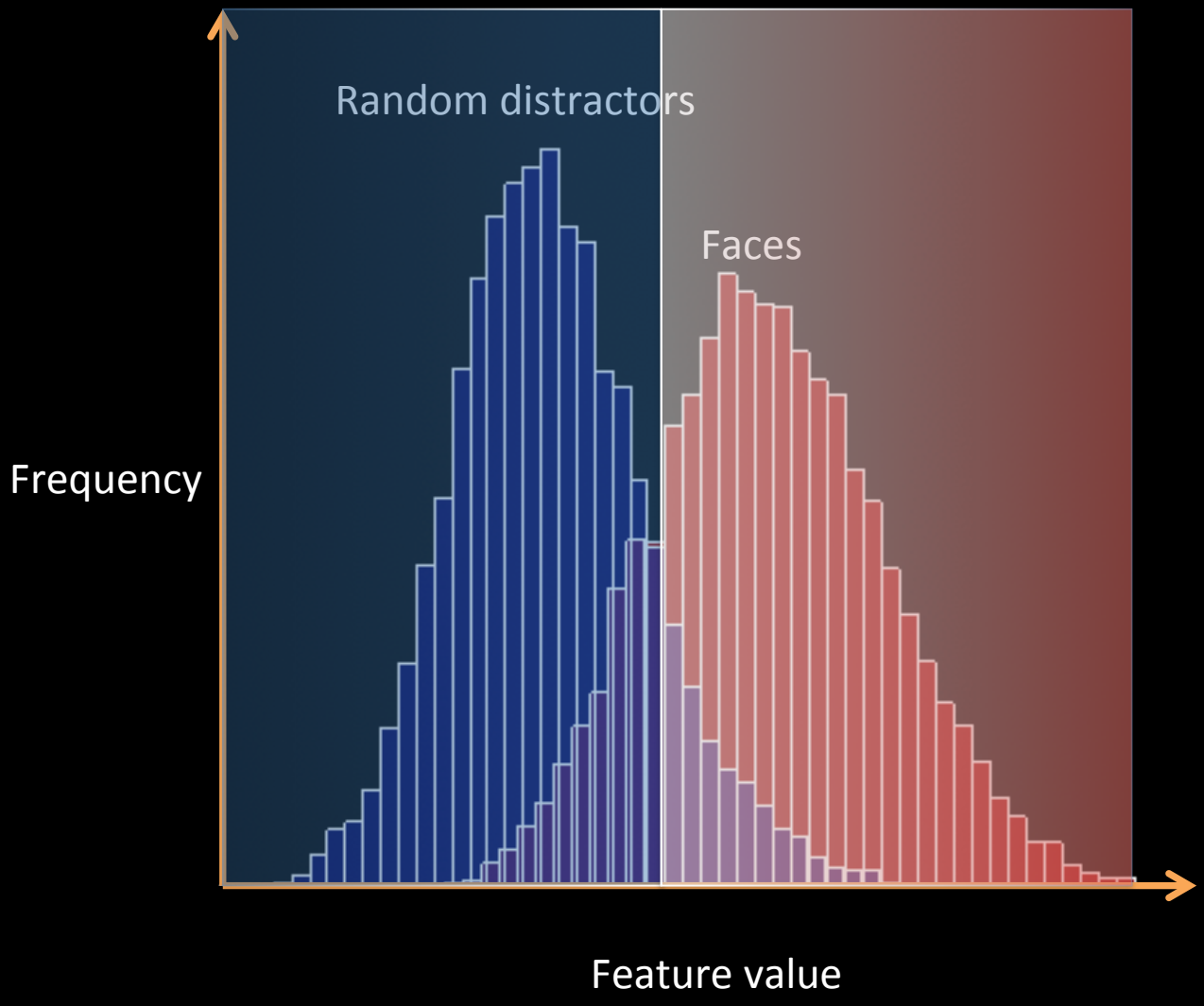
The face neuron

Top stimuli from the test set



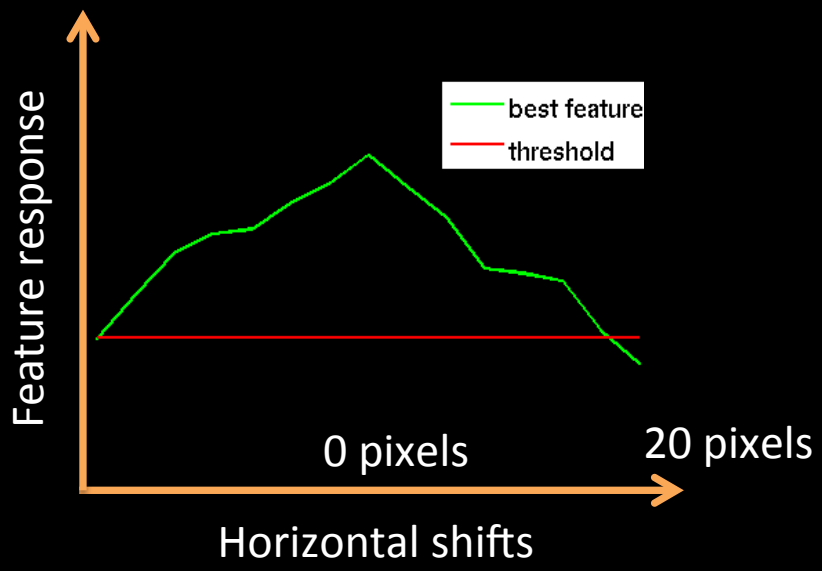
Optimal stimulus
by numerical optimization

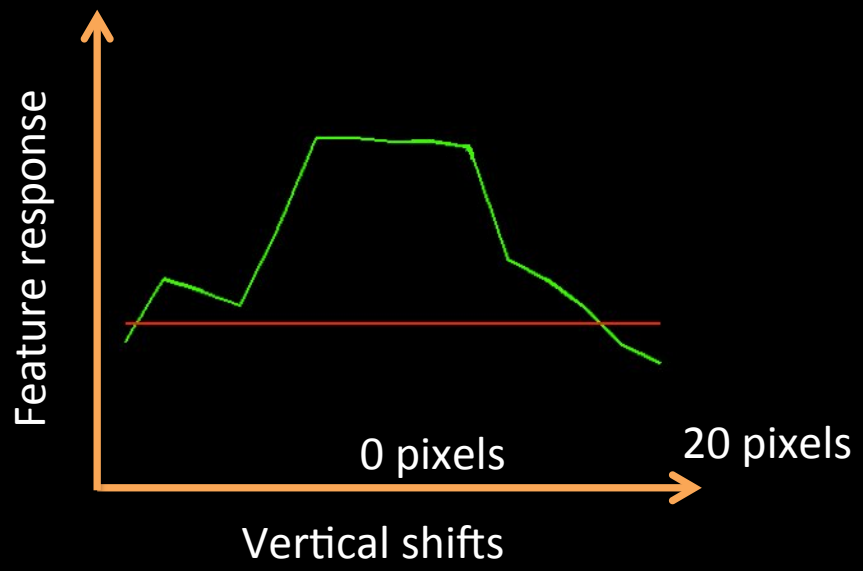
Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

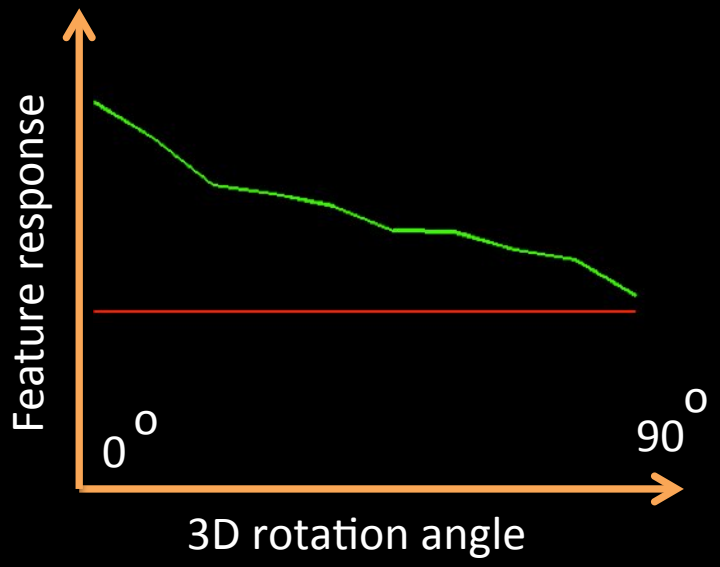


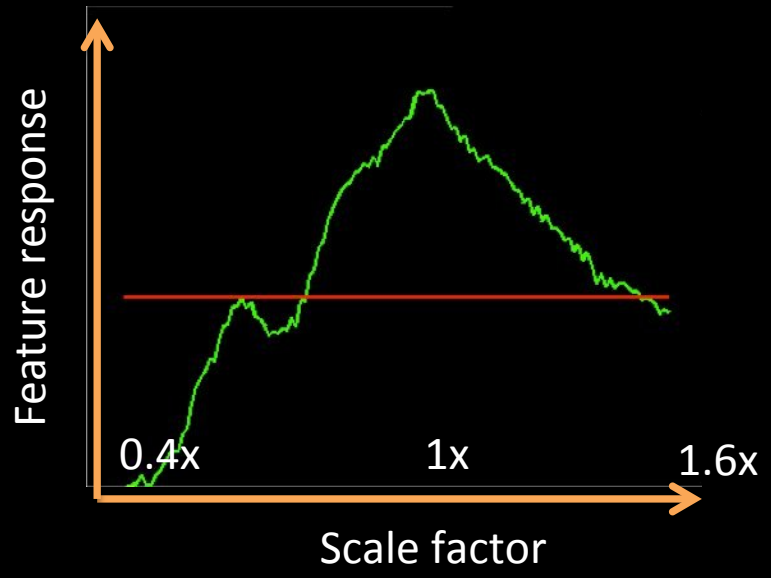
Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Invariance properties







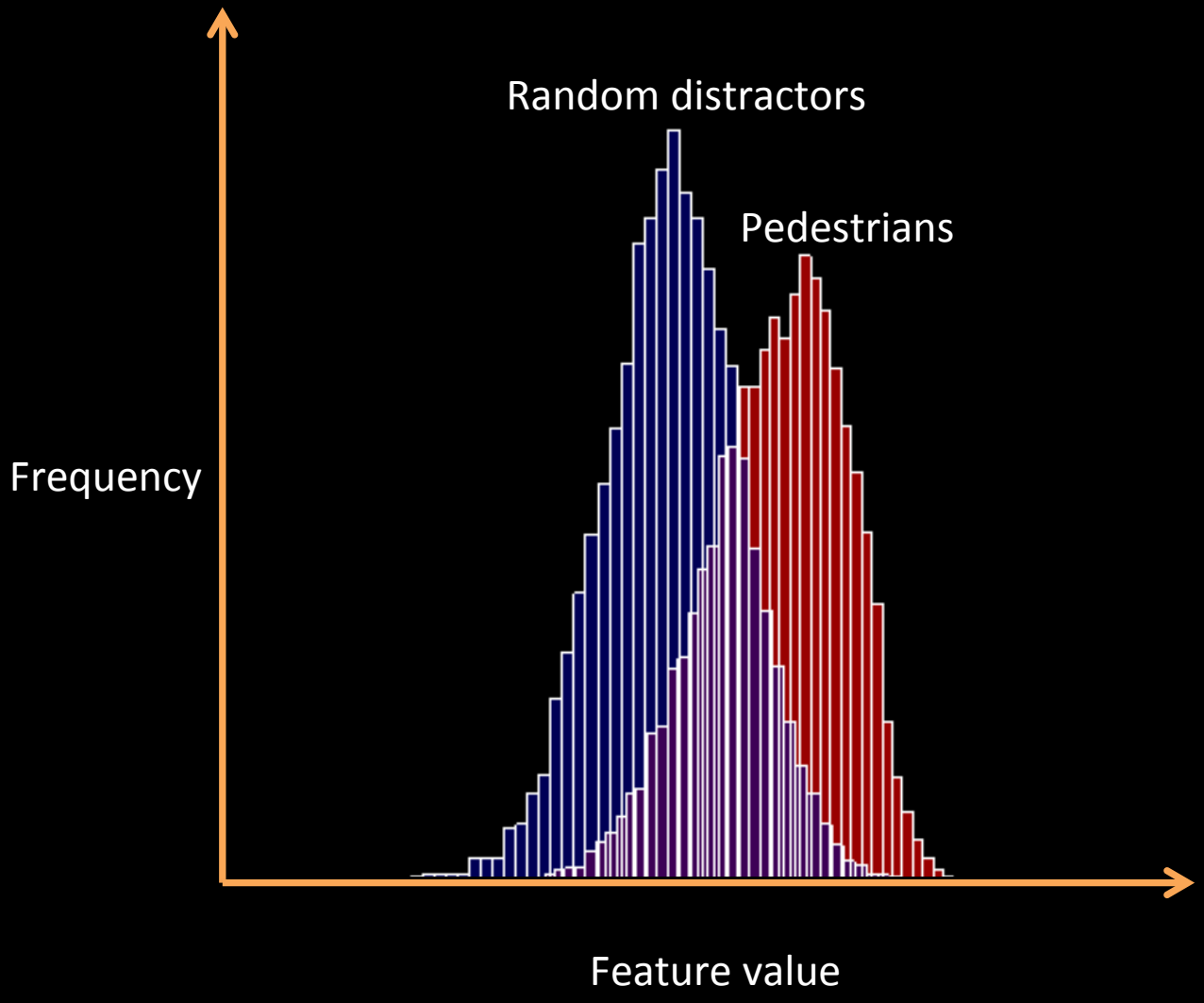


Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Top stimuli from the test set

Optimal stimulus
by numerical optimization

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

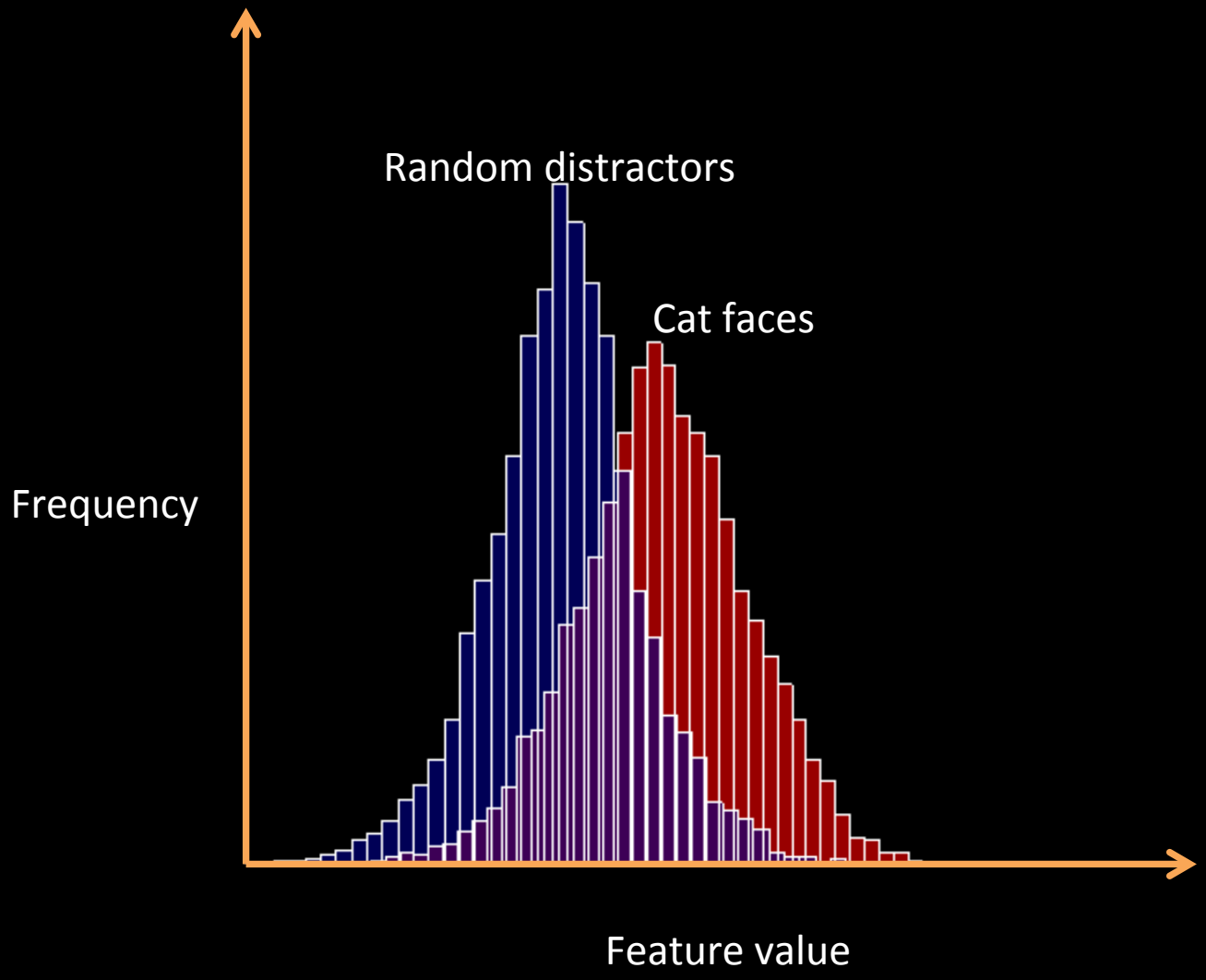


Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Top stimuli from the test set

Optimal stimulus
by numerical optimization

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

The New York Times

Tuesday, June 26, 2012 Last Update: 9:57 AM ET

Turkish Border Is Crucial to Syrian Fight as Rebels Evolve

By NEIL MacFARQUHAR

A network of activists has taken advantage of the tensions between Turkey and Syria to build a supply chain for those opposed to the government of President Bashar al-Assad.

· Photographs | Video
· Assad Supporters Suspected in Beirut Unrest 8:07 AM ET

Post a Comment | Read (15)

Turkey Warns Syrian Forces Not to Approach Border

By SEBNEM ARSU, ALAN COWELL and PAUL GEITNER
12 minutes ago

Buoyed by support from NATO allies, the Turkish

LIVE DAILY AT 10 AM ET

Jim Wilson/The New York Times

Why News Corporation may divide itself in two. | The anxiety of being a digital parent. | Google's artificial brain scours the Internet in search of ... cats.

DEALBOOK

News Corporation Considers Dividing Itself

By MICHAEL J. DE LA MERCED

The embattled media company may separate its publishing arm from its larger entertainment division.

Euro Zone Leaders to Debate Joint Banking Union

By STEPHEN CASTLE 8:51 AM ET

Details of a plan for tighter fiscal unity, including the creation

OPINION »

THE SUPREME COURT ACTS

Editorial: The basis of the harsh Arizona immigration law is rejected by the court.

Editorial: The justices allow the biggest donors to keep buying elections.

· Brooks: The Particular
· Bruni: Capt
· Nocera: The
· Room for D Muslim Bro
· Gil Troy: Pr Politicking

MARKETS » As of 10:03 AM ET

S&P 500	Dow	Nasdaq
1,317.65	12,515.03	2,851.37
+3.83	+12.37	+15.81
+0.30%	+0.10%	+0.56%

GET QUOTES My Portfolio »

Stock, ETFs, Funds Go

ImageNet classification

22,000 categories

14,000,000 images

Hand-engineered features (SIFT, HOG, LBP),
Spatial pyramid, SparseCoding/Compression

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

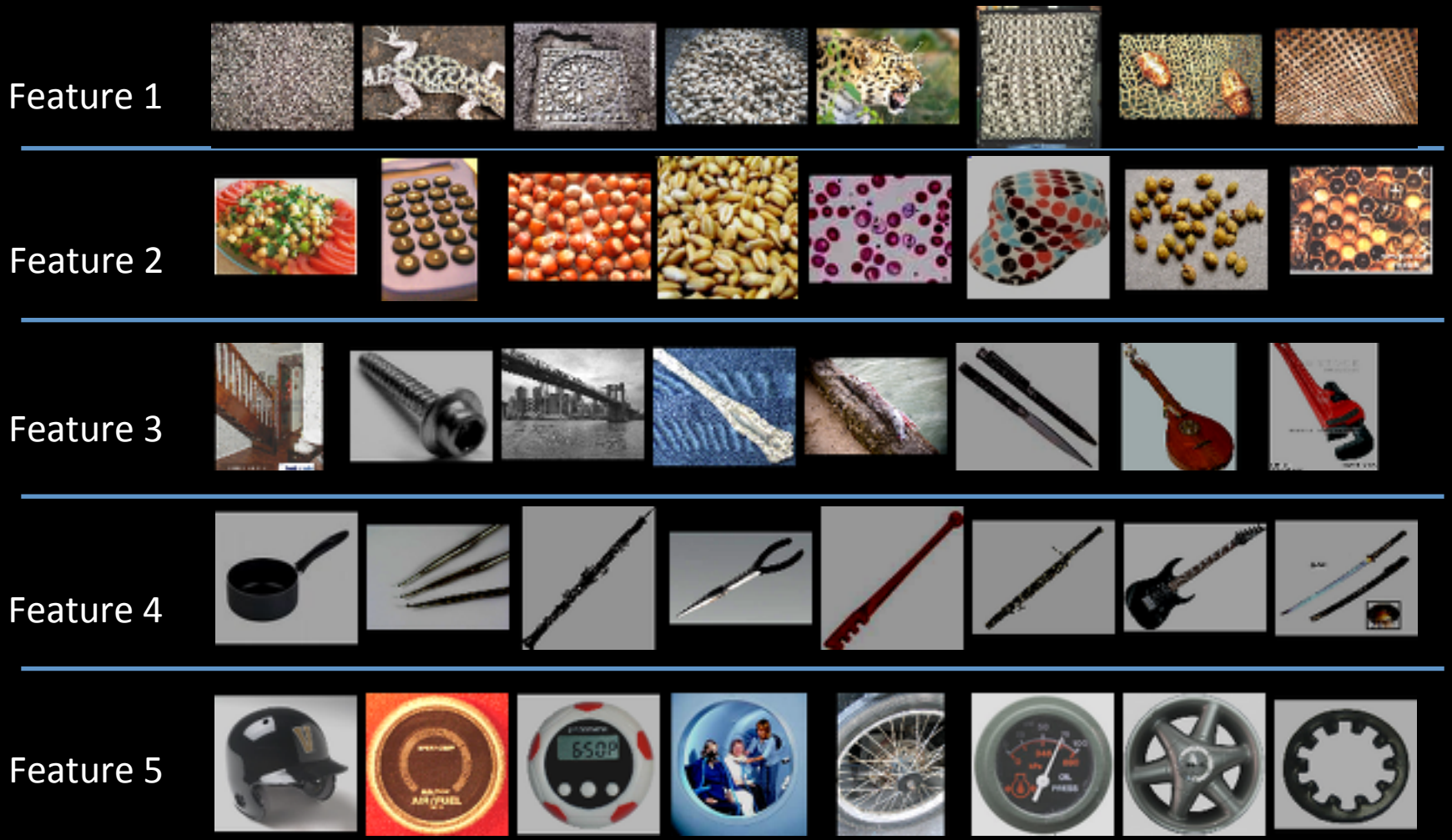
22,000 is a lot of categories…

…
smoothhound, smoothhound shark, Mustelus mustelus
American smooth dogfish, Mustelus canis
Florida smoothhound, Mustelus norrisi
whitetip shark, reef whitetip shark, Triaenodon obseus
Atlantic spiny dogfish, Squalus acanthias
Pacific spiny dogfish, Squalus suckleyi
hammerhead, hammerhead shark
smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, Sphyrna tiburo
angel shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish
roughtail stingray, Dasyatis centroura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish
Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea
…

Stingray

Mantaray

Best stimuli



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

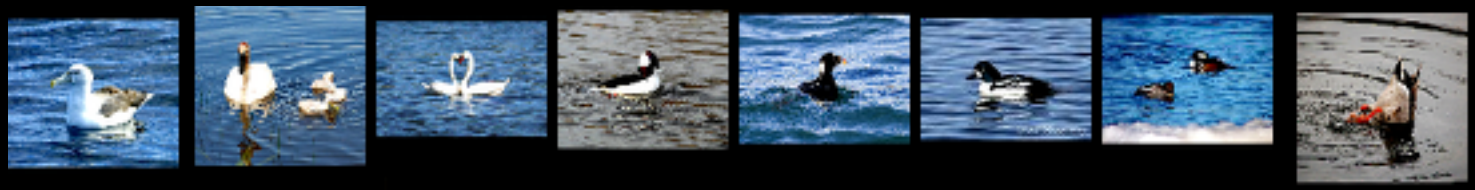
Best stimuli

Feature 6

Feature 7

Feature 8

Feature 9



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Best stimuli

Feature 10

Feature 11

Feature 12

Feature 13

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

0.005%

Random guess

9.5%

State-of-the-art
(Weston, Bengio ‘11)

?

Feature learning
From raw pixels

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

0.005%
Random guess

9.5%
State-of-the-art
(Weston, Bengio '11)

15.8%
Feature learning
From raw pixels

ImageNet 2009 (10k categories): Best published result: 17%
(Sanchez & Perronnin '11),
Our method: 20%

Using only 1000 categories, our method > 50%

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

Other results

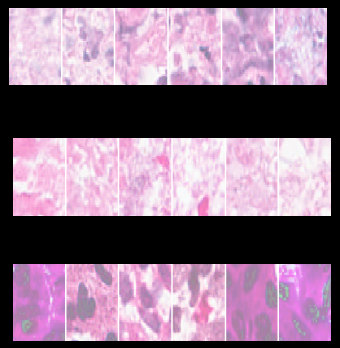
- We also have great features for
  - Speech recognition
  - Word-vector embedding for NLPs

Conclusions

- RICA learns invariant features
- Face neuron with totally unlabeled data with enough training and data
- State-of-the-art performances on
  - Action Recognition
  - Cancer image classification
  - ImageNet

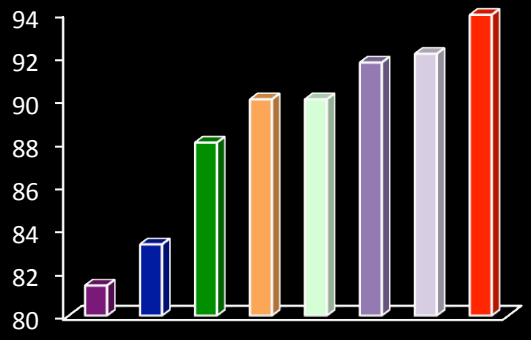
ImageNet

0.005%	9.5%	15.8%
Random guess	Best published result	Our method



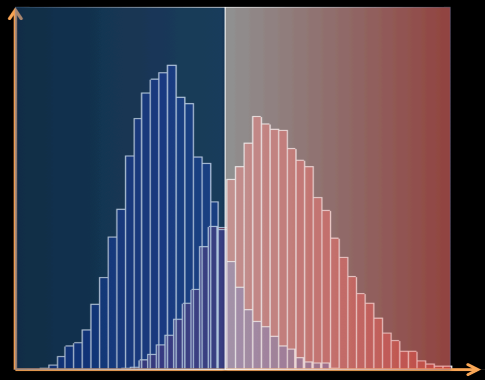
Cancer classification

Action recognition

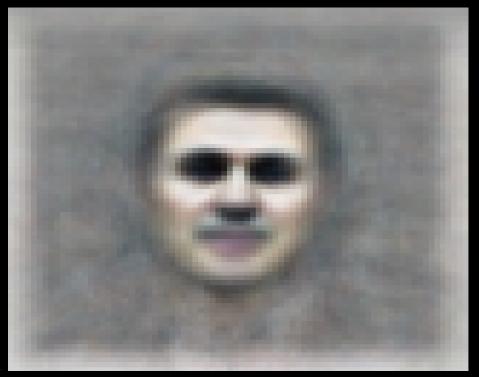


Action recognition benchmarks

Feature visualization



Face neuron



References

- Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, A.Y. Ng. **Building high-level features using large-scale unsupervised learning**. *ICML*, 2012.
- Q.V. Le, J. Ngiam, Z. Chen, D. Chia, P. Koh, A.Y. Ng. **Tiled Convolutional Neural Networks**. *NIPS*, 2010.
- Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng. **Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis**. *CVPR*, 2011.
- Q.V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, A.Y. Ng. **On optimization methods for deep learning**. *ICML*, 2011.
- Q.V. Le, A. Karpenko, J. Ngiam, A.Y. Ng. **ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning**. *NIPS*, 2011.
- Q.V. Le, J. Han, J. Gray, P. Spellman, A. Borowsky, B. Parvin. **Learning Invariant Features for Tumor Signatures**. *ISBI*, 2012.
- I.J. Goodfellow, Q.V. Le, A.M. Saxe, H. Lee, A.Y. Ng, **Measuring invariances in deep networks**. *NIPS*, 2009.

http://ai.stanford.edu/~quocle