Understanding Deep Learning and Neural Semantics

Xiaogang Wang

Department of Electronic Engineering,
The Chinese University of Hong Kong

Machine Learning

$x \longrightarrow \boxed{F(x)} \longrightarrow y$

y:
- Class label (Classification)
- Vector (Estimation)

Object recognition →

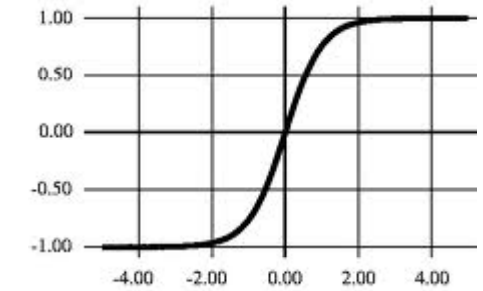
{dog, cat, horse, flower, …}

Low-resolution image

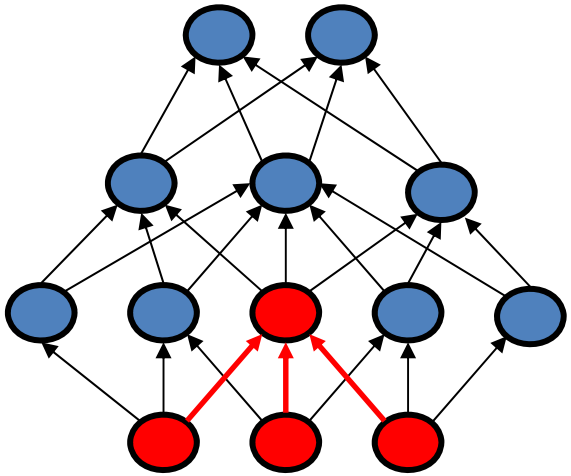
Super resolution →

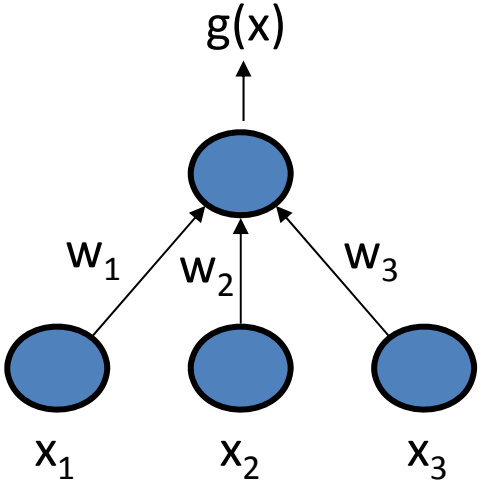
High-resolution image

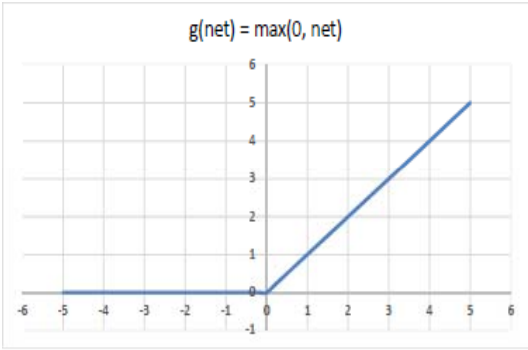
Neural Network



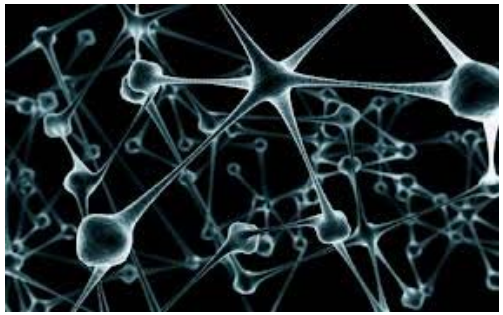
f(*net*)

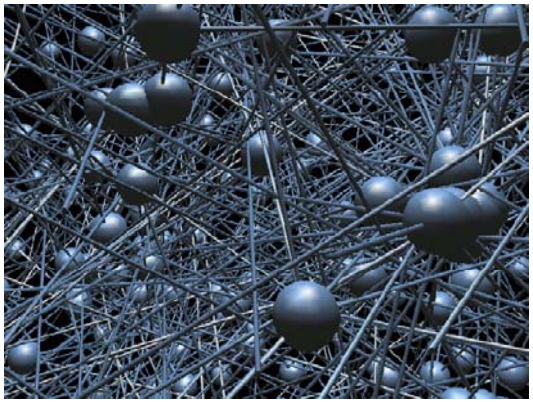






$$g(\mathbf{x}) = f(\sum_{i=1}^{d} x_i w_i + w_0) = f(\mathbf{w}^t \mathbf{x})$$





Highly complex neural networks with many layers, millions or billions of neurons, and sophisticated architectures

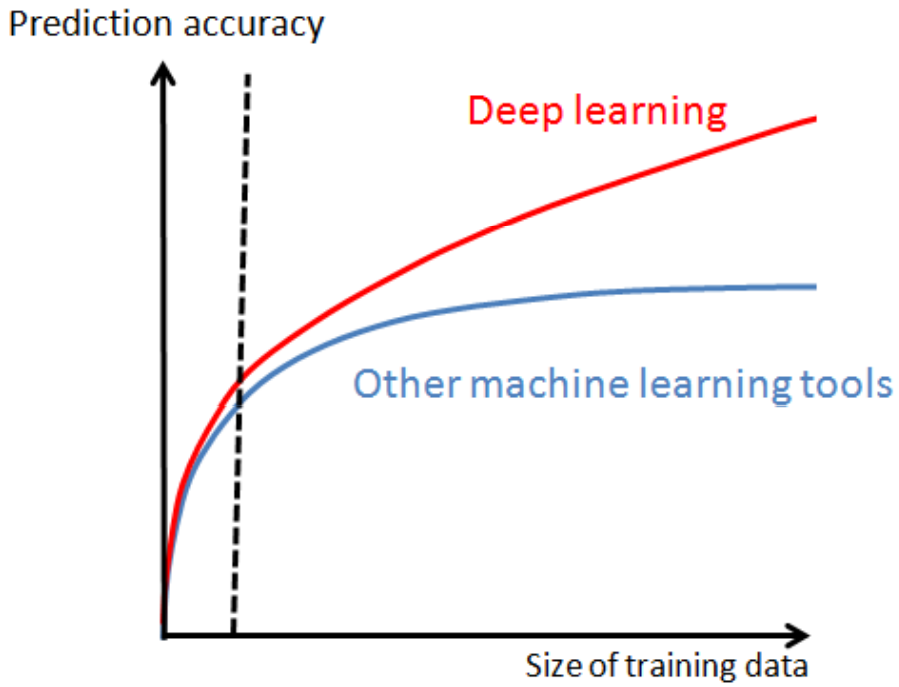
Fit billions of training samples

Trained with GPU clusters with millions of processors

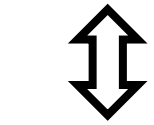
Deep learning

Machine Learning with Big Data

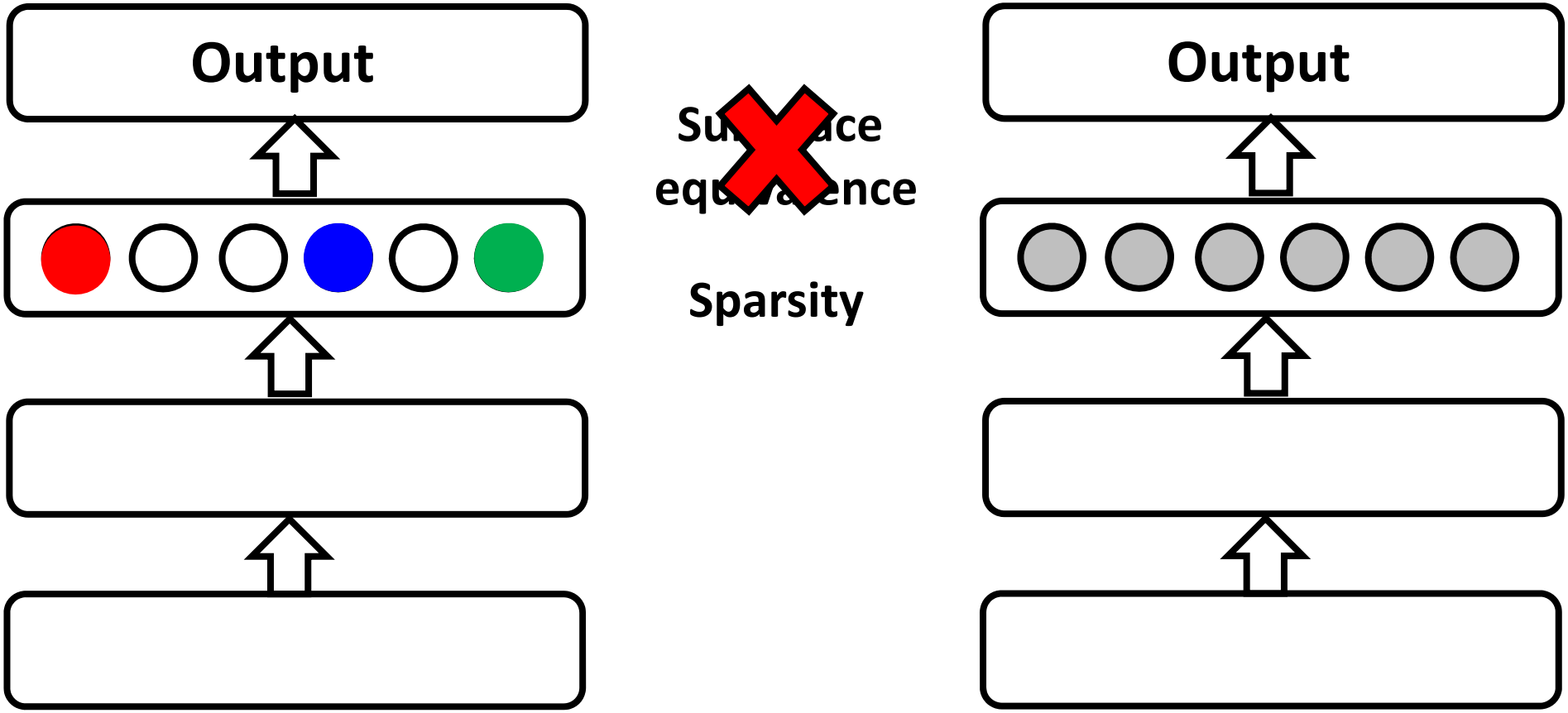
- Machine learning with small data: **overfitting**, reducing model complexity (capacity), adding regularization
- Machine learning with big data: **underfitting**, increasing model complexity, optimization, computation resource



Selectiveness



Sparsity

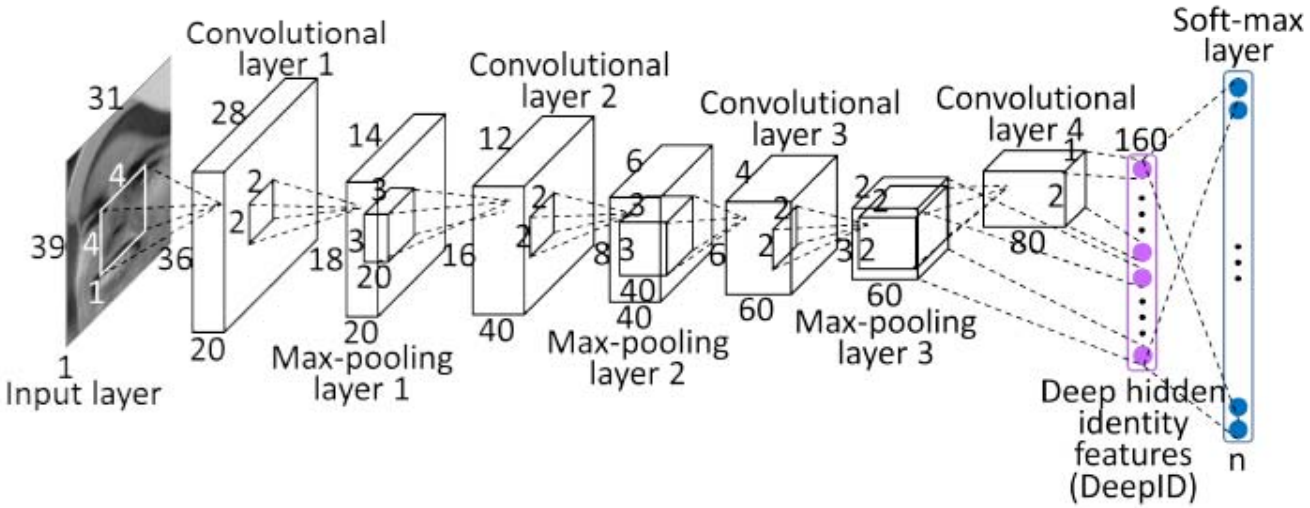


Outline

- **Face recognition and analysis**
- Object tracking
- Human pose estimation

DeepID2: Joint Identification (Id)-Verification (Ve) Signals

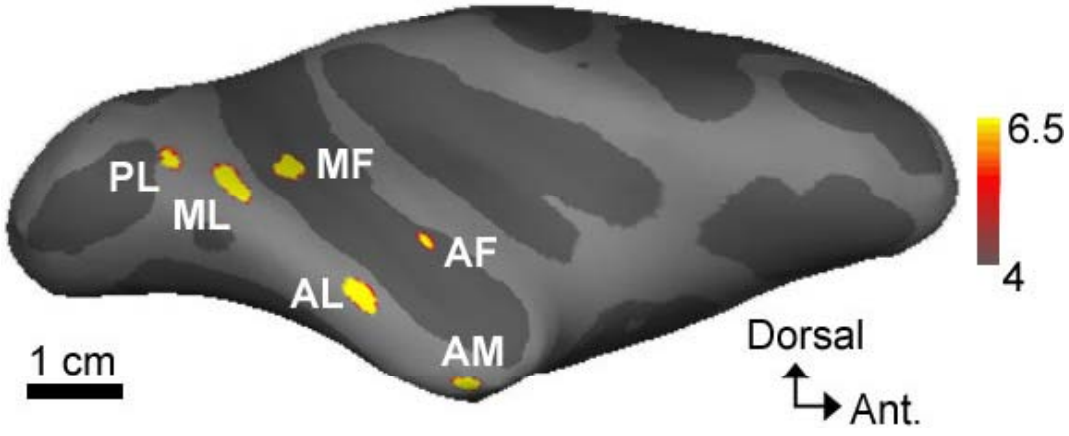
$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \|f_i - f_j\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$



(Id)

Y. Sun, X. Wang, and X. Tang. NIPS, 2014.

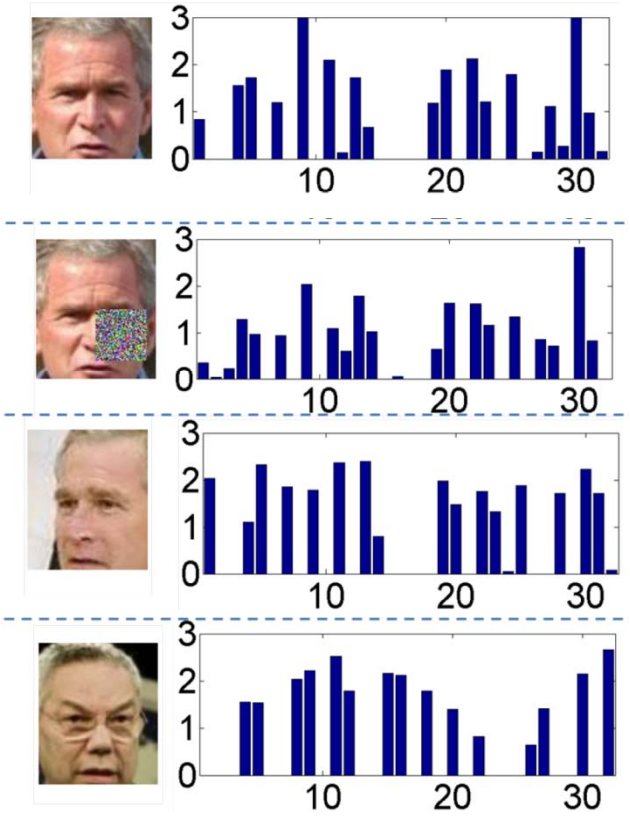
Biological Motivation



- Monkey has a face-processing network that is made of six interconnected face-selective regions
- Neurons in some of these regions were view-specific, while some others were tuned to identity across views
- View could be generalized to other factors, e.g. expressions?

Winrich A. Freiwald and Doris Y. Tsao, "Functional compartmentalization and viewpoint generalization within the macaque face-processing system," *Science,* 330(6005):845–851, 2010.

Deeply learned features are moderately sparse

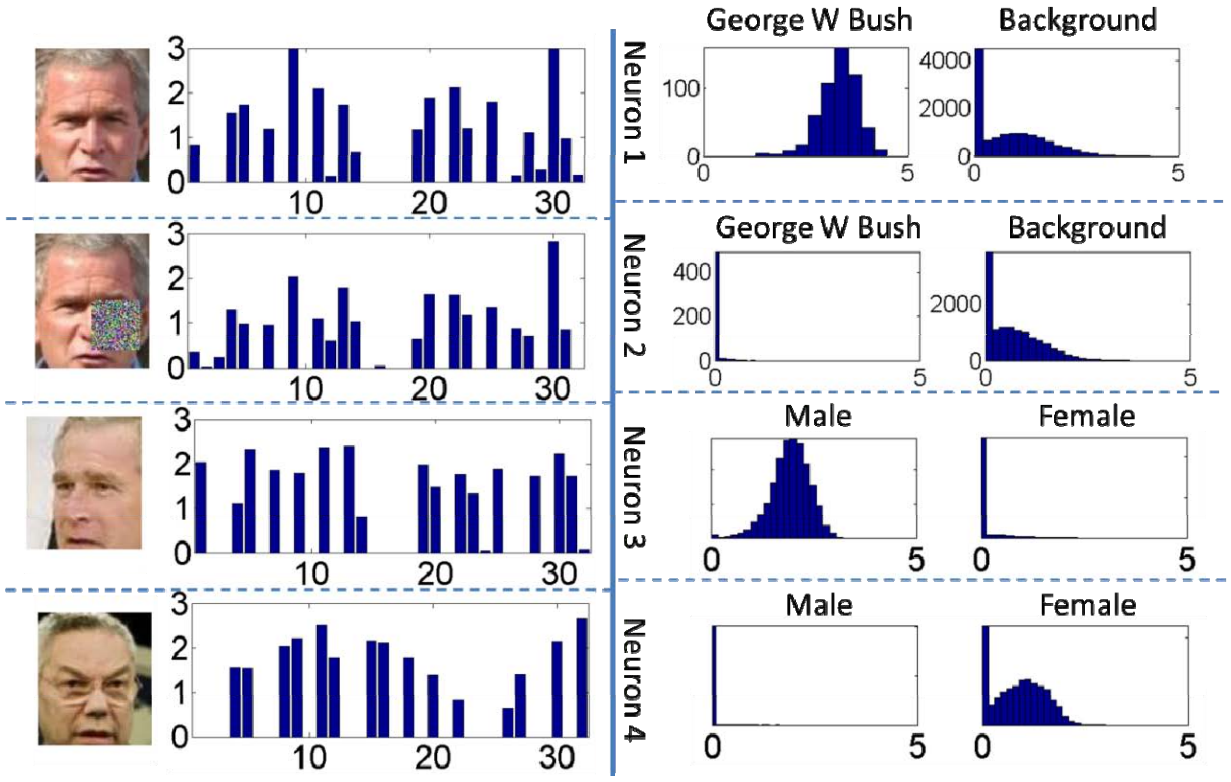


- The **binary codes** on activation patterns are very effective on face recognition
- Save storage and speedup face search dramatically
- Activation patterns are more important than activation magnitudes in face recognition

	Joint Bayesian (%)	Hamming distance (%)
Combined model (real values)	99.47	n/a
Combined model (binary code)	99.12	97.47

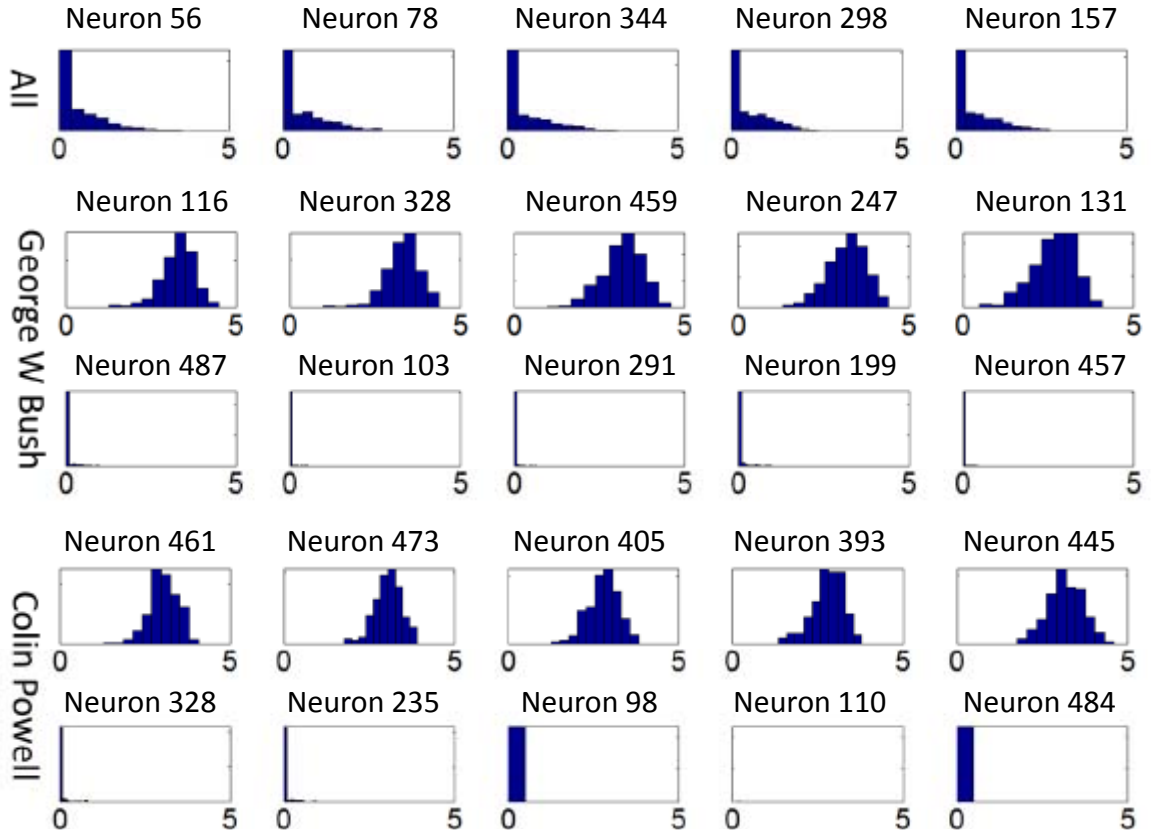
Deeply learned features are selective to identities and attributes

- With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute



Deeply learned features are selective to identities and attributes

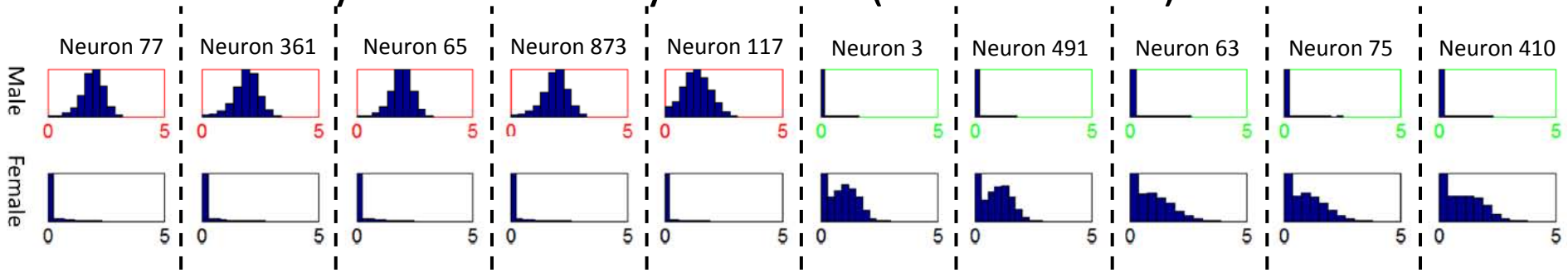
- Excitatory and inhibitory neurons (on identities)



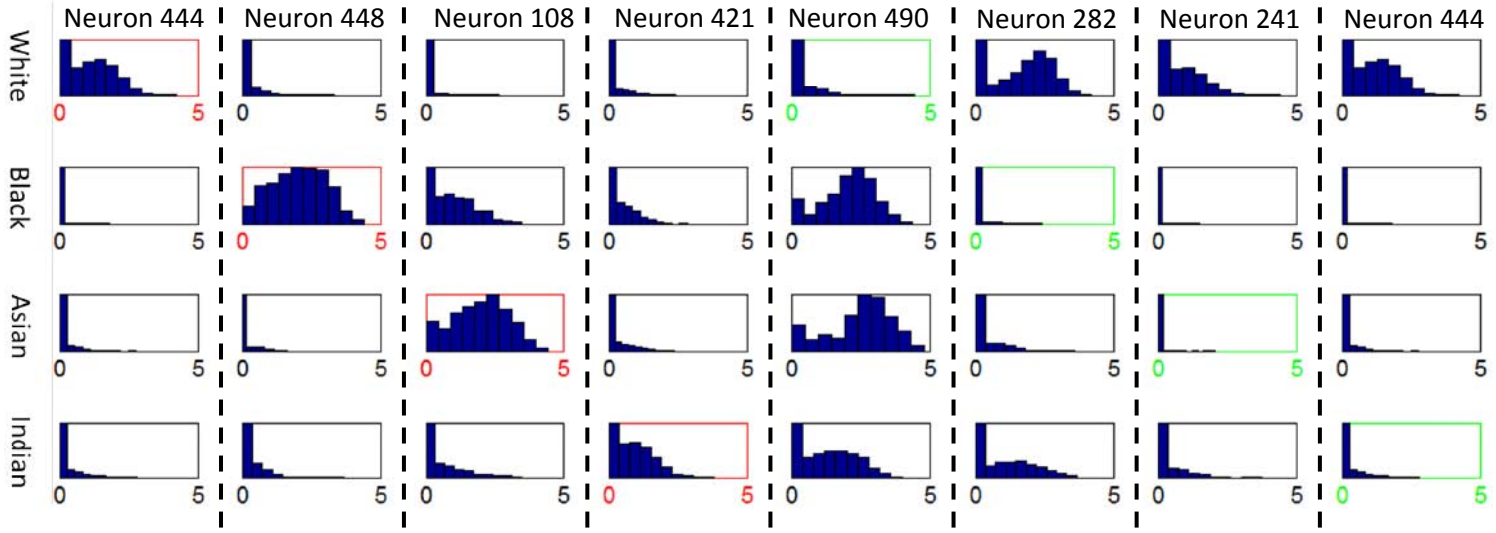
Histograms of neural activations over identities with the most images in LFW

Deeply learned features are selective to identities and attributes

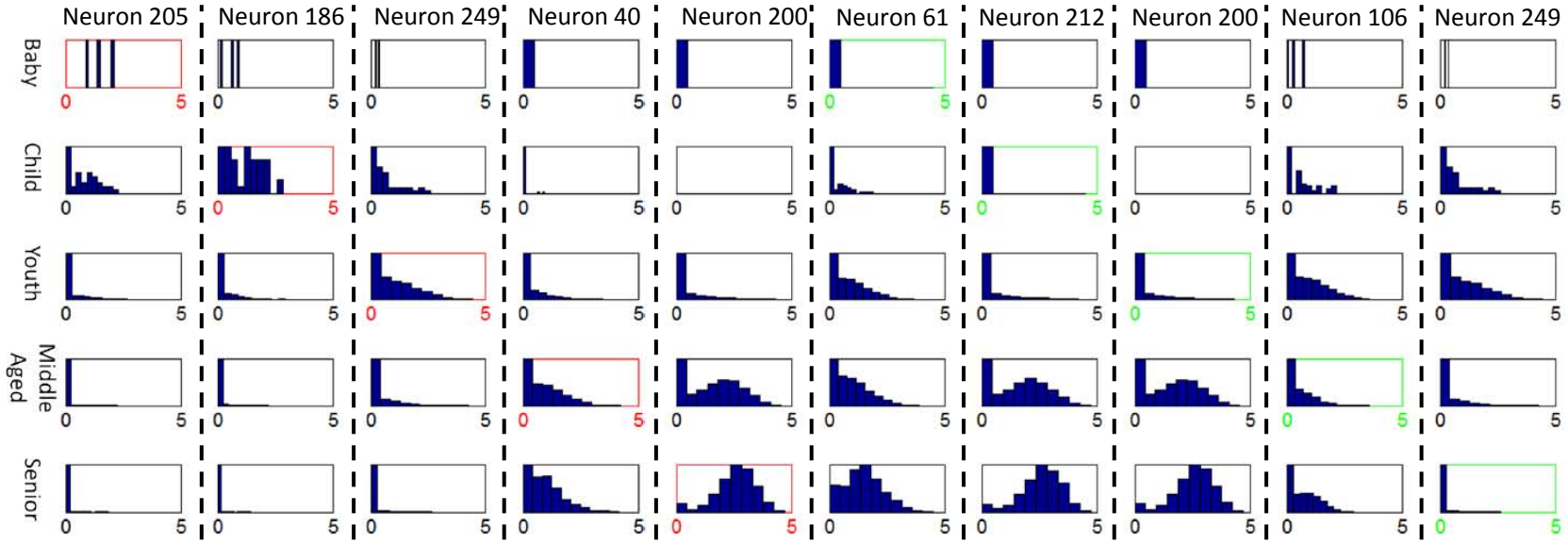
- Excitatory and inhibitory neurons (on attributes)



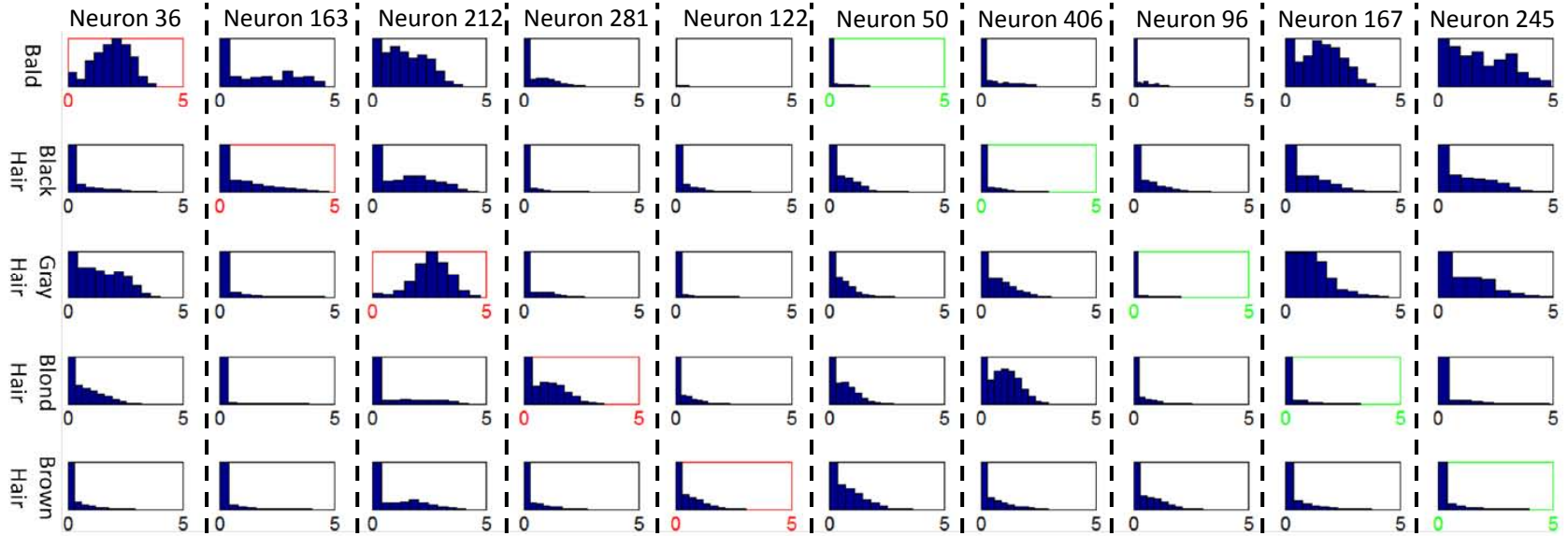
Histograms of neural activations over gender-related attributes (Male and Female)



Histograms of neural activations over race-related attributes (White, Black, Asian and India)



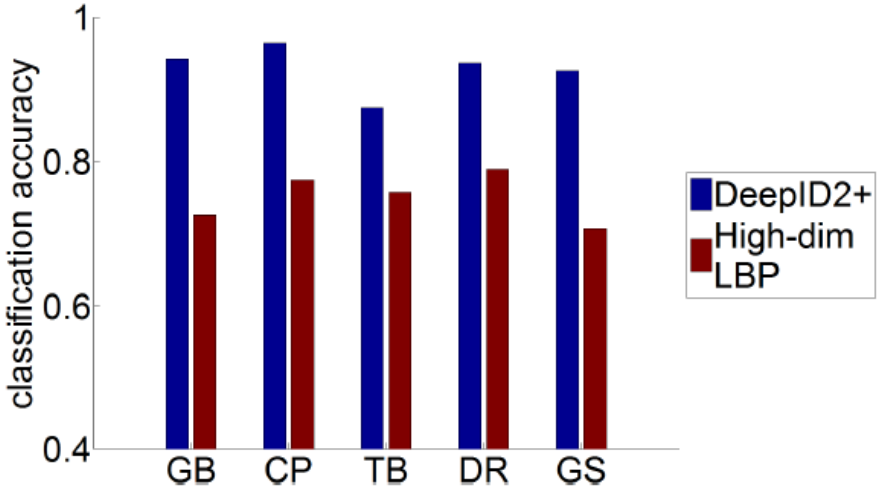
Histogram of neural activations over age-related attributes (Baby, Child, Youth, Middle Aged, and Senior)



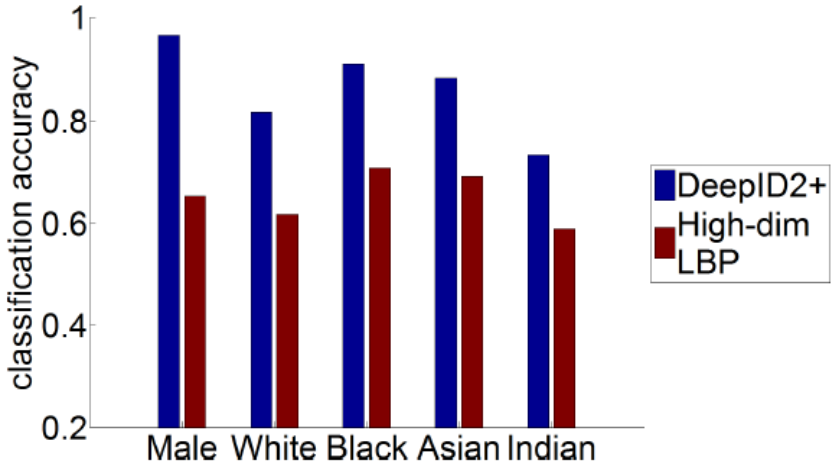
Histogram of neural activations over hair-related attributes (Bald, Black Hair, Gray Hair, Blond Hair, and Brown Hair.

Deeply learned features are selective to identities and attributes

- With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute



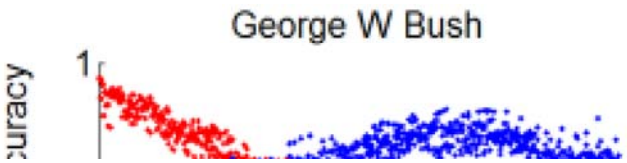
Identity classification accuracy on LFW with one single DeepID2+ or LBP feature. GB, CP, TB, DR, and GS are five celebrities with the most images in LFW.



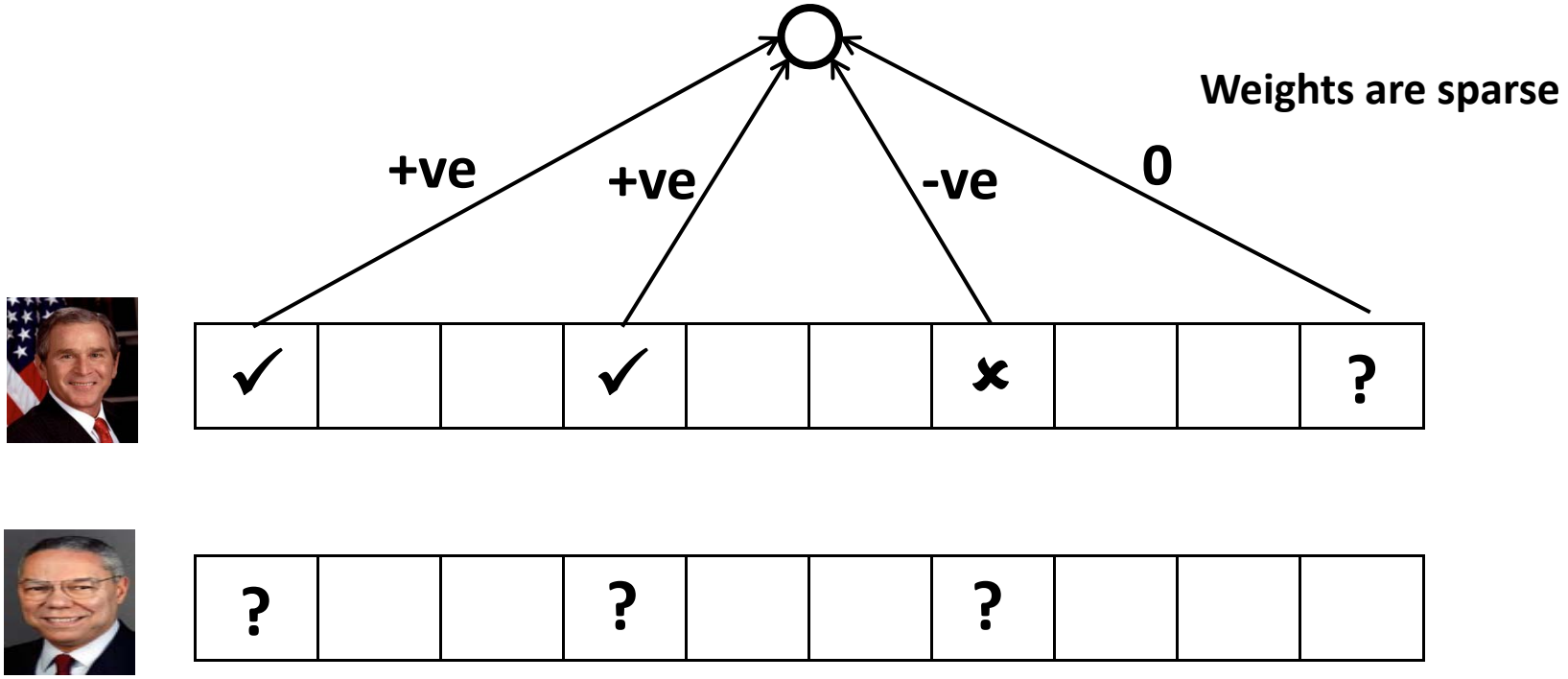
Attribute classification accuracy on LFW with one single DeepID2+ or LBP feature.

Excitatory and Inhibitory neurons

DeepID2+



High-dim LBP

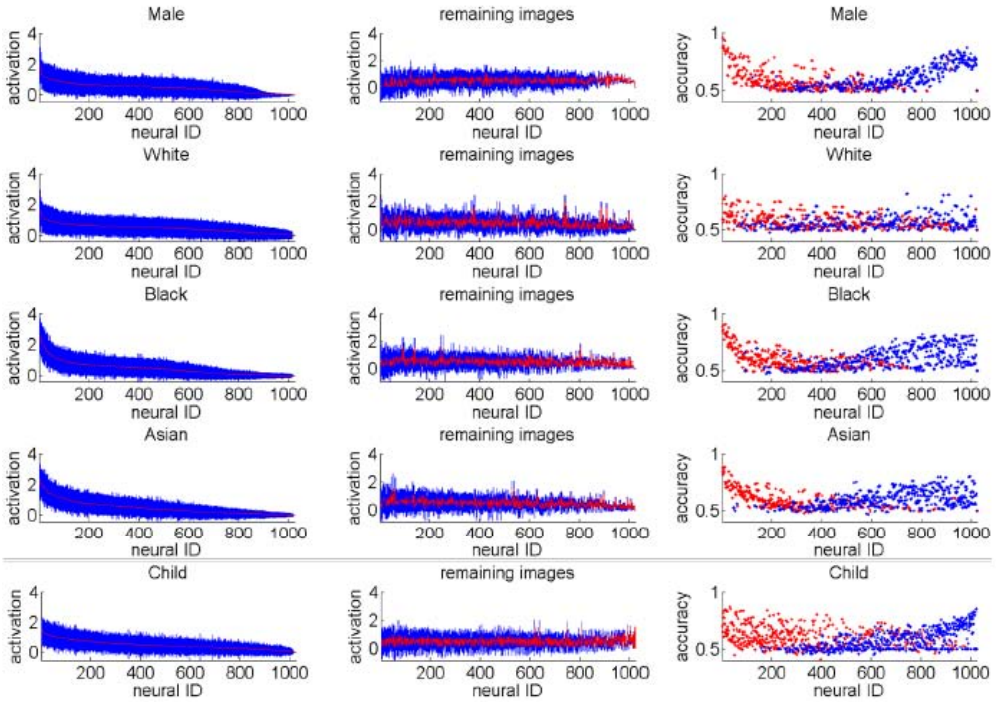


✓ Always respond to a person

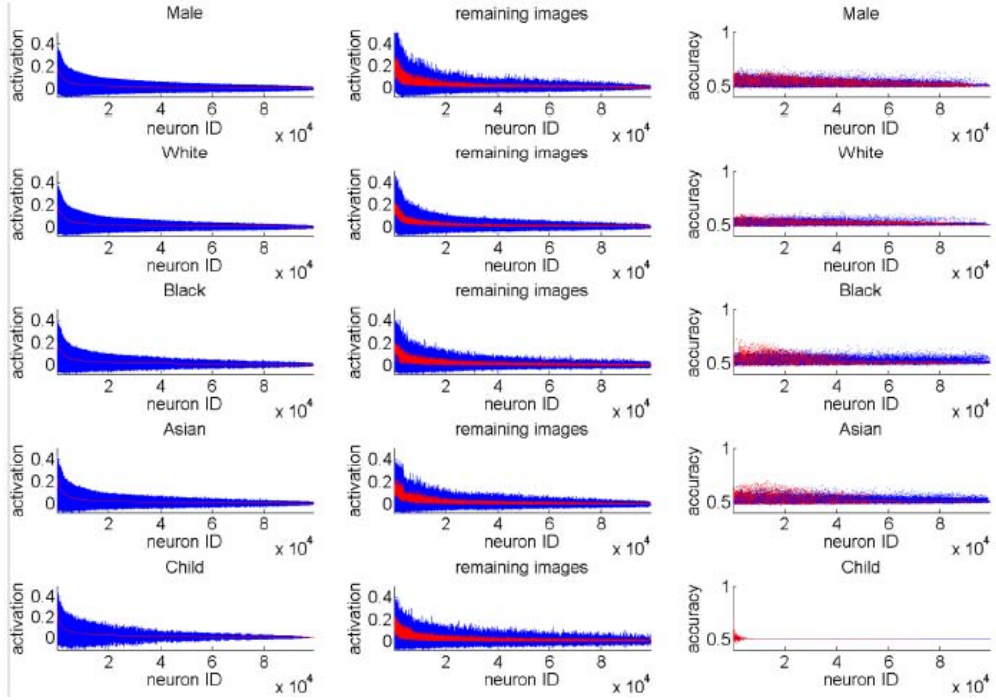
✗ Always no response to a person

? Uncertain

Excitatory and Inhibitory neurons



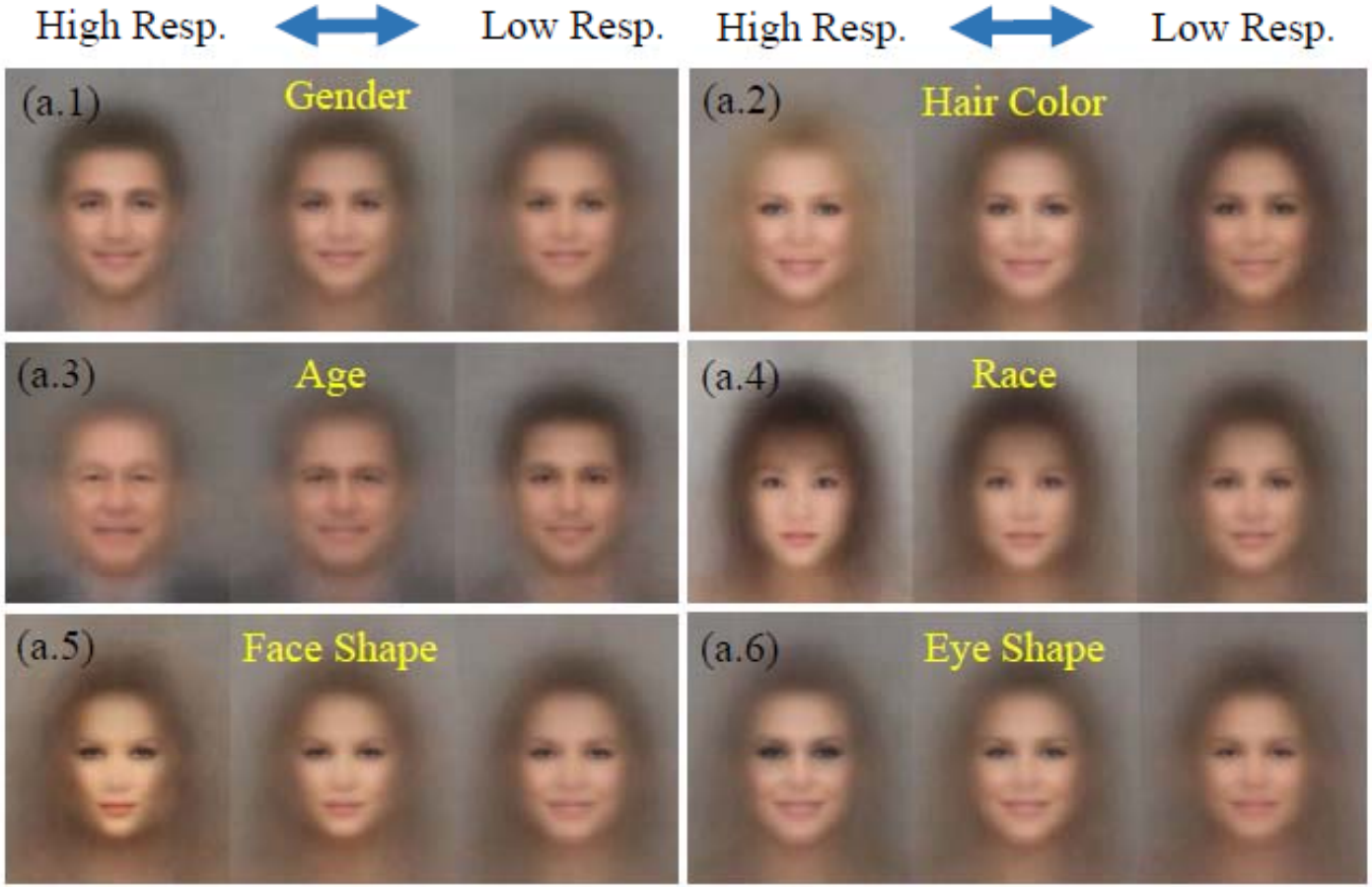
DeepID2+

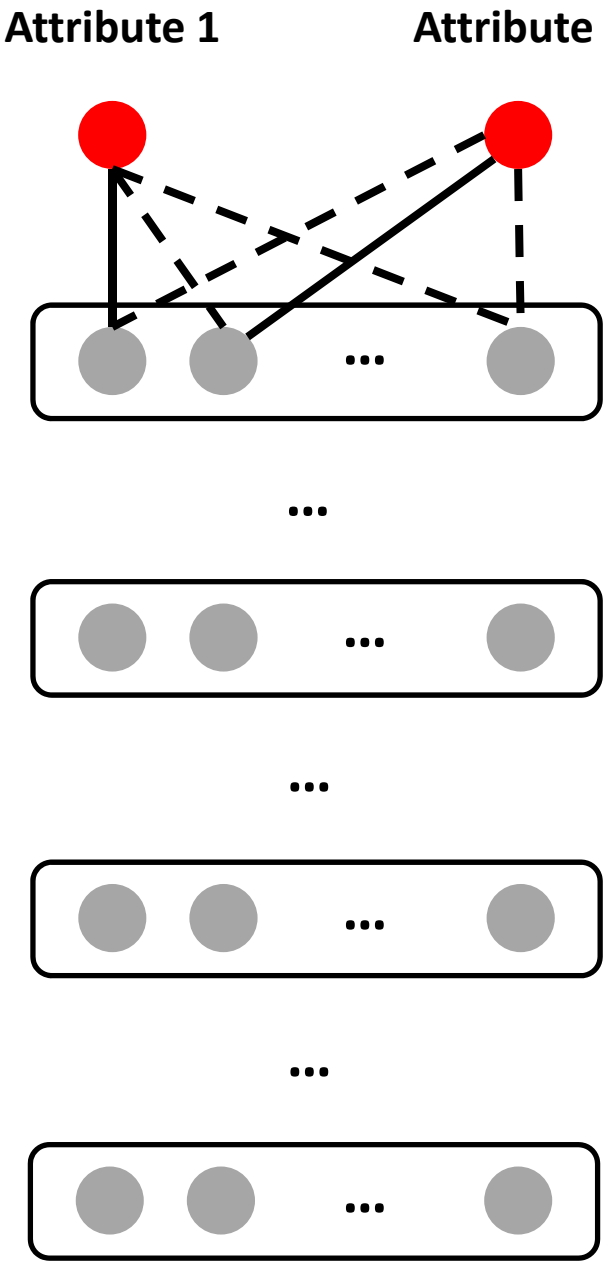


High-dim LBP

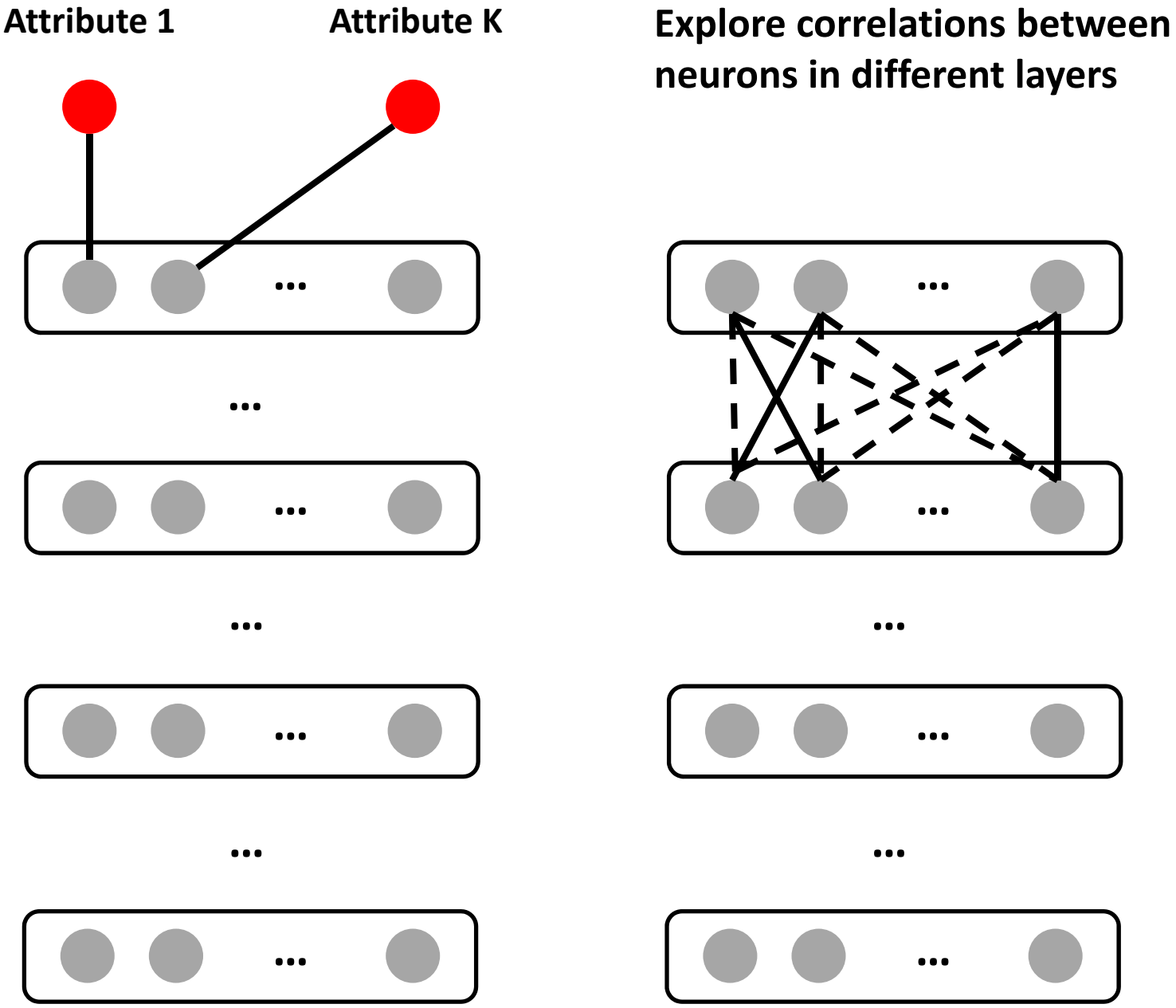
Deeply learned features are selective to identities and attributes

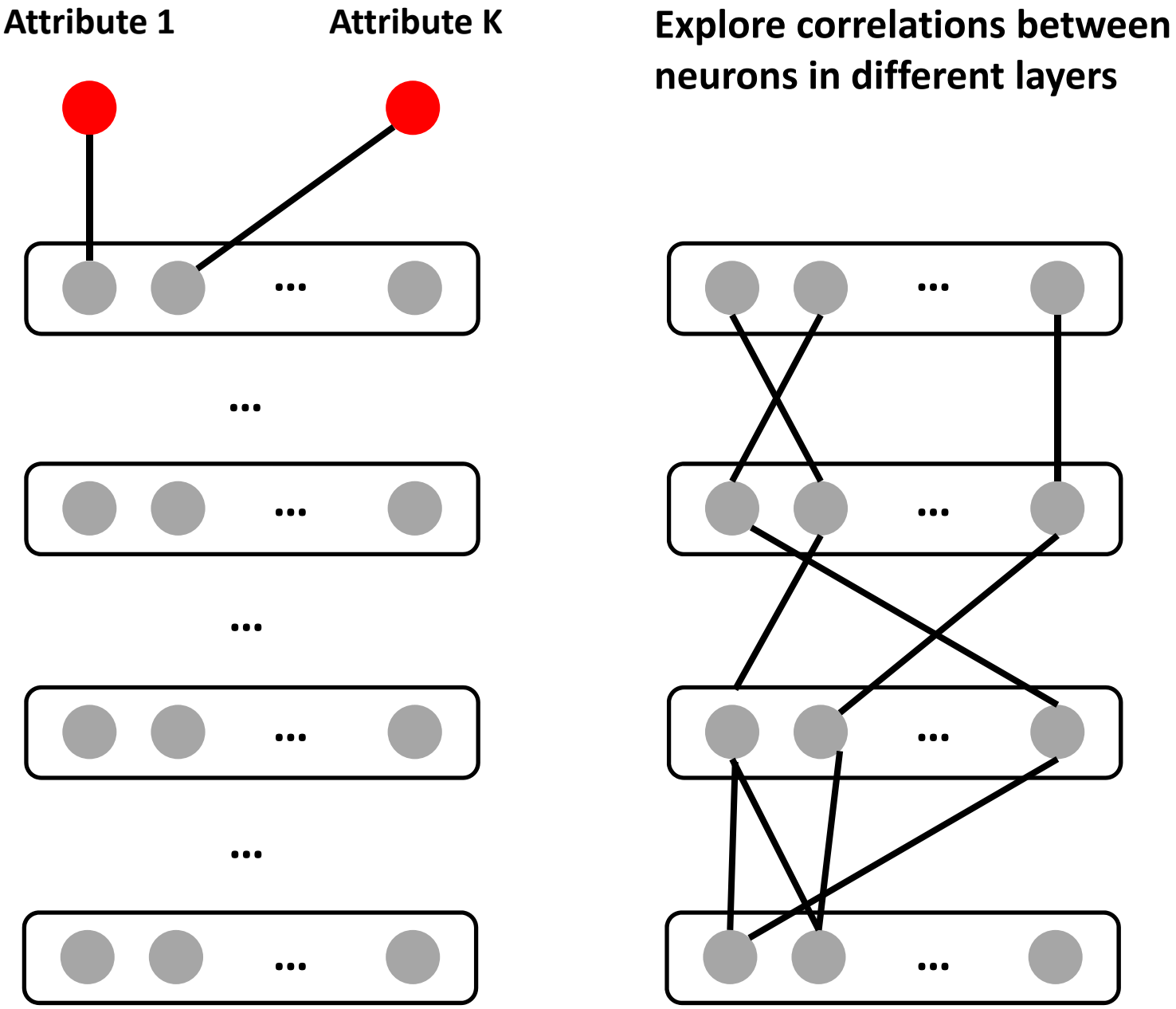
- Visualize the semantic meaning of each neuron



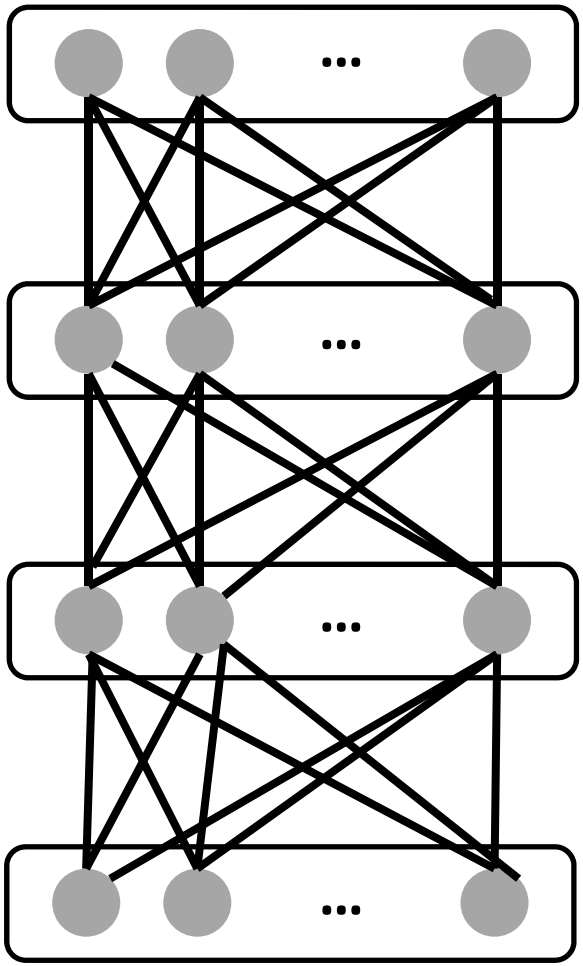


Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Sparsifying Neural Network Connections for Face Recognition,” arXiv:1512.01891, 2015


Attribute 1
Attribute K
Explore correlations between neurons in different layers


Attribute 1
Attribute K
Explore correlations between neurons in different layers

Alternatively learning weights and net structures



1. Train a dense network from scratch
2. Sparsify the top layer, and **re-train** the net
3. Sparsify the second top layer, and **re-train** the net

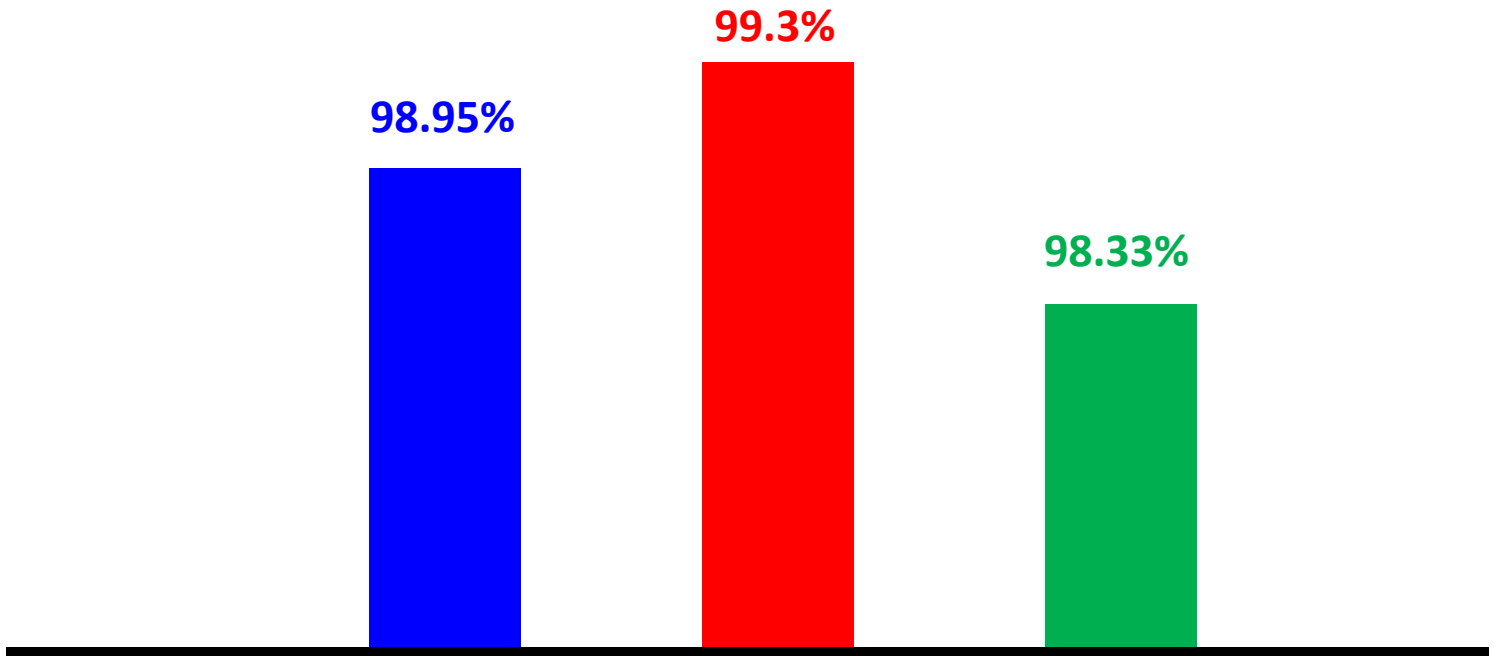
...

Conel, JL. The postnatal development of the human cerebral cortex. Cambridge, Mass: Harvard University Press, 1959.

Original deep neural network

Sparsified deep neural network and only keep 1/8 amount of parameters after joint optimization of weights and structures

Train the sparsified network from scratch



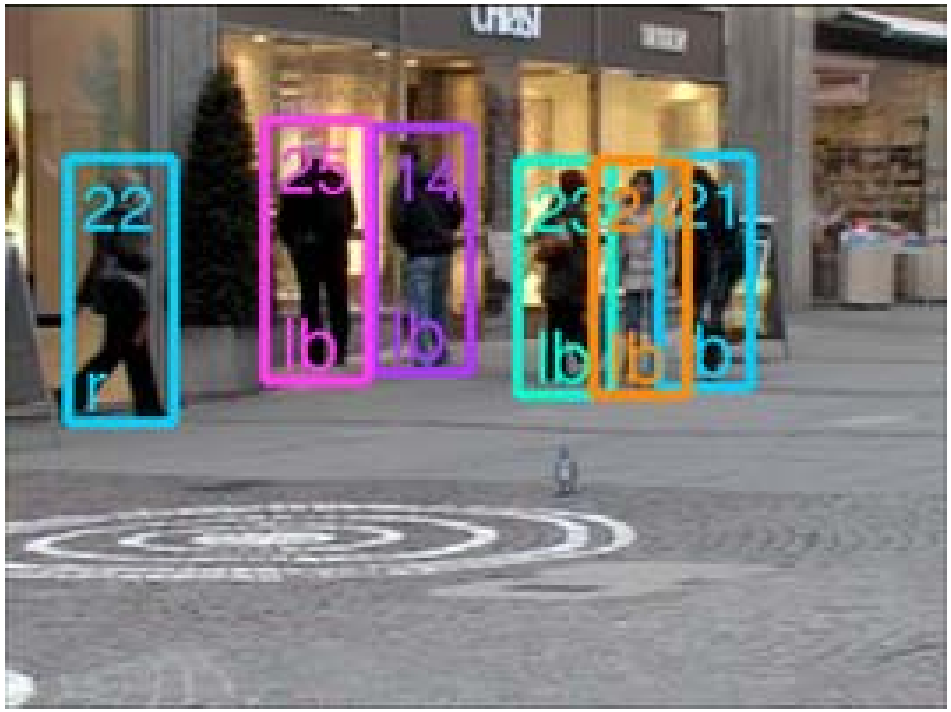
The sparsified network has enough learning capacity, but the original denser network helps it reach a better intialization

Outline

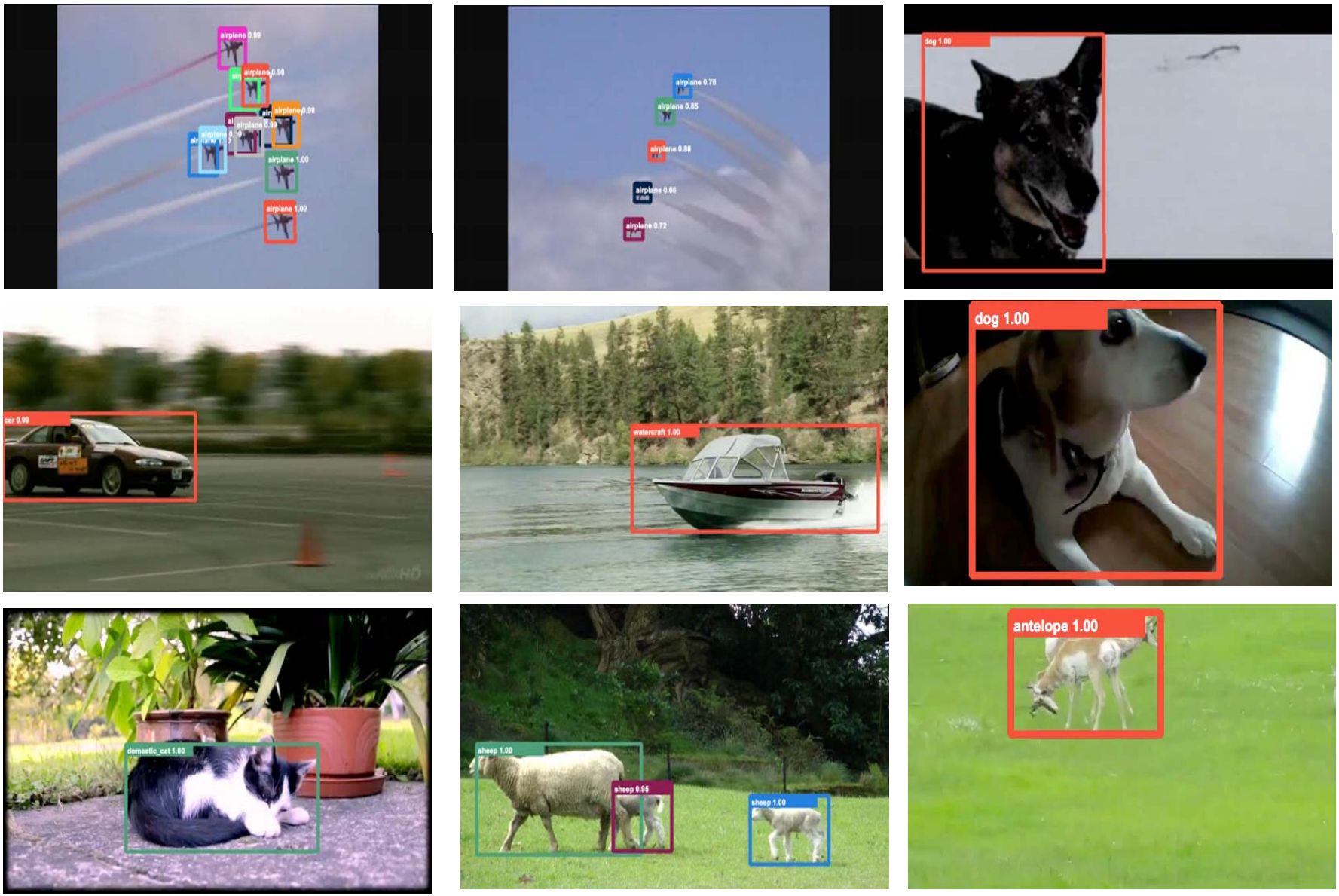
- Face recognition and analysis
- **Object tracking**
- Human pose estimation

Motivations

- Tracking by detection is the state-of-the-art
- How to get detectors for general objects with annotations only in the first frame?

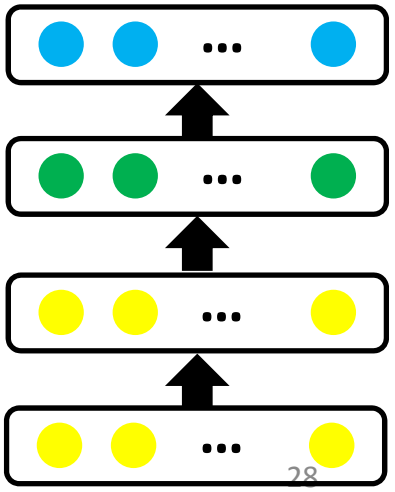


ImageNet Challenge



Motivations

- We observe that for a deep CNN pre-trained on ImageNet, its neurons have strong selectiveness on object categories
- Such CNN provides a large pool of detectors. Its neurons or its subsets of neurons serve as detectors
- The annotation on the first frame can select neurons

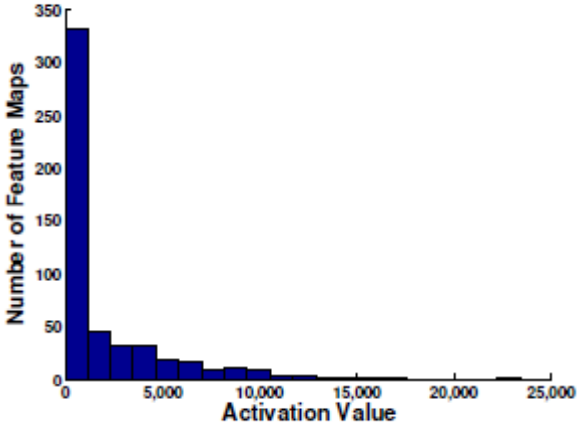


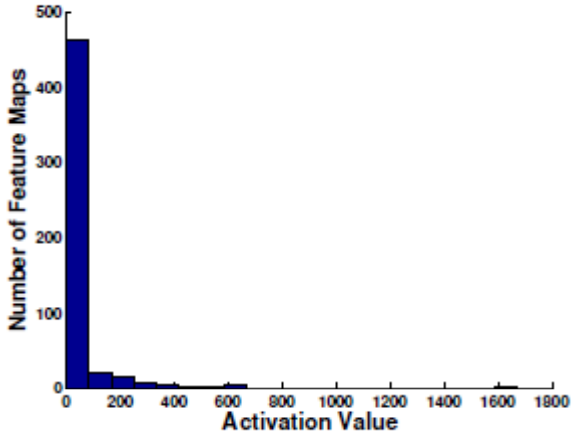
Motivations

- Explore the features pre-trained on massive data and classification task on ImageNet
- A top convolution layer is more robust and encodes more semantic features and serves as a category detector
- A lower convolution layer carries more discriminative information and can better separate the target from distractors with similar appearance
- Both layers are jointly used with a switch mechanism during tracking
- A tracking target, only a subset of neurons are relevant

L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual Tracking with Fully Convolutional Networks," ICCV 2015.

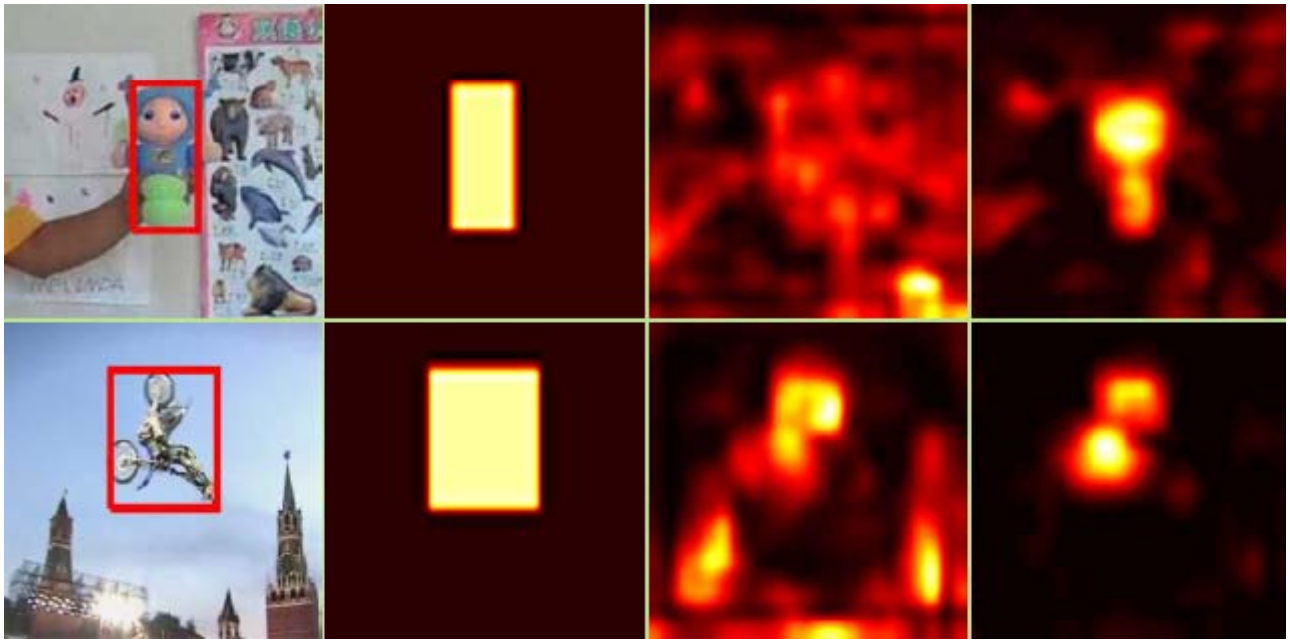
Observation 1: Although the receptive field of CNN feature maps is large, activated feature maps are sparse and localized. Activated regions are highly correlated to the regions of semantic objects





Activation value histograms of feature maps in top (left) and lower (right) layers

Observation 2: Many CNN feature maps are noisy or unrelated for the task of discriminating a particular target from its background

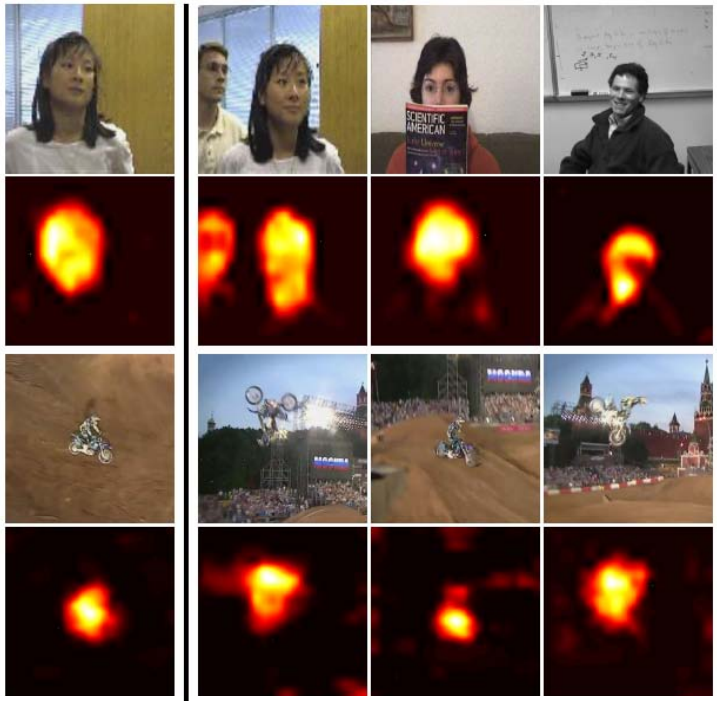


(a) (b) (c)

(a) Ground truth foreground mask, average feature maps of convolution layers; average selected feature maps of convolution layers

Selection of feature maps

- Select feature maps by reconstructing foreground masks and their significance calculated with BP

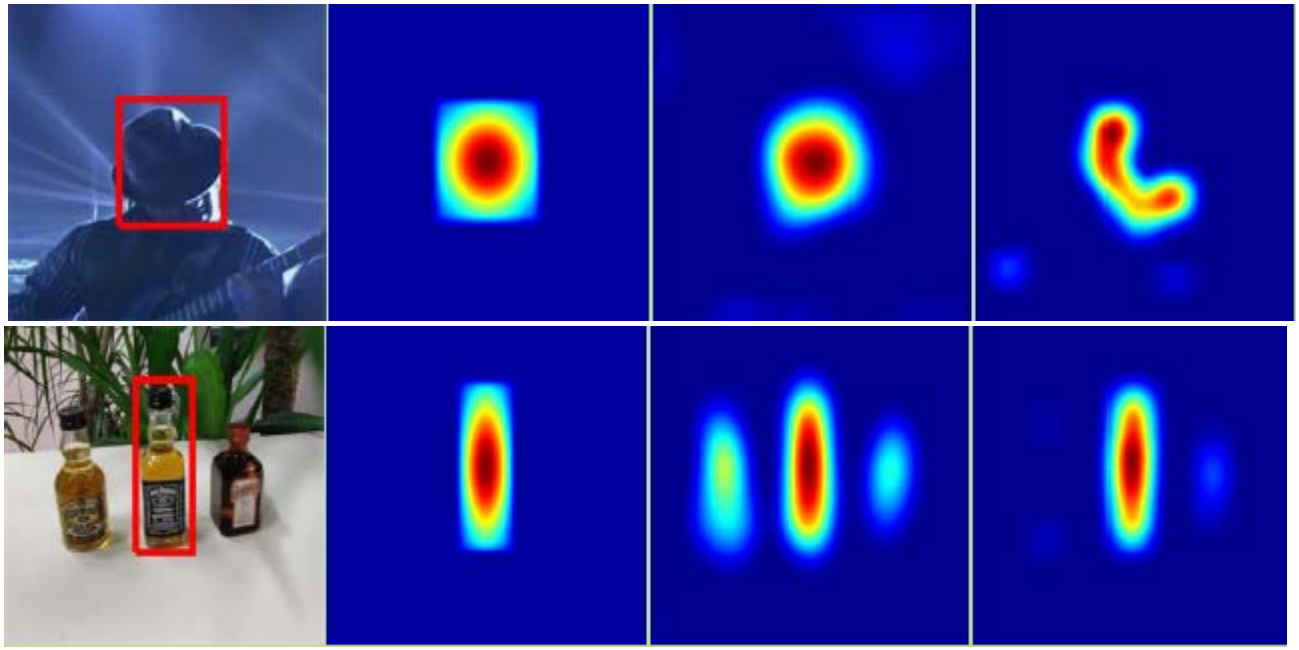


The sparse coefficients are computed using the images in the first column and directly applied to the other columns without change

Section 2:

Performance of Feature Map Selection

Observation 3: Higher layers capture semantic concepts on object categories, whereas lower layers encode more discriminative features to capture intra class variations

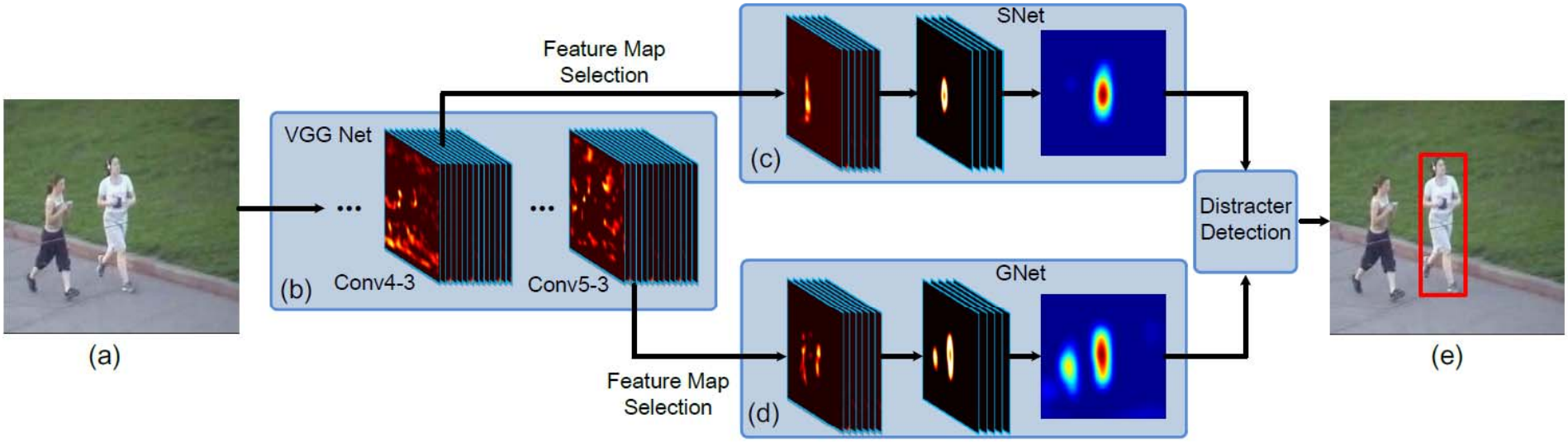


(a) (b) (c)

(a) Ground truth target heat map; (b) Predicted heat maps using feature maps of top convolution layers of VGG; (c) Predicted heat maps using feature maps of lower convolution layers of VGG

Fully convolutional network based tracker (FCN)

- GNet: capture the category information of the target and is built on the top layers of VGG
- SNet: discriminate the target from background with similar appearance and is built on the lower layers of VGG

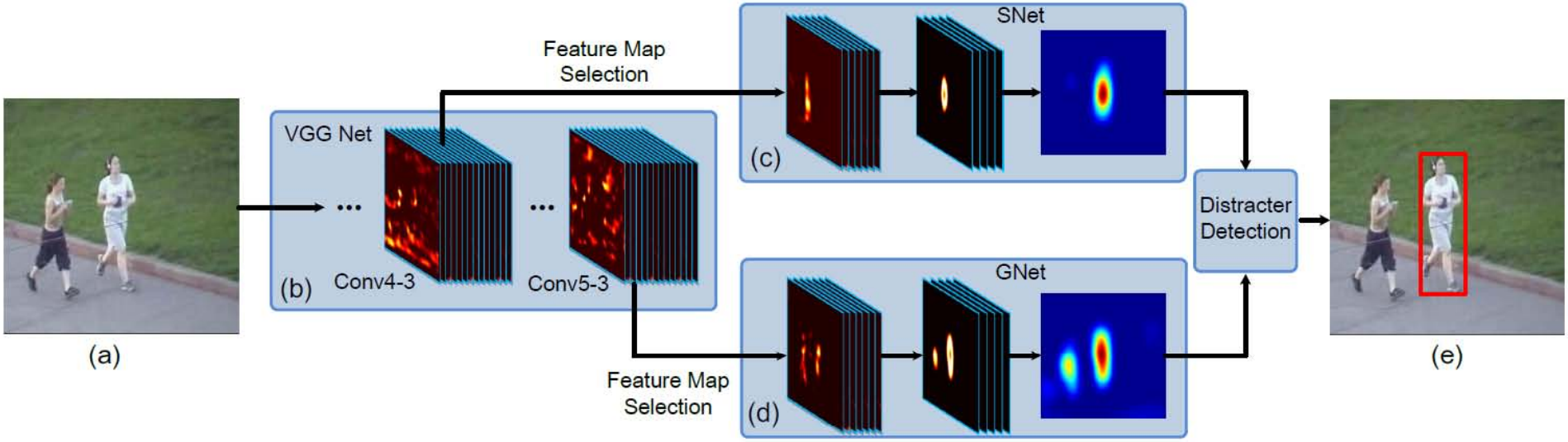


(b) VGG network; (c) SNet; (d) Gnet; (e) Tracking results

Both GNet and SNet are initialized in the first frame to perform foreground heat map regression for the target: GNet is fixed and SNet is updated every 200 frames

SNet is used if the background distractor is larger than a threshold; otherwise GNet is used

For a new frame, a region of interest (ROI) centered at the last target location containing both target and background context is cropped and propagated through the fully convolutional network

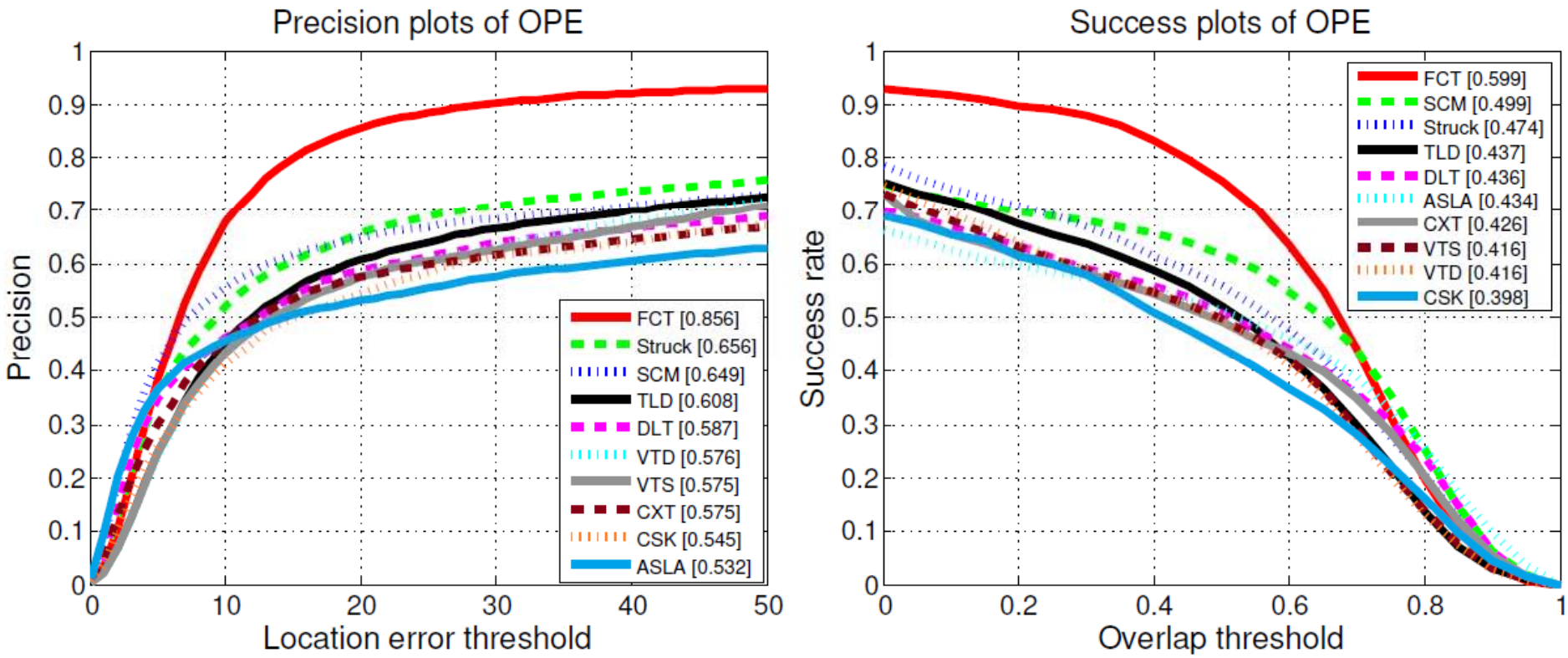


(b) VGG network; (c) SNet; (d) Gnet; (e) Tracking results

Section 1:

Performance of SNet and GNet

Precision plots and success plots of OPE for the top 10 trackers

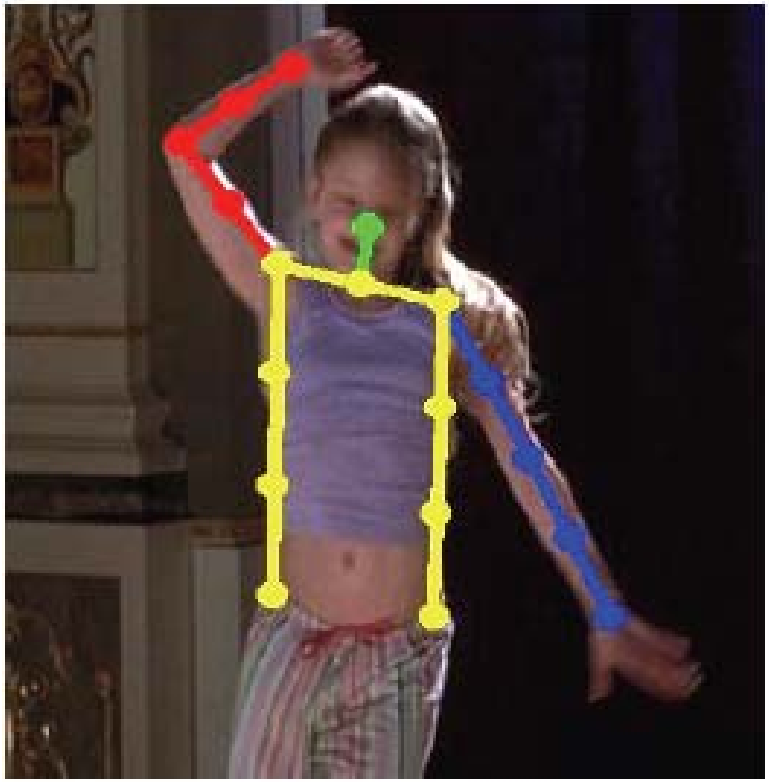


Section 3:
Comparison Resutls

Outline

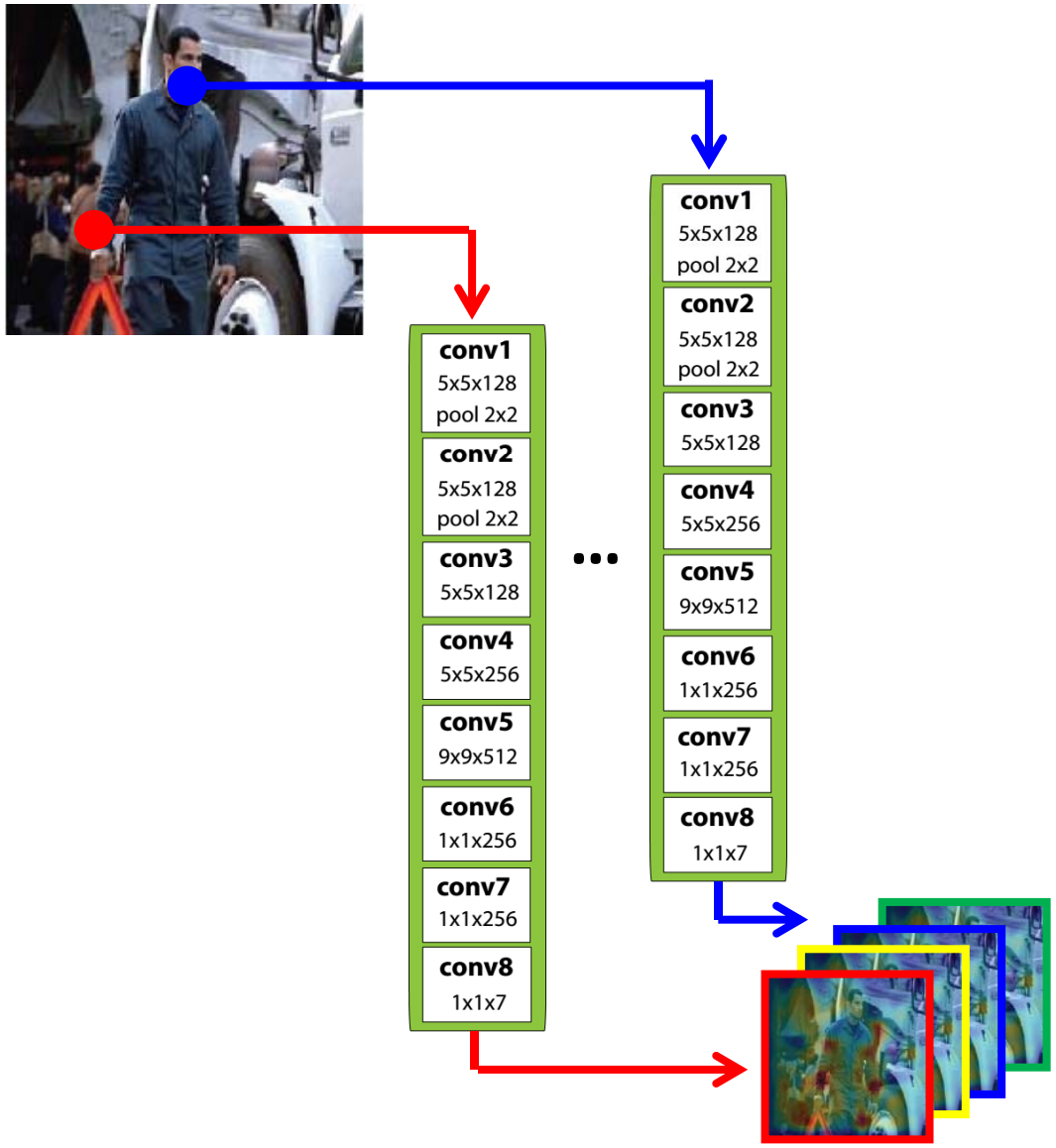
- Face recognition and analysis
- Object tracking
- **Human pose estimation**

Human Pose Estimation

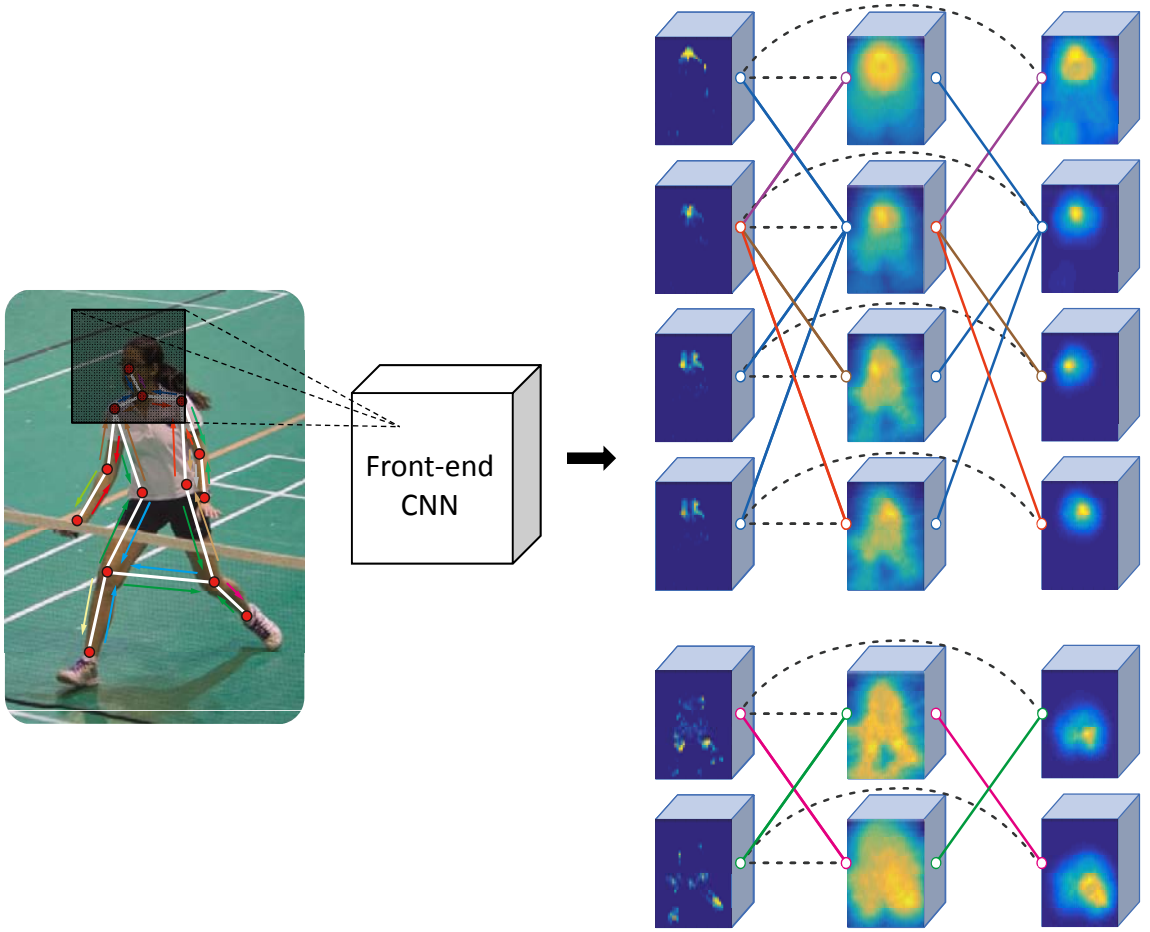


Pose estimation result generated by our deep learning algorithm

Using CNN to localize individual joints separately is not reliable

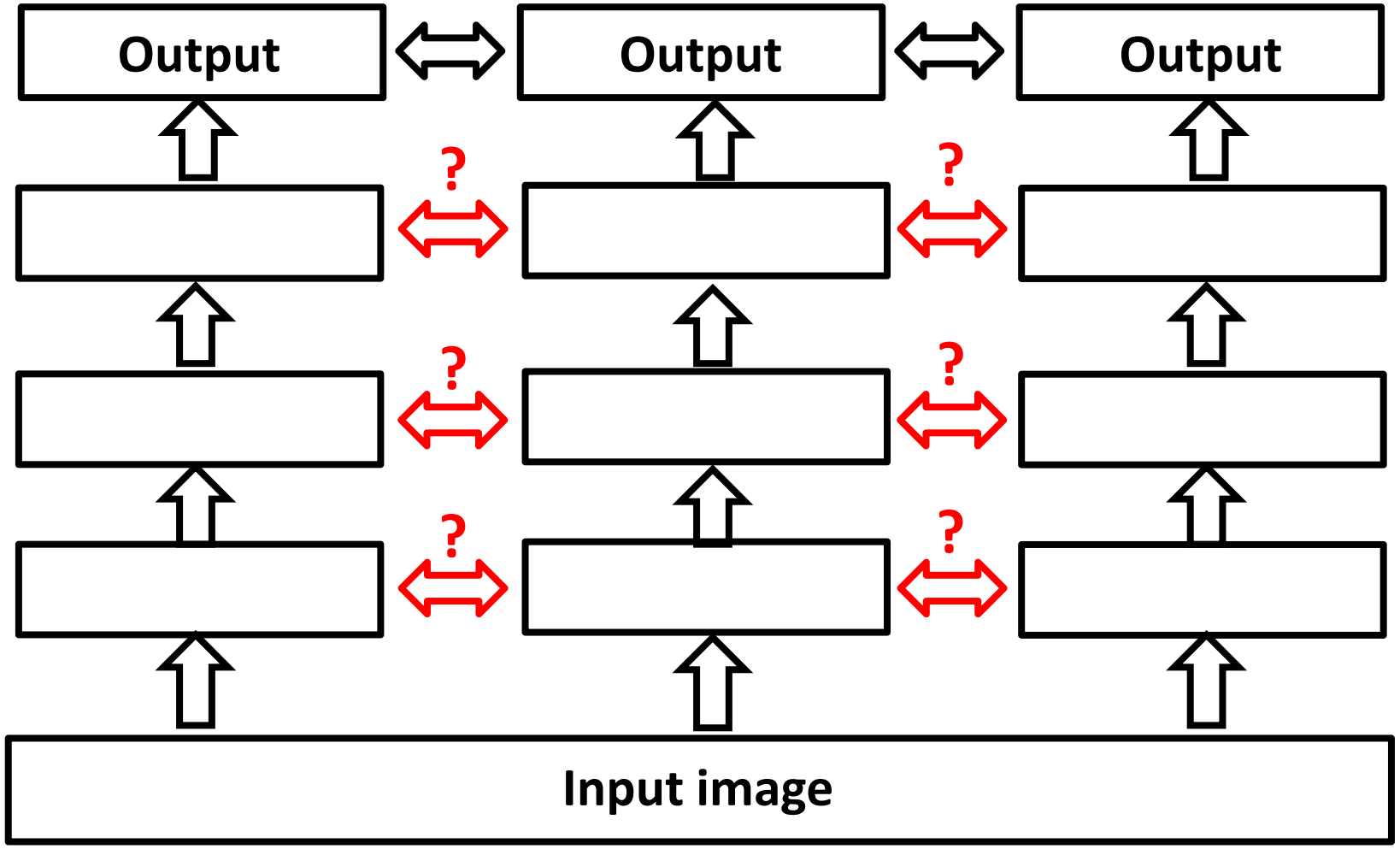


Model structures on score maps or predicted labels (with much information loss)



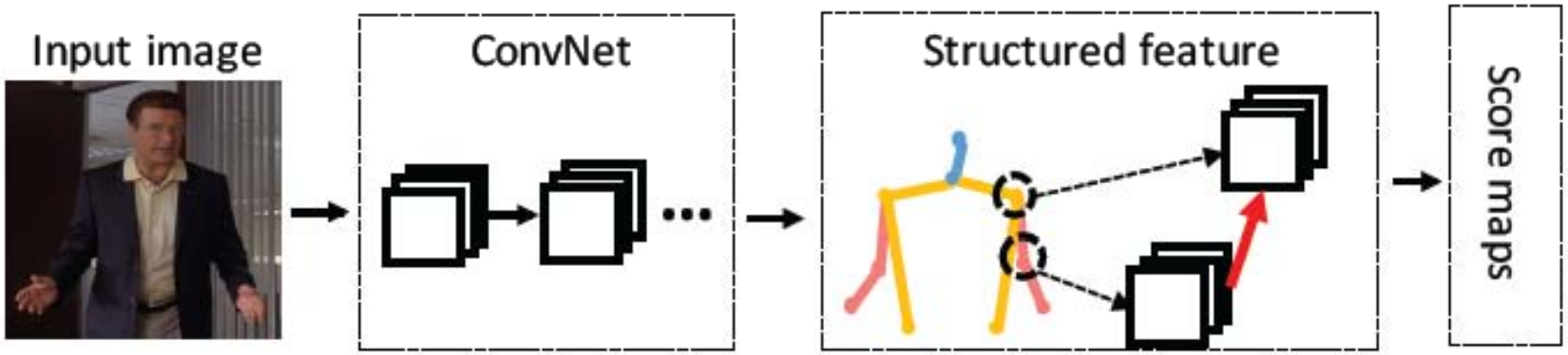
Message passing on score maps

Model interaction between neurons in the same layer?



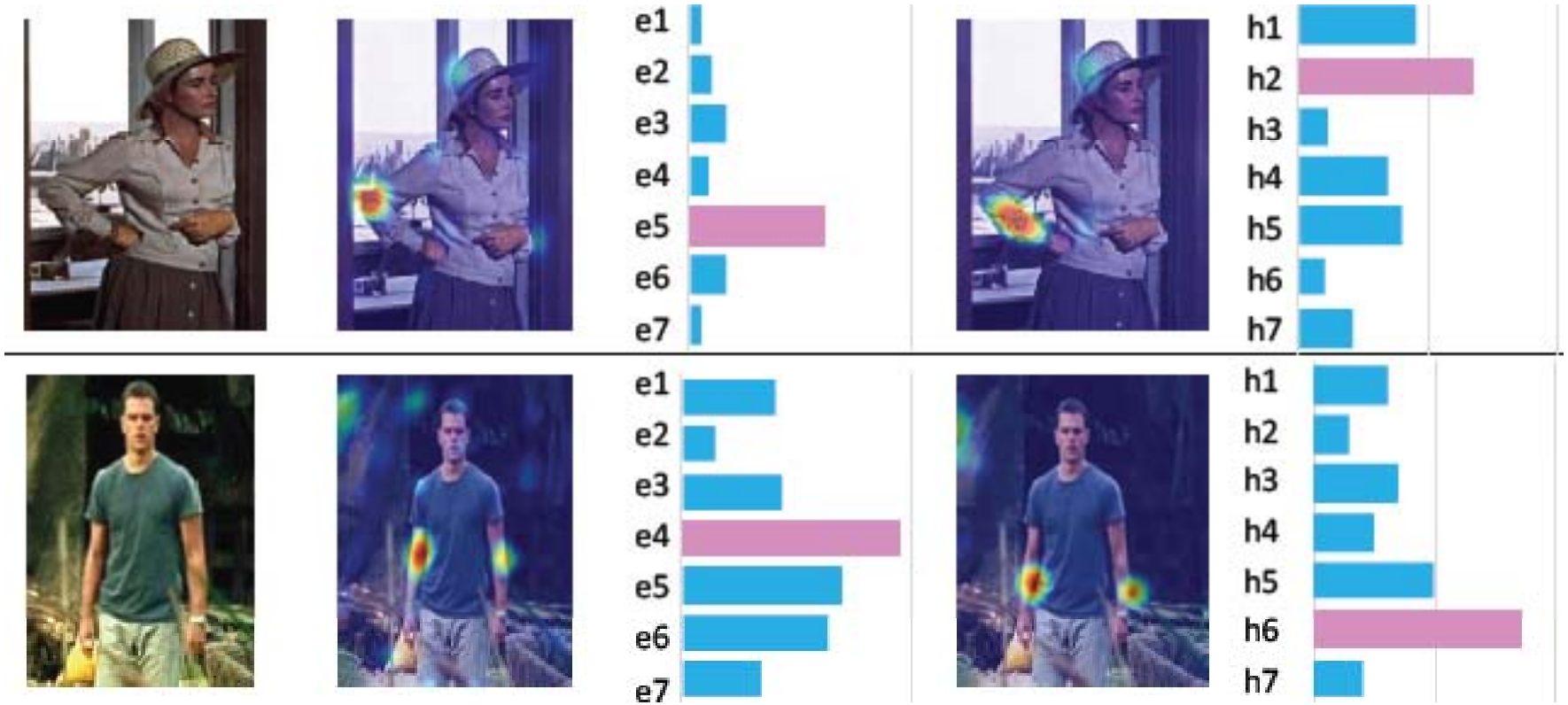
Structured Feature Learning

- Rich information is preserved at feature map level
- Reason the correlations among body joints at the feature level

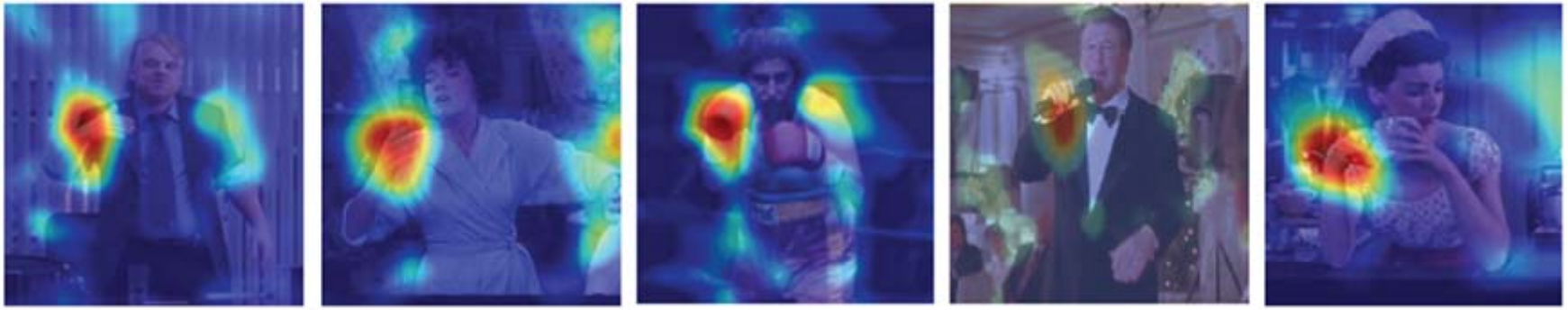


X. Chu, W. Ouyang, W. Yang, and X. Wang, "Structured Feature Learning for Pose Estimation," CVPR 2016.

- **Understand the semantic meanings of feature maps**



High responding images for channel 1 for neck

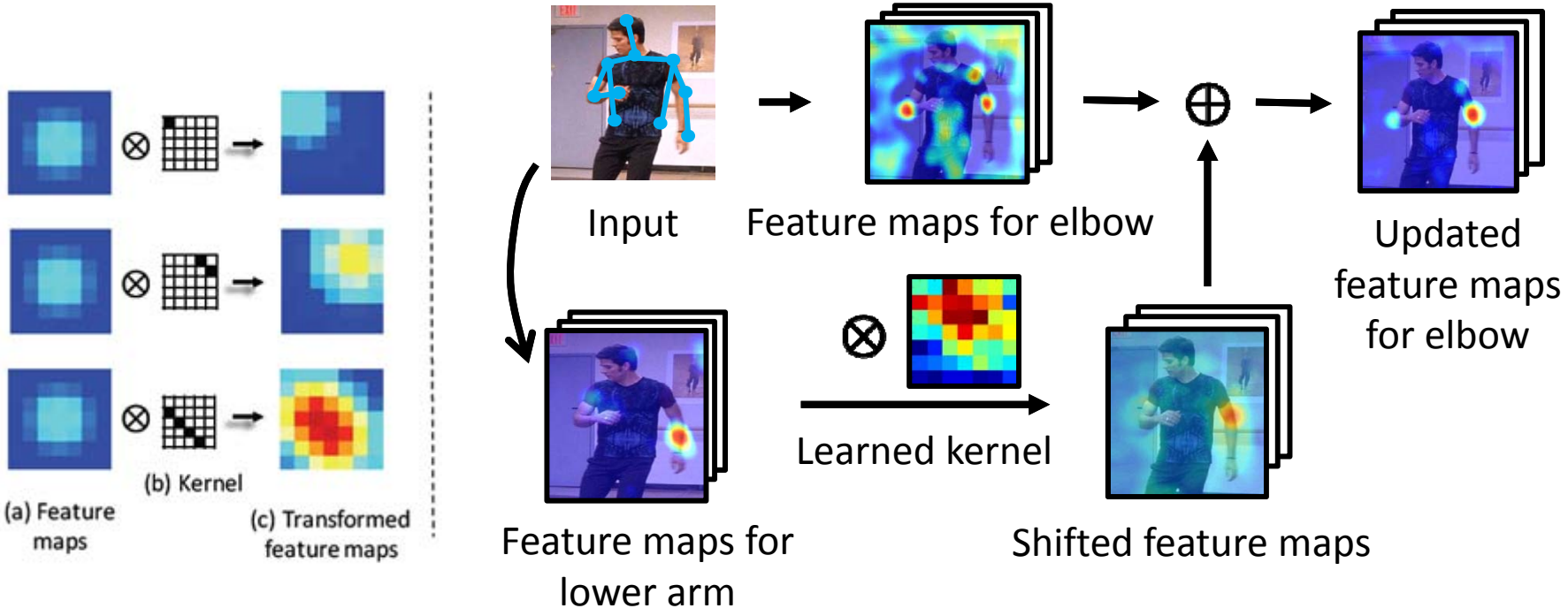


High responding images for channel 2 for left shoulder

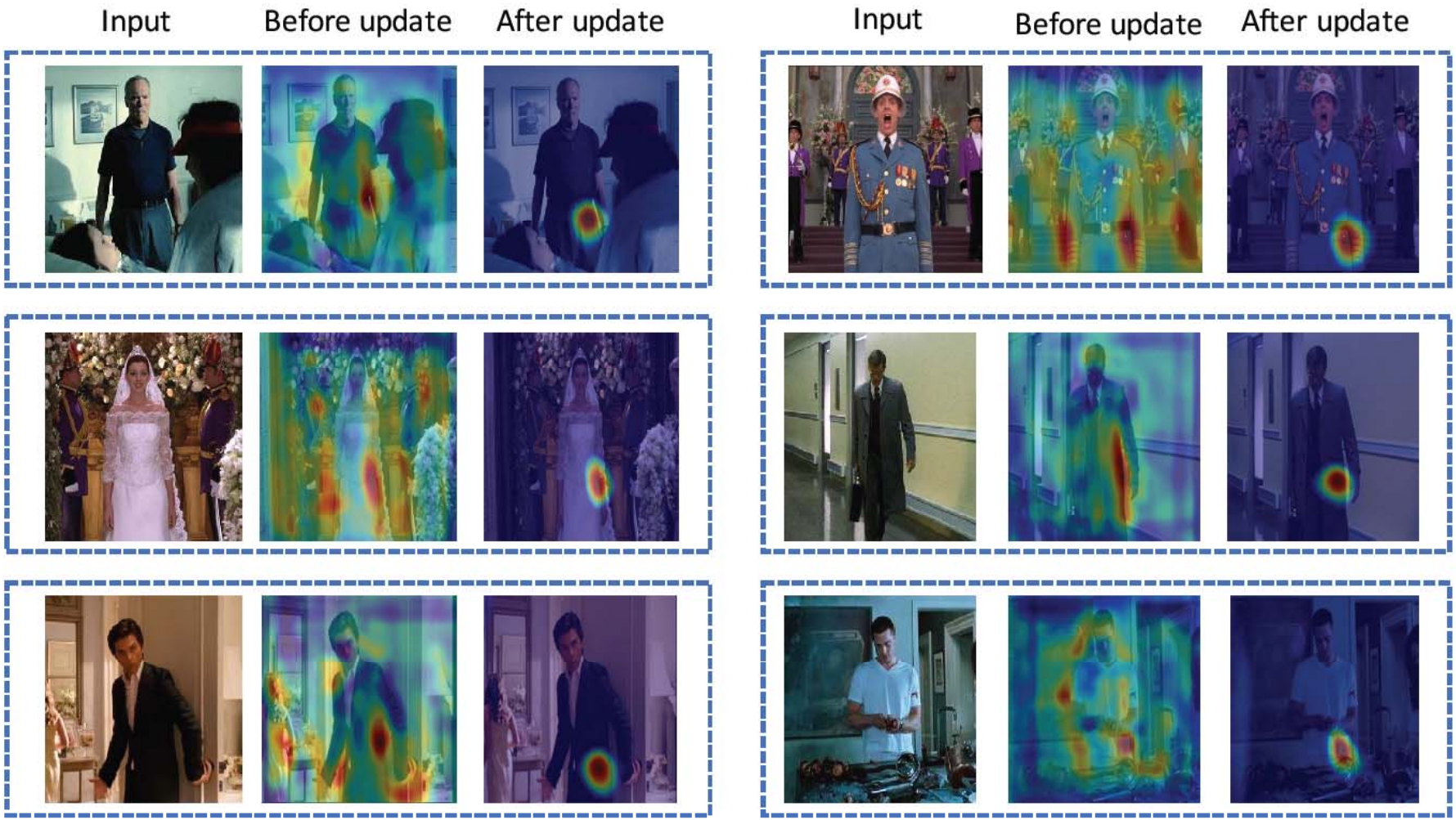
High responding images for channel 3 for lower arm

Geometrical Transform Kernels

- Pass information through convolution between feature maps and geometrical transform kernels



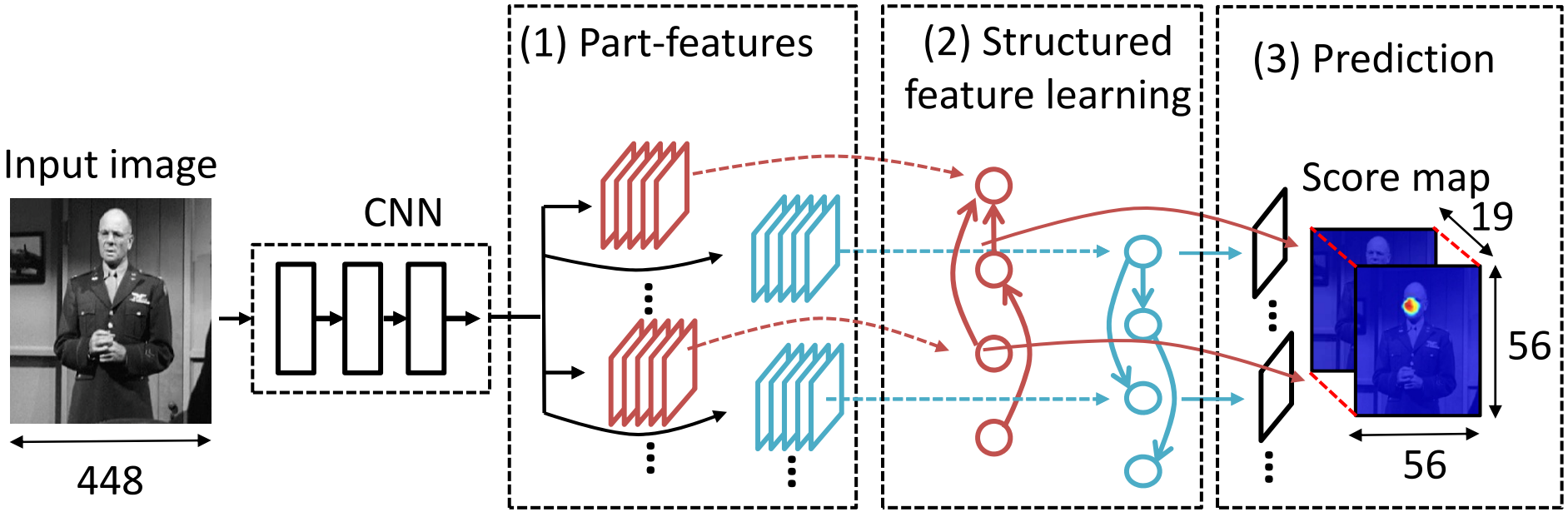
Feature map update --- Torso

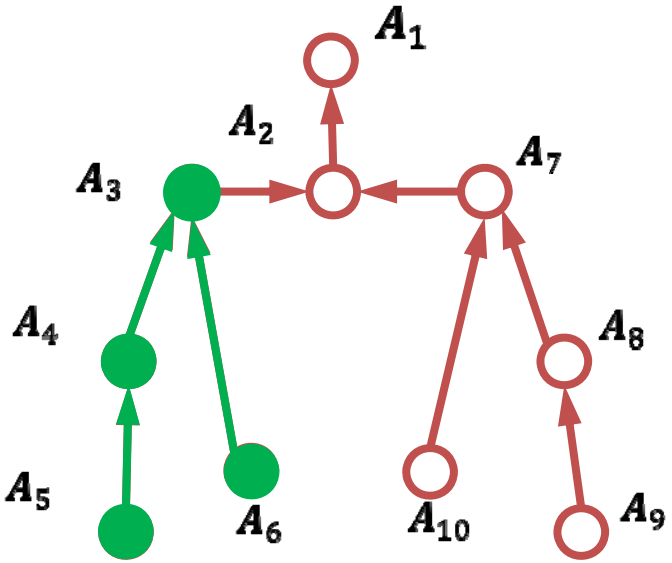


Feature map update --- Shoulder

Bidirectional Tree

- Fully connected graph is not a good solution
  - Large transform kernels are required to model joints in distance
  - Relationship between some joints are unstable
- Propagate information through intermediate joints on a designed graph
- On a bi-directional tree, feature channels at a joint well receives information from other joints


Input image
448
CNN
(1) Part-features
(2) Structured feature learning
(3) Prediction
Score map
19
56
56



$$A_6 = f(h_{fcn6} \otimes w^{a6}) \qquad A_6' = A_6$$

$$A_5 = f(h_{fcn6} \otimes w^{a5}) \qquad A_5' = A_5$$

$$A_4 = f(h_{fcn6} \otimes w^{a4}) \qquad A_4' = f(A_4 + A_5' \otimes w^{a5,a4})$$

$$A_3 = f(h_{fcn6} \otimes w^{a3}) \qquad A_3' = f(A_3 + A_4' \otimes w^{a4,a3} + A_6' \otimes w^{a6,a3})$$

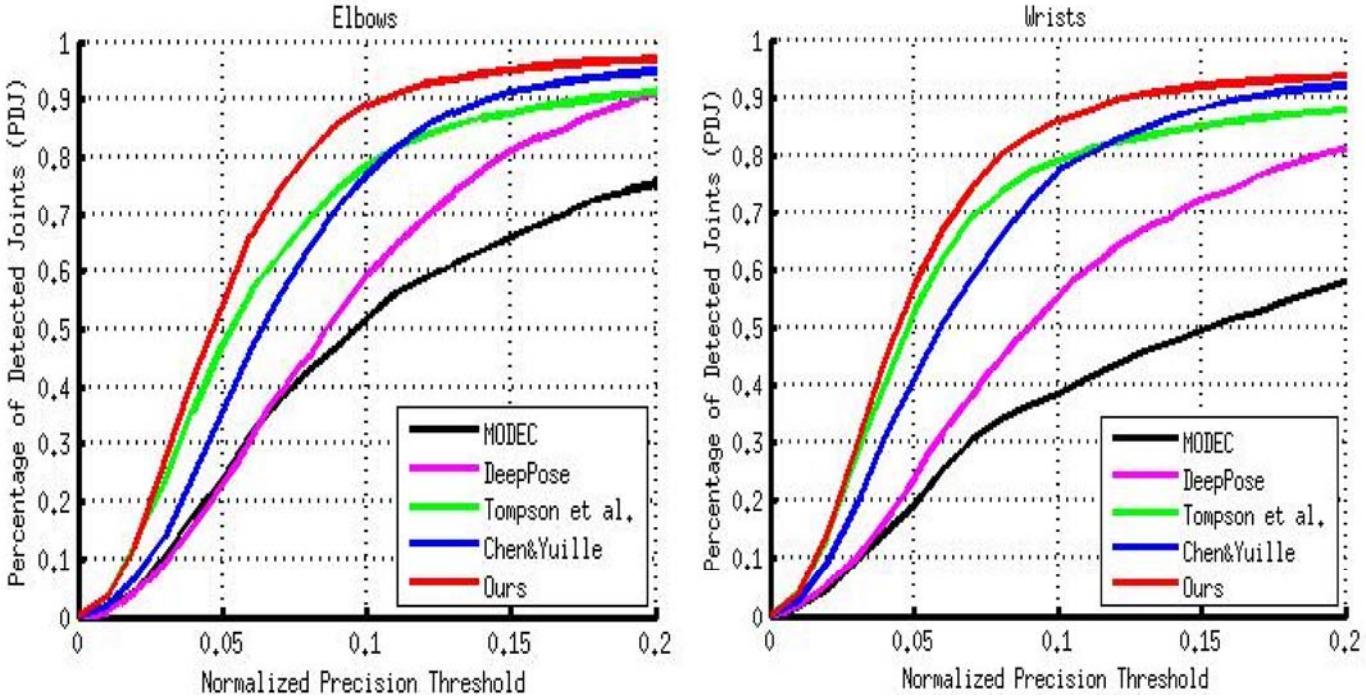
Fully connected graph is not a suitable structure for our method

Fully connected graph: feature maps for shoulder which collect information directly form all the other joints.

Tree graph: feature maps for shoulder which collect information directly from upper arm and indirectly from elbow, lower arm and wrist.

Results(FLIC dataset)

Experiment	Head	Torso	U.arms	L.arms	Mean
MODEC [25]	-	-	84.4	52.1	68.3
Tompson *et al.* [31]	-	-	93.7	80.9	87.3
Chen& Yuille [7]	-	-	97.0	86.8	91.9
Ours	98.6	93.9	97.9	92.4	95.2



Results(LSP dataset)

Experiment	Torso	Head	U.arms	L.arms	U.legs	L.legs	Mean
Andriluka *et al.* [2]	80.9	74.9	46.5	26.4	67.1	60.7	55.7
Yang&Ramanan [37]	82.9	79.3	56.0	39.8	70.3	67.0	62.8
Pishchulin *et al.* [22]	87.5	78.1	54.2	33.9	75.7	68.0	62.9
Eichner&Ferrari *et al.* [10]	86.2	80.1	56.5	37.4	74.3	69.3	64.3
Ouyang *et al.* [18]	85.8	83.1	63.3	46.6	76.5	72.2	68.6
Pishchulin *et al.* [23]	88.7	85.1	61.8	45.0	78.9	73.2	69.2
Chen&Yuille [7]	92.7	87.8	69.2	55.4	82.9	77.0	75.0
Ours	**95.4**	**89.4**	**76.0**	**64.3**	**87.6**	**83.5**	**80.8**

Conclusions

- The success of deep learning is not to be simply understood as using a large number of parameters to fit data. The neural responses have semantic interpretation, i.e. selectiveness on classes and object instances
- Such semantic interpretation has a wide range of applications, including sparsifying networks, object tracking, and human pose estimation
- Understanding neural semantics help to develop new net architectures and training strategies

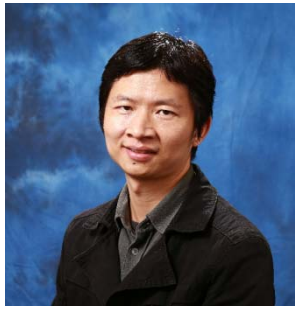
Yi Sun

Ziwei Liu

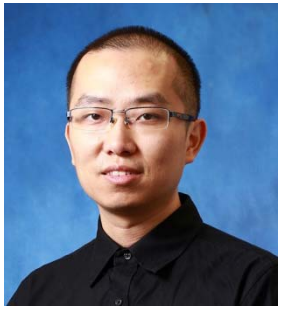
Lijun Wang

Xiao Chu

Jing Shao



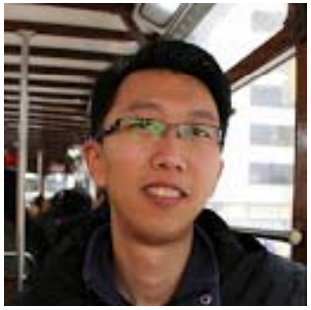
Wanli Ouyang



Hongsheng Li

Xiaogang Wang

Xiaoou Tang



Chen-Change Loy

Huchuan Lu