E9: 309 ADL 09-12-2020

http://leap.ee.iisc.ac.in/sriram/teaching/ADL2020/

Housekeeping

Desirable Topics that belong Module-II

Desirable Choose a domain

Midterm project II - Abstract submission deadline 14/12/2020

Presentation deadline - Dec. 29th, 30th (time will be announced)

GAN - Reading

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

- Creswell, Antonia, et al. "Generative adversarial networks: An overview." *IEEE Signal Processing Magazine* 35.1 (2018): 53-65.

https://jonathan-hui.medium.com/proof-gan-optimal-point-658116a236fb

Topics thus far …

Visual and Time Series Modeling: Semantic Models, Recurrent neural models and LSTM models, Encoder-decoder models, Attention models.

Representation Learning, Causality And Explainability: t-SNE visualization, Hierarchical Representation, semantic embeddings, gradient and perturbation analysis, Topics in Explainable learning, Structural causal models.

Unsupervised Learning: Restricted Boltzmann Machines, Variational Autoencoders, Generative Adversarial Networks.

New Architectures: Capsule networks, End-to-end models, Transformer Networks.

Applications: Applications in in NLP, Speech, Image/Video domains in all modules.

Need of explainable and interpretable learning

- Deep networks are large complex networks (black boxes)
  - Often give the performance needed for commercialization
  - But may fail to provide meaningful insight to the choice of decisions
- Significant need for explainability if technology has to be used in
  - healthcare (life and death situations)
  - banking and credit
  - forensics and law

Today's lecture

- Analyzing trained neural networks
  - ✓ Hierachical representations
  - ✓ Transferable representations

Empirical

Maximizing activations

Visualizing Higher-Layer Features of a Deep Network

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent
Dept. IRO, Université de Montréal
P.O. Box 6128, Downtown Branch, Montreal, H3C 3J7, QC, Canada
first.last@umontreal.ca

Technical Report 1341
Département d'Informatique et Recherche Opérationnelle

2009

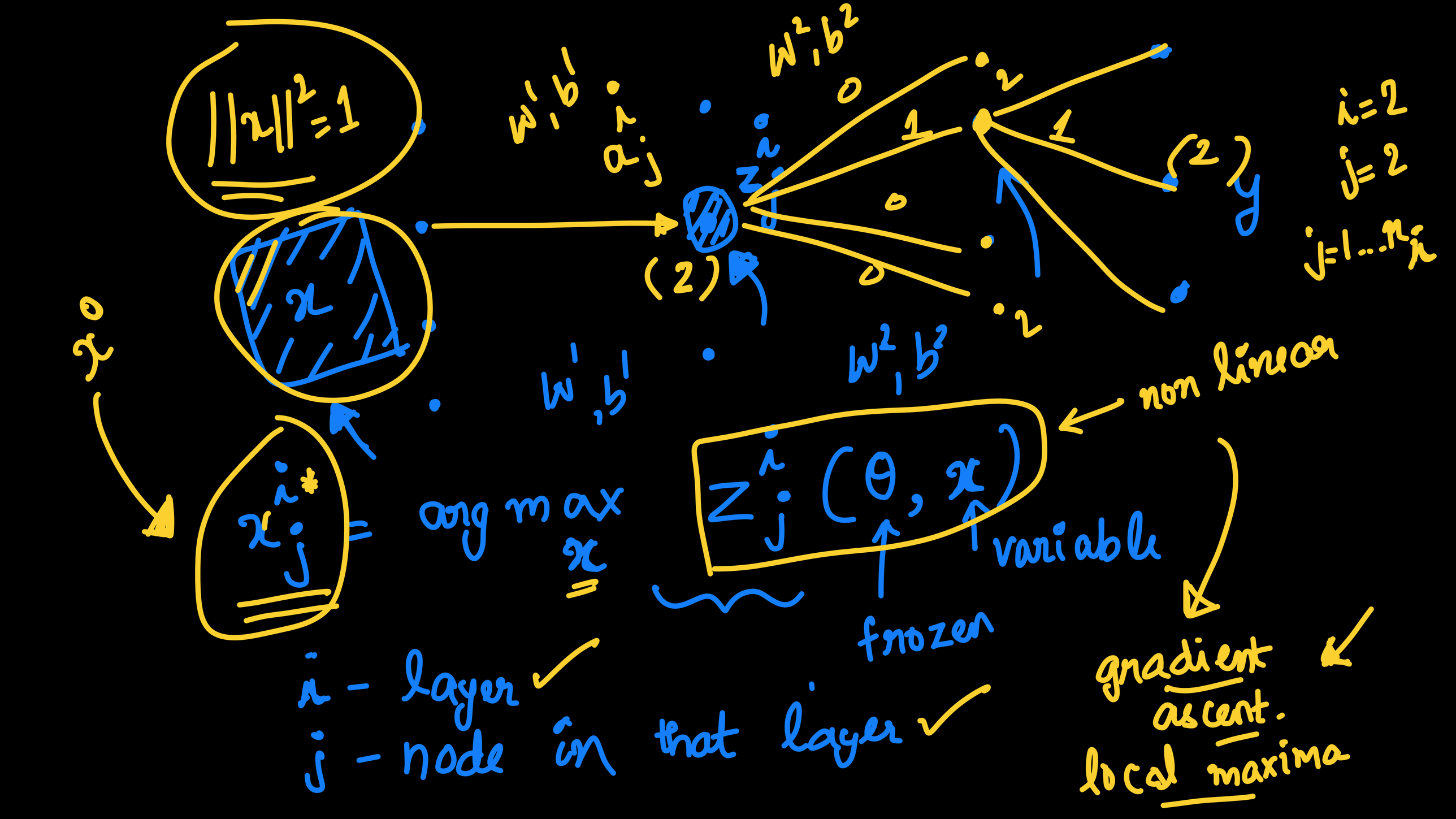
Learning the input pattern of a trained network

- Choose a trained neural network
- Find input patterns that maximize the activations from that neuron
- Solved using gradient ascent

$x = \{x_1 \dots x_N\}$

$\theta = \{W^1, b^1, \dots, W^L, b^L\}$?

fixed

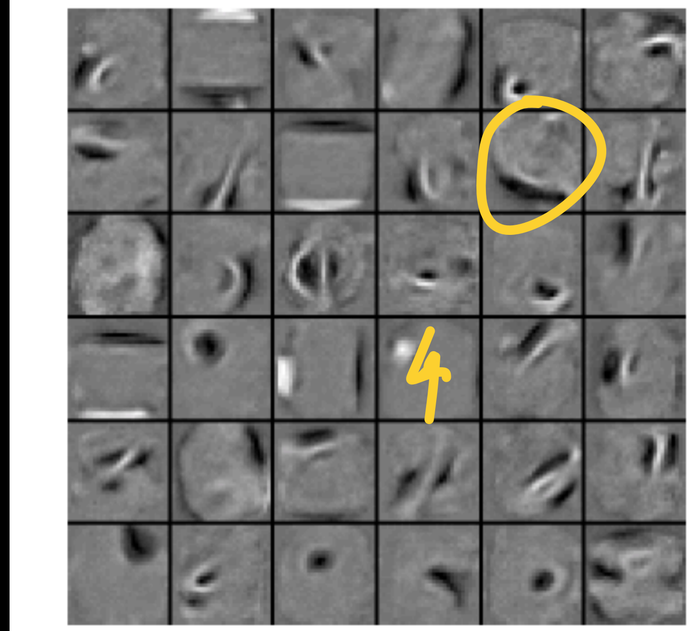

$\|x\|^2 = 1$
W^1, b^1
W^2, b^2
a^1_j
z^2_j
x
x^0
x^{1*}_j
$= \arg\max_x$
$z^2_j(\theta, x)$
non linear
variable
frozen
i = 2
j = 2
$j = 1 \ldots n_2$
i - layer ✓
j - node in that layer ✓
gradient ascent.
local maxima

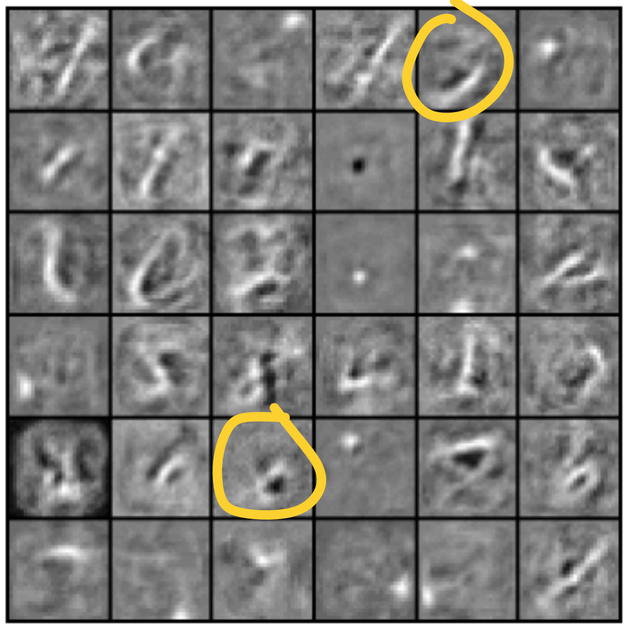
Neuron's

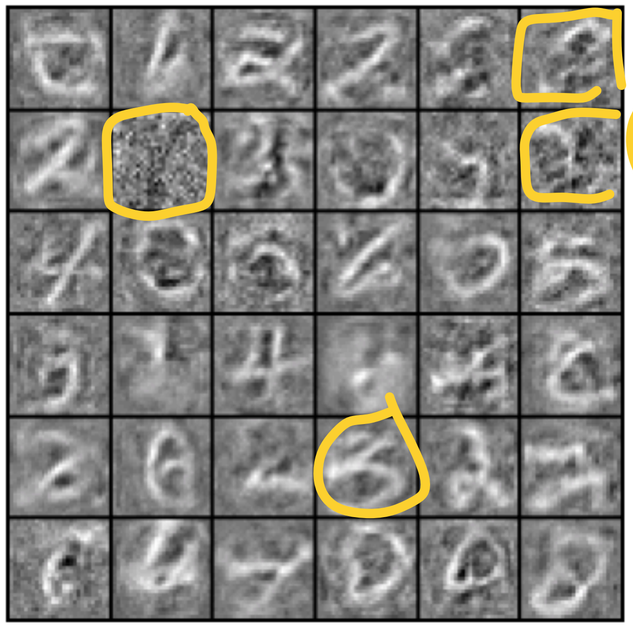
1st layer

2nd layer

3rd layer







t-SNE embeddings for visualization

IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS VOL. 23, NO. 1, JANUARY 2017

Visualizing the Hidden Activity of Artificial Neural Networks

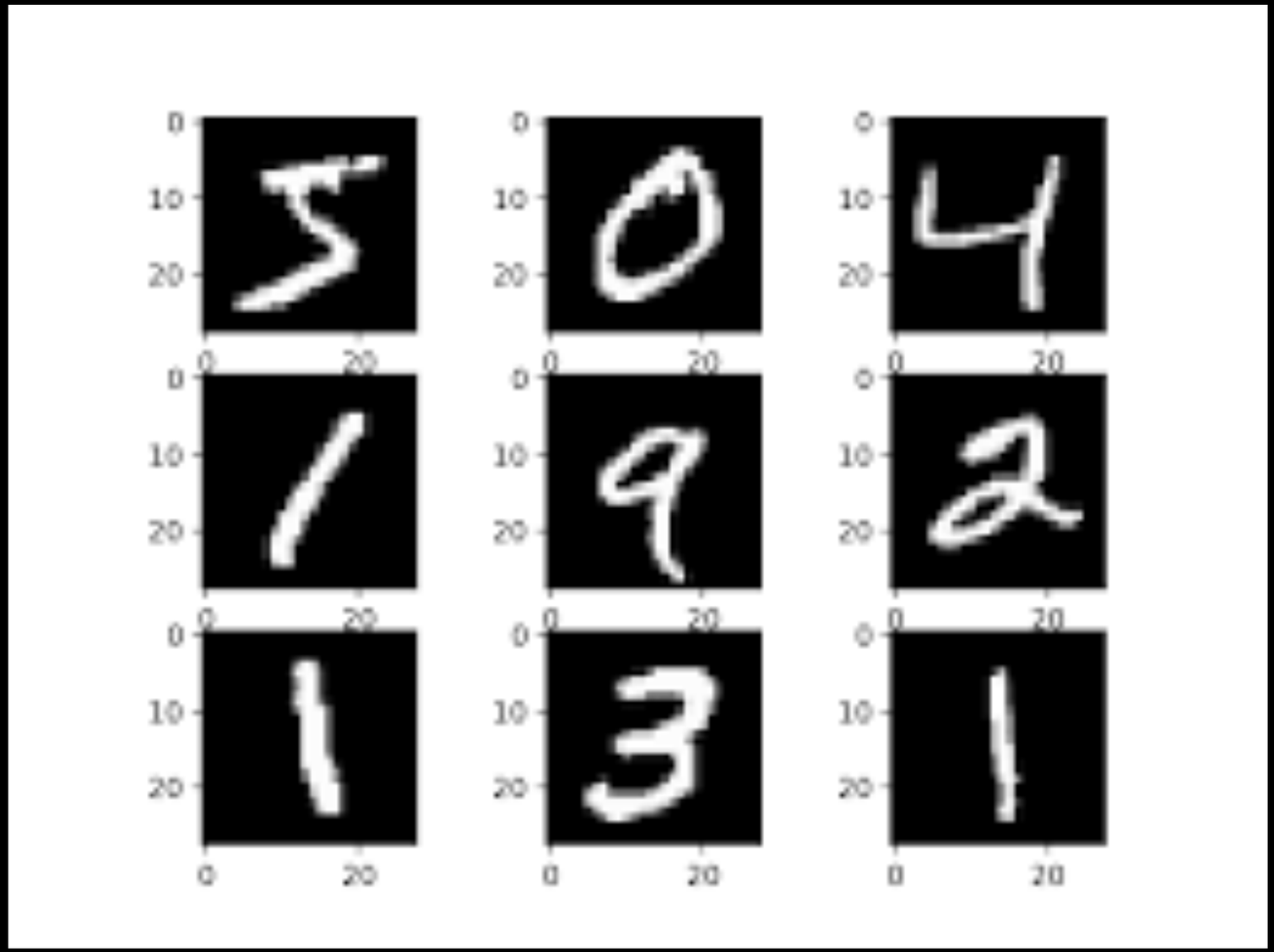
Paulo E. Rauber, Samuel G. Fadel, Alexandre X. Falcão, and Alexandru C. Telea

t-SNE embeddings for visualization

SVHN dataset

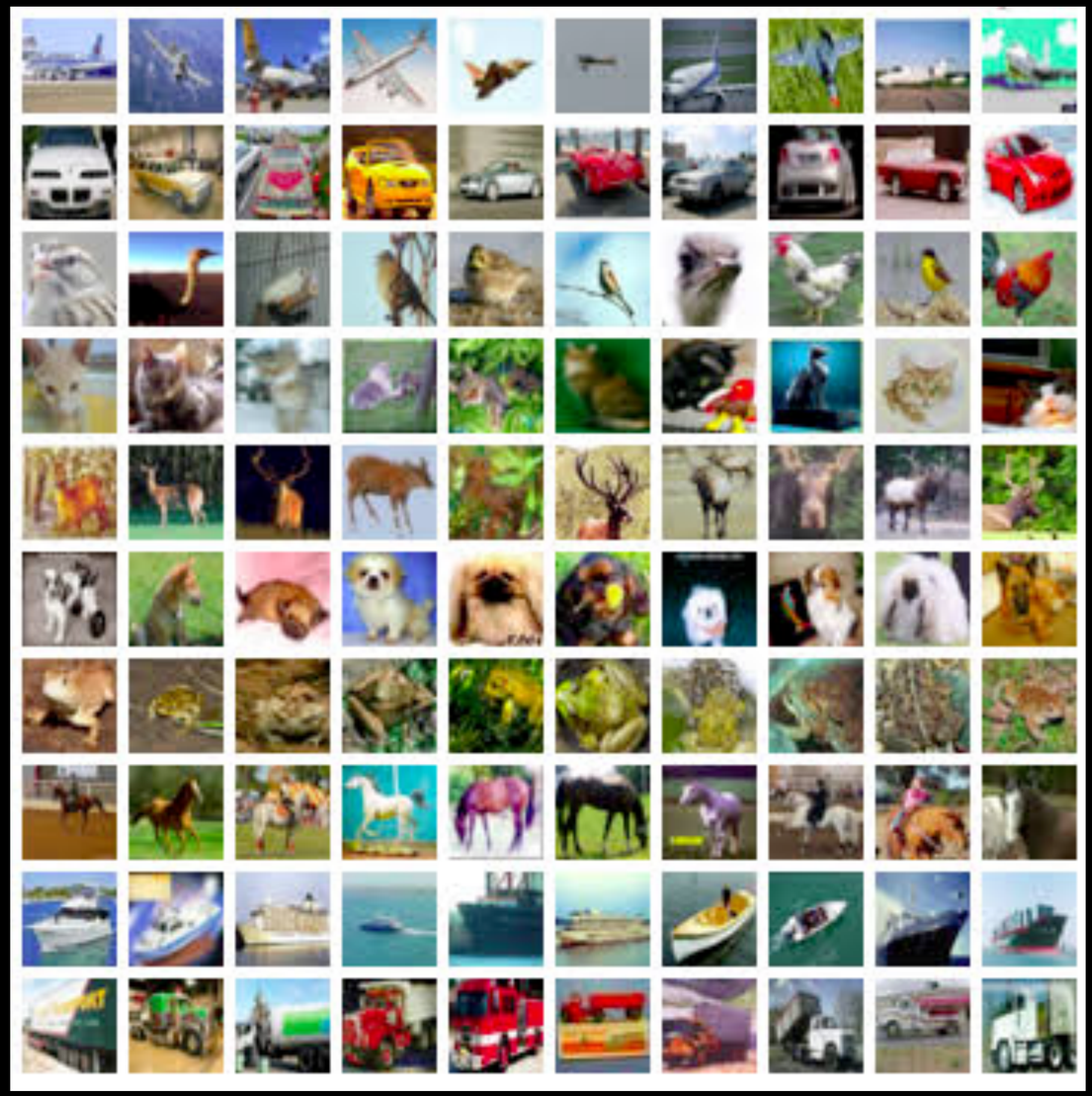
t-SNE embeddings for visualization

MNIST dataset

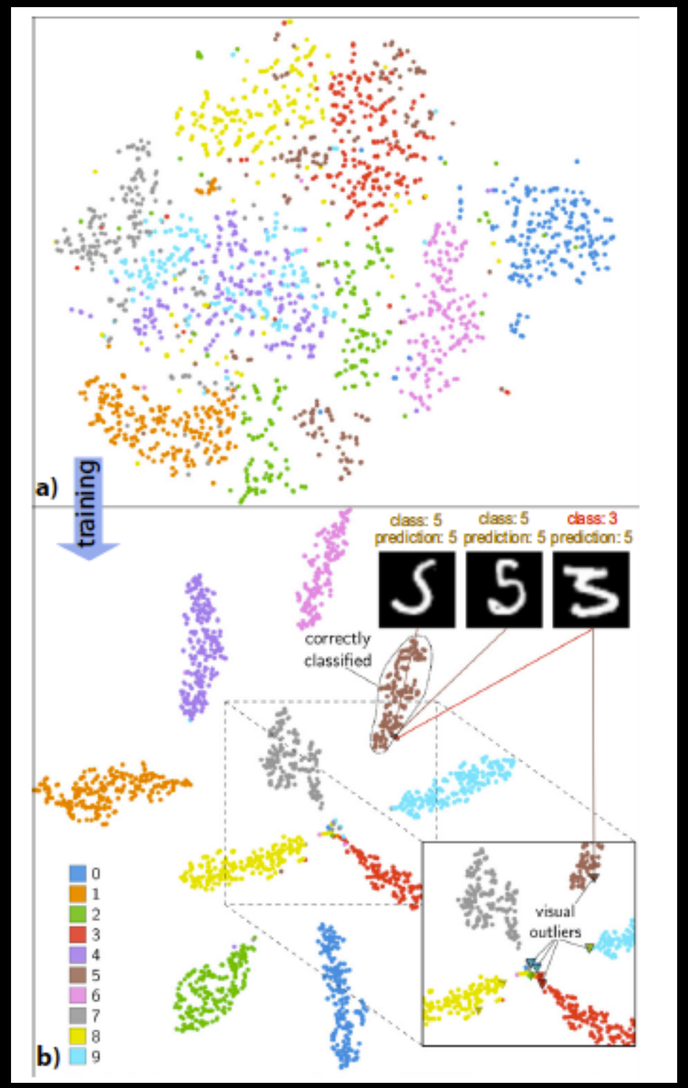


t-SNE embeddings for visualization

CIFA10 dataset



Understanding Deep Networks



tSNE
projection
of last layer
of the neural network.

Fig. 3. Projection of the last MLP hidden layer activations, MNIST test subset. a) Before training (NH: 83.78%). b) After training (NH: 98.36%, AC: 99.15%). Inset shows classification of visual outliers.

Understanding Deep Networks

Fig. 4. Projection of the last MLP hidden layer activations before training, *SVHN* test subset (NH: 20.94%). Poor class separation is visible.

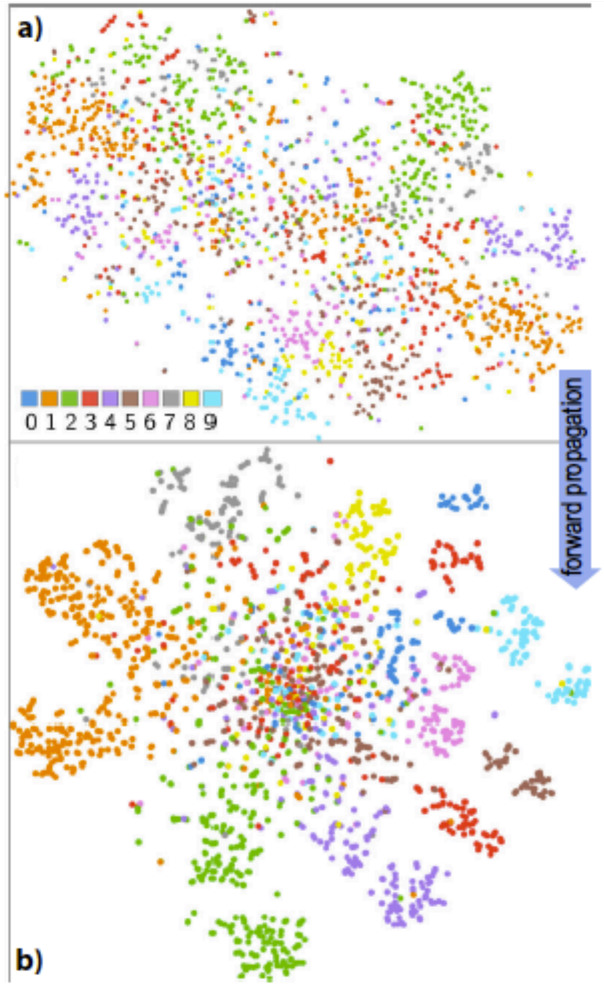


Fig. 5. Projection of the MLP hidden layer activations after training, SVHN test subset. a) First hidden layer (NH: 52.78%). b) Last hidden layer (NH: 67%).

Understanding deep networks

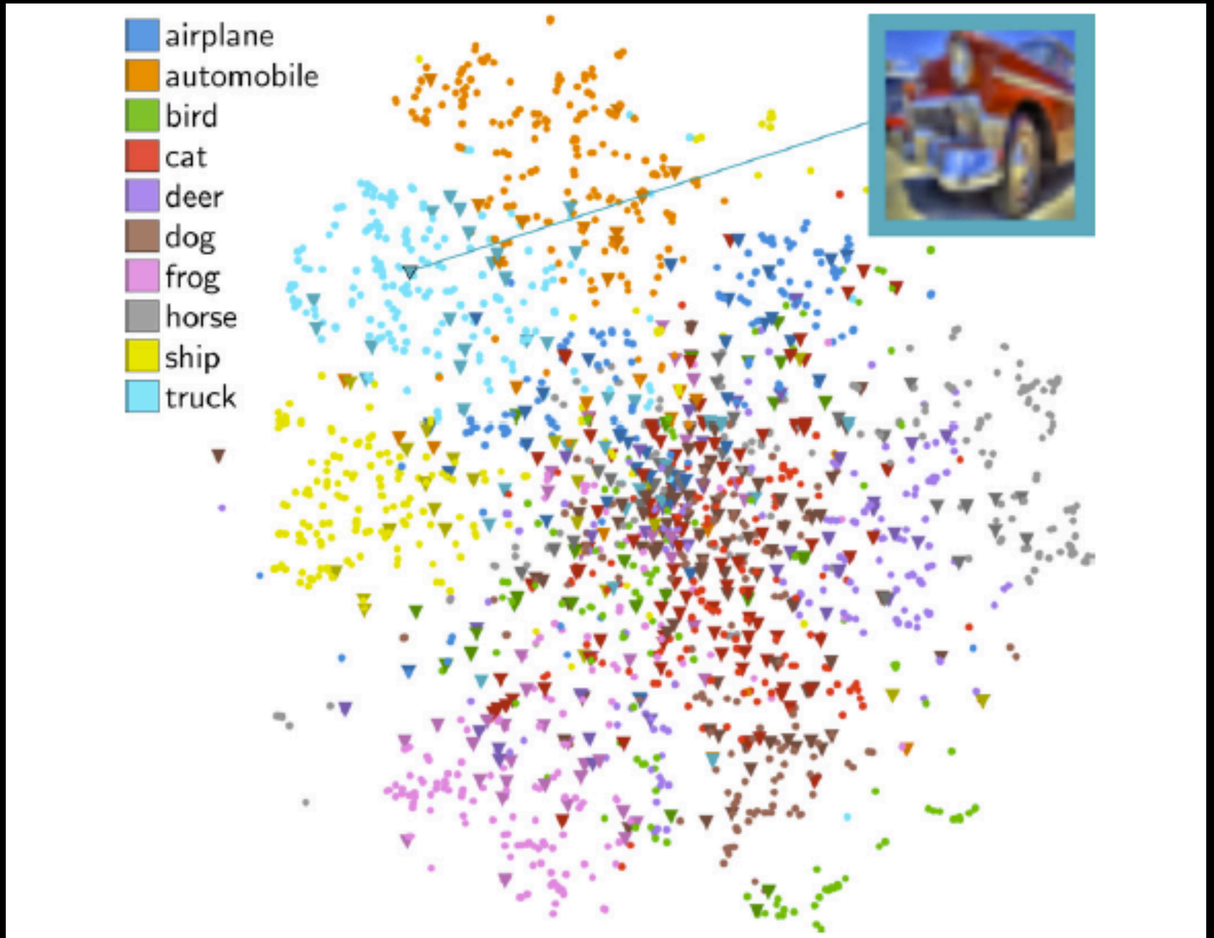


Fig. 9. Projection of last CNN hidden layer activations after training, *CIFAR*-10 test subset (NH: 53.43%, AC: 78.7%).

Understanding Deep Networks

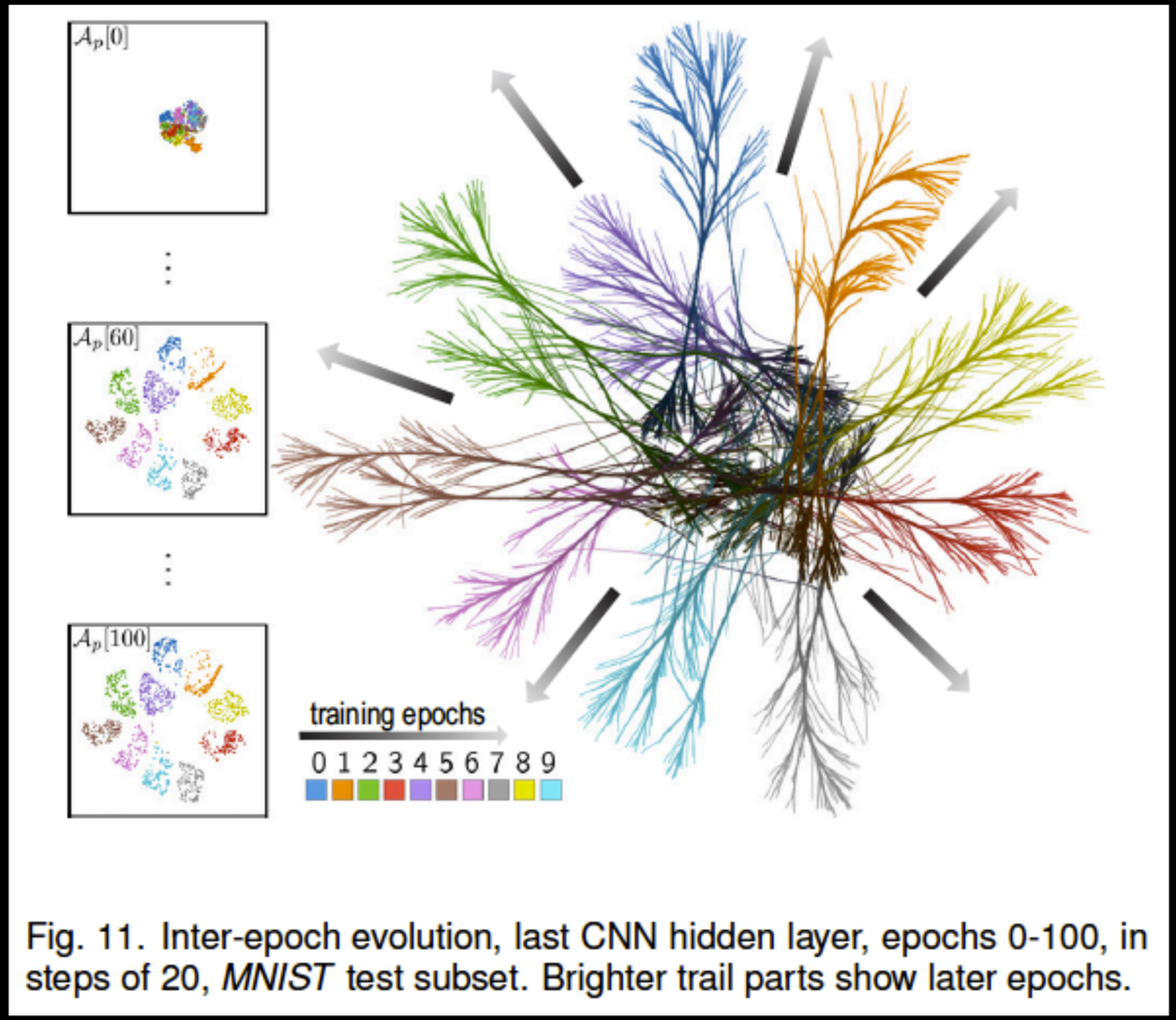


Fig. 11. Inter-epoch evolution, last CNN hidden layer, epochs 0-100, in steps of 20, *MNIST* test subset. Brighter trail parts show later epochs.

Understanding Deep Networks

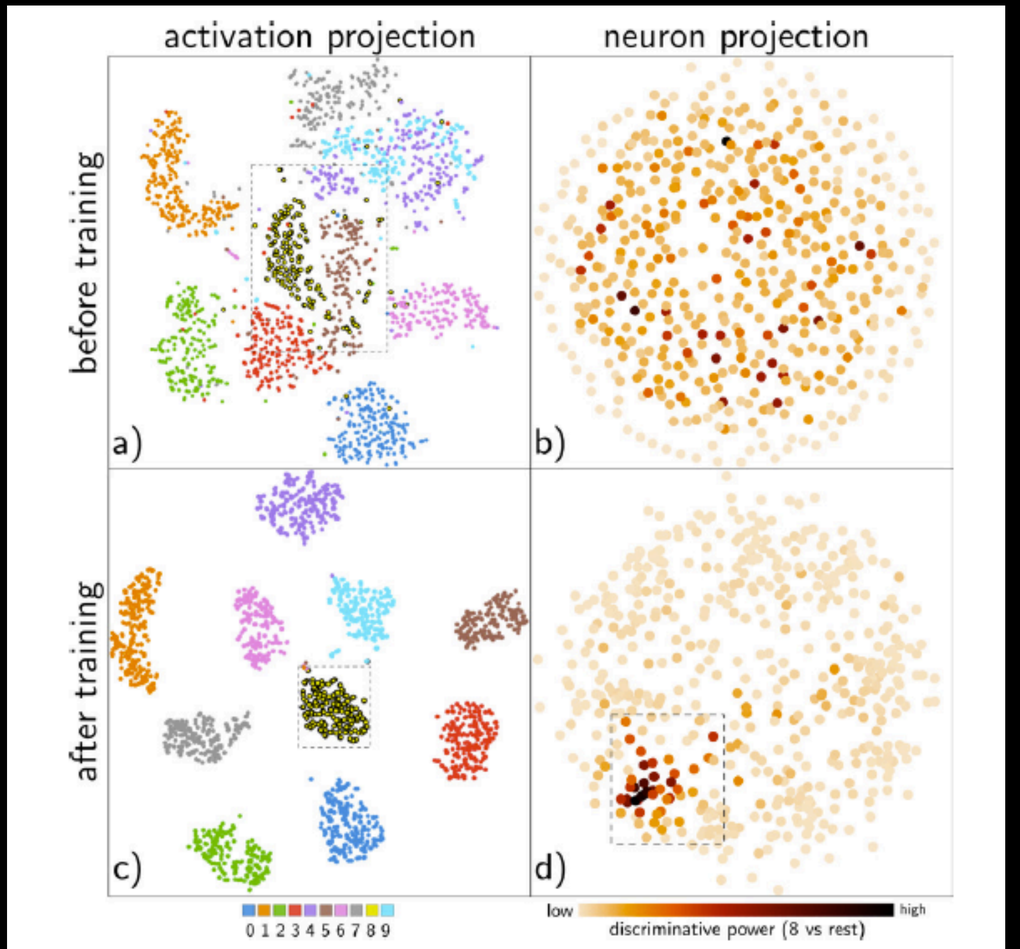


Fig. 12. Activation and neuron projections of last CNN hidden layer activations before and after training, *MNIST* test subset. Neuron projection colors show the neurons' power to discriminate class 8 *vs* rest.

Hierachical representations in deep networks

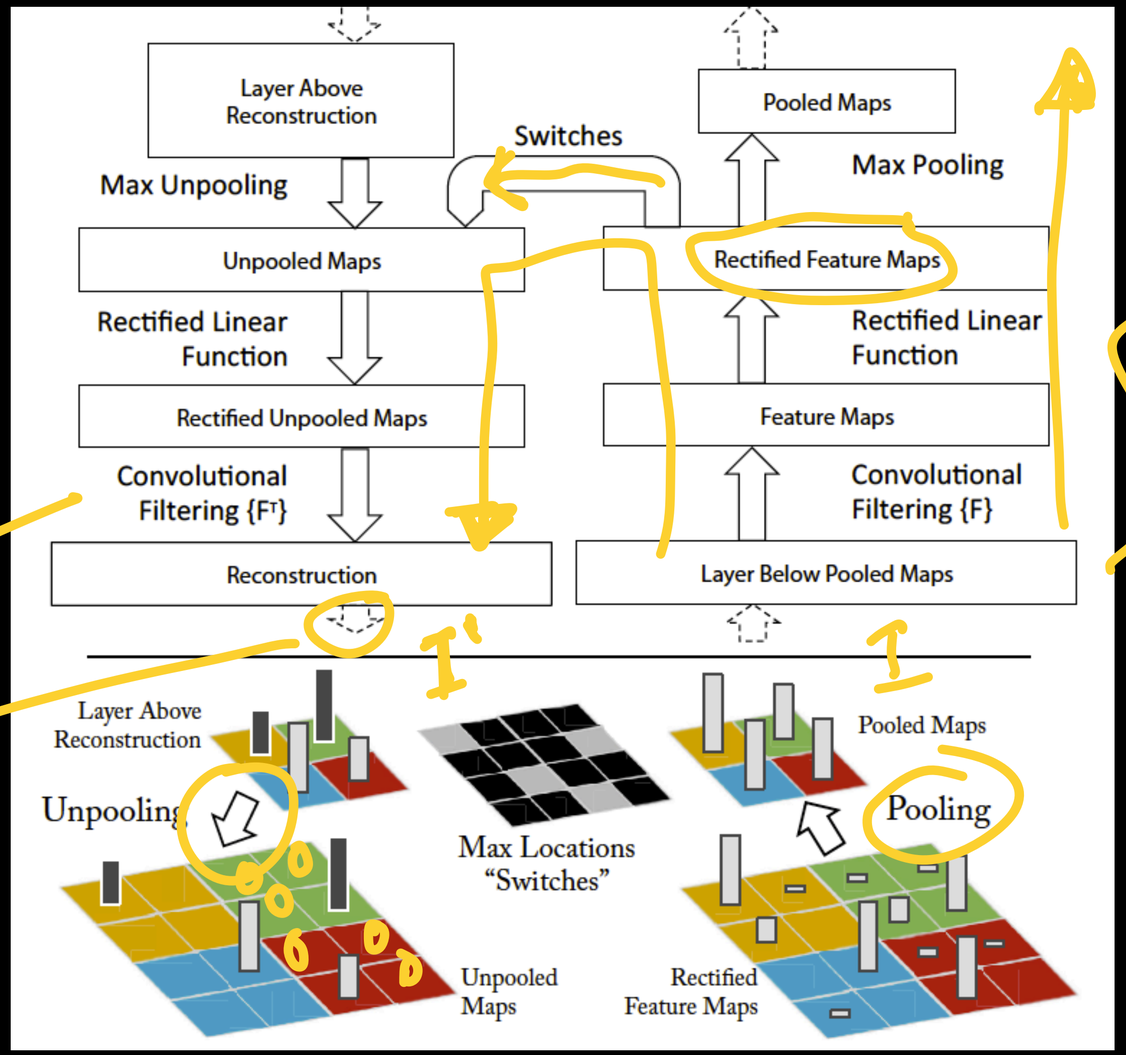
Visualizing and Understanding Convolutional Networks

Matthew D. Zeiler and Rob Fergus

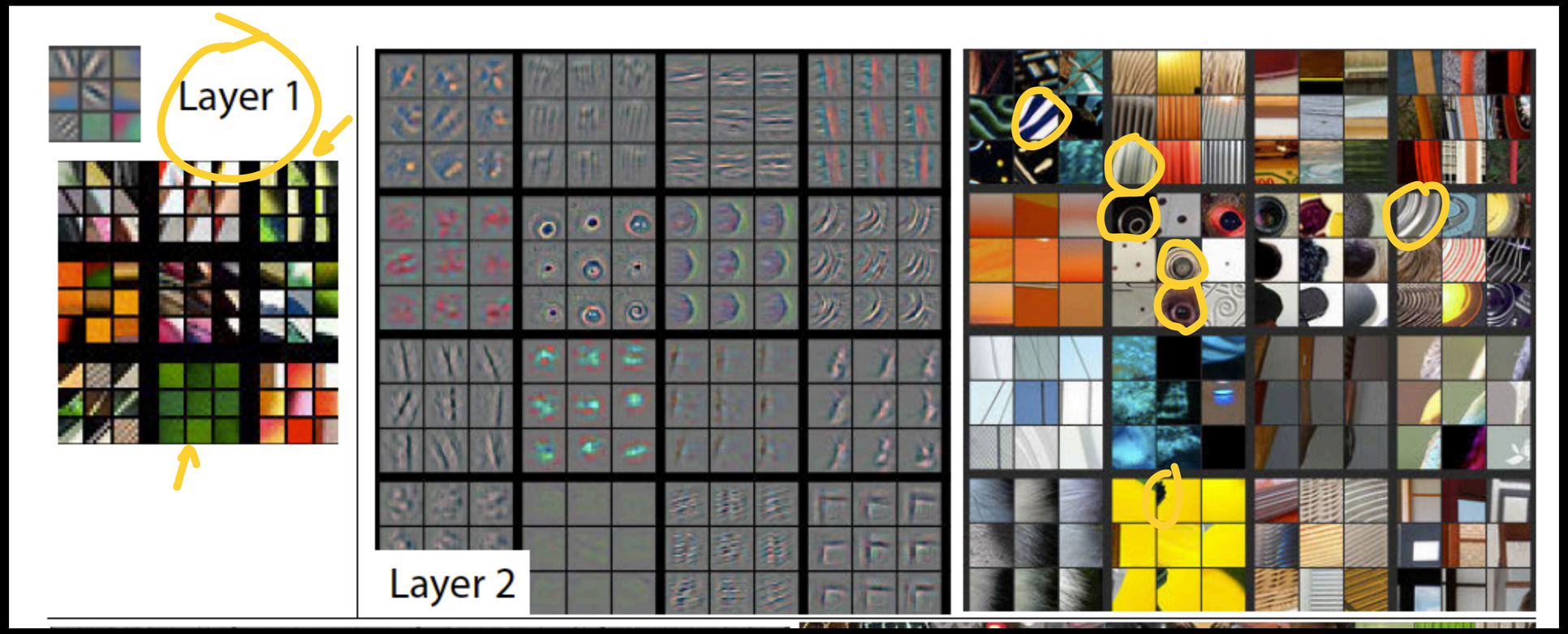
Dept. of Computer Science,
New York University, USA
{zeiler,fergus}@cs.nyu.edu

2014

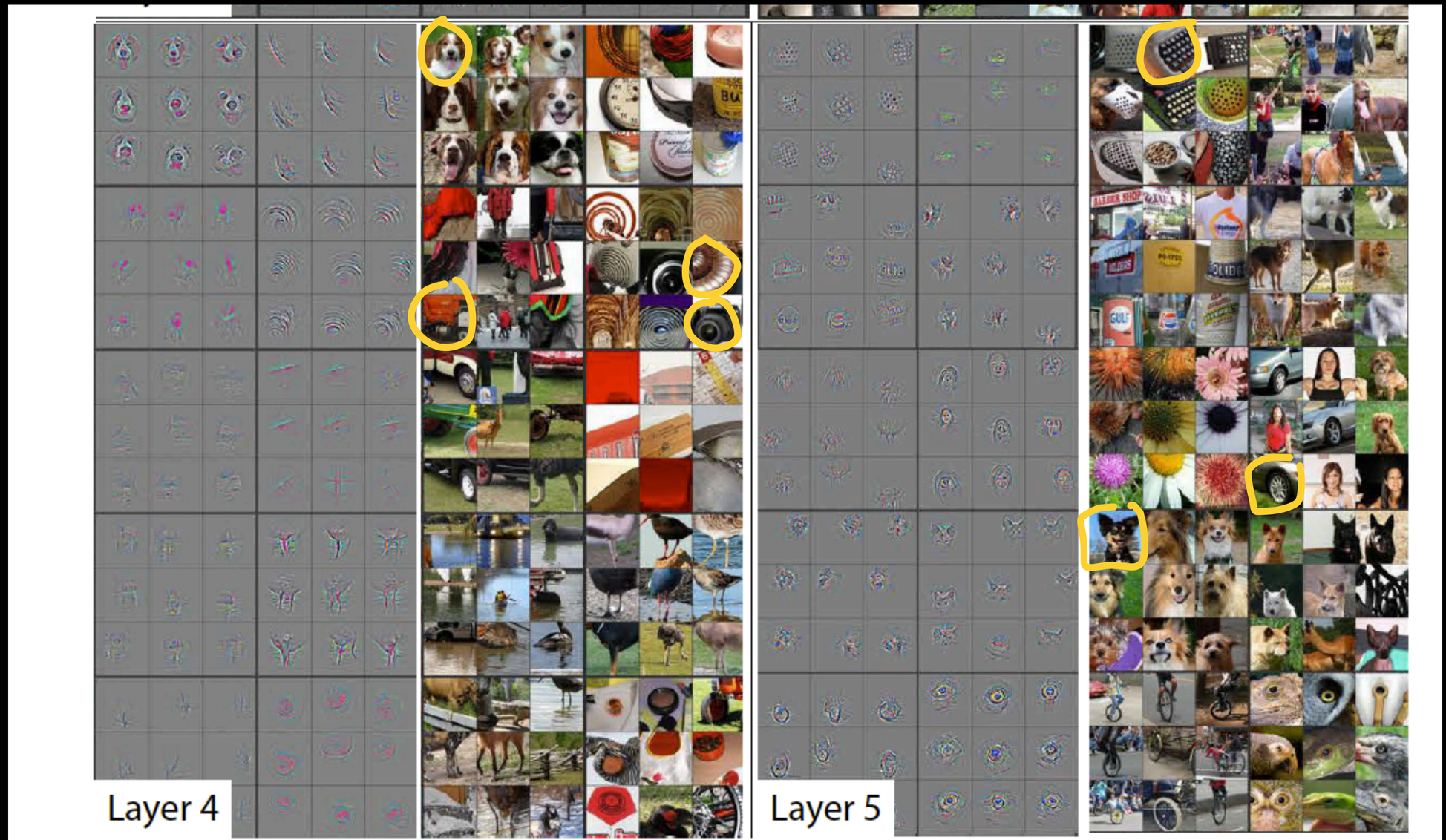
Hierachical representations in deep networks



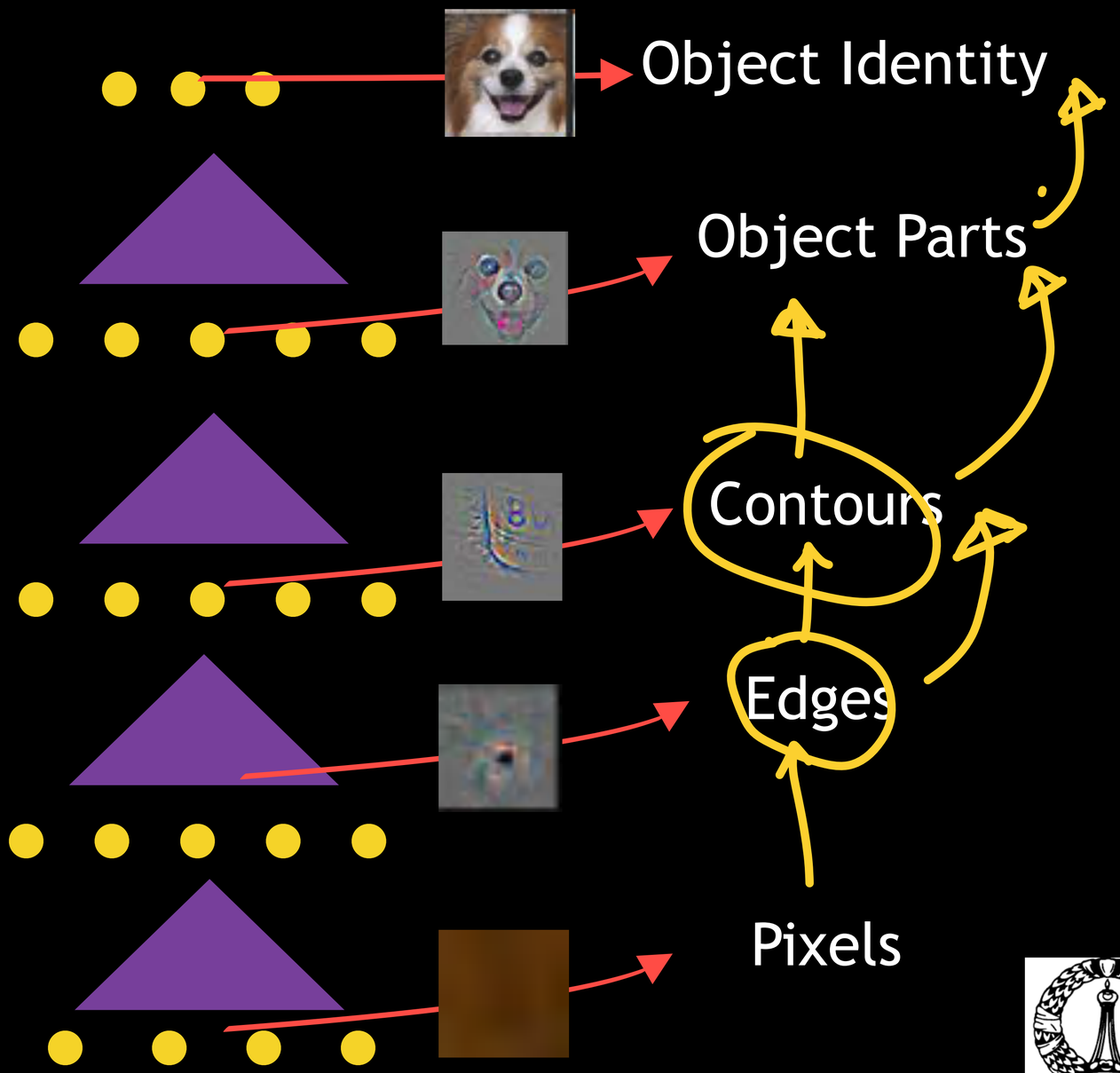
Understanding Deep Networks



Understanding Deep Networks



Hierachical representations in deep networks



[Zeiler, 2014]

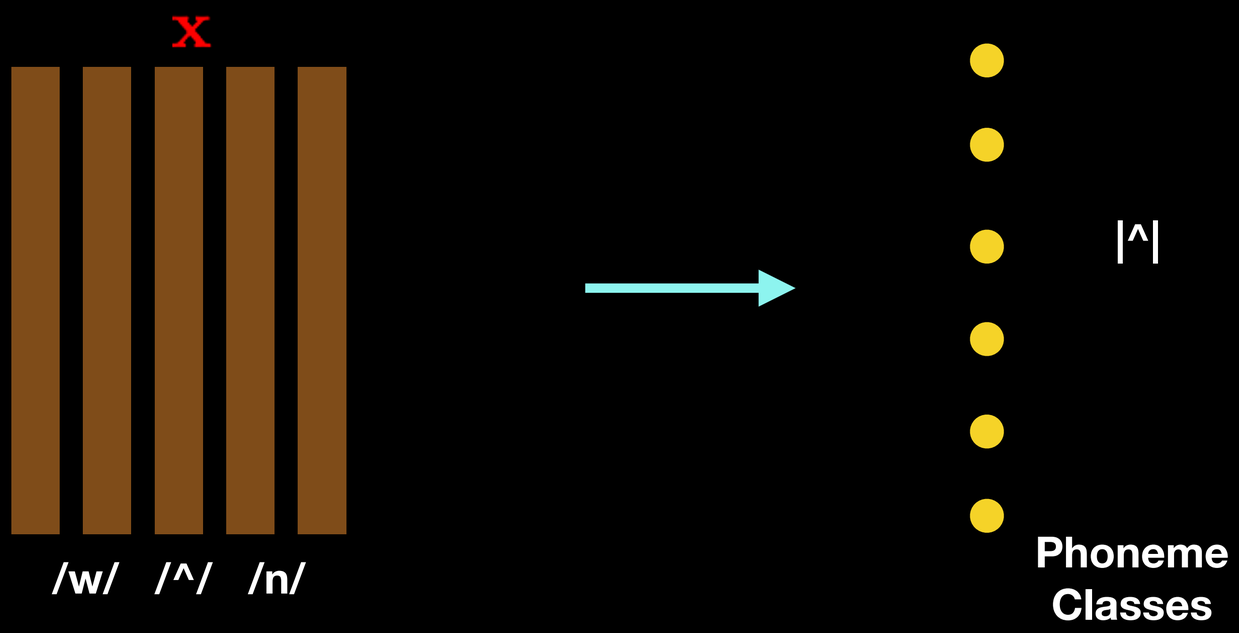
UNDERSTANDING HOW DEEP BELIEF NETWORKS PERFORM ACOUSTIC MODELLING

Garcia-Romero, Daniel, et al. "Speaker diarization using deep neural network embeddings." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

Abdel-rahman Mohamed, Geoffrey Hinton, and Gerald Penn

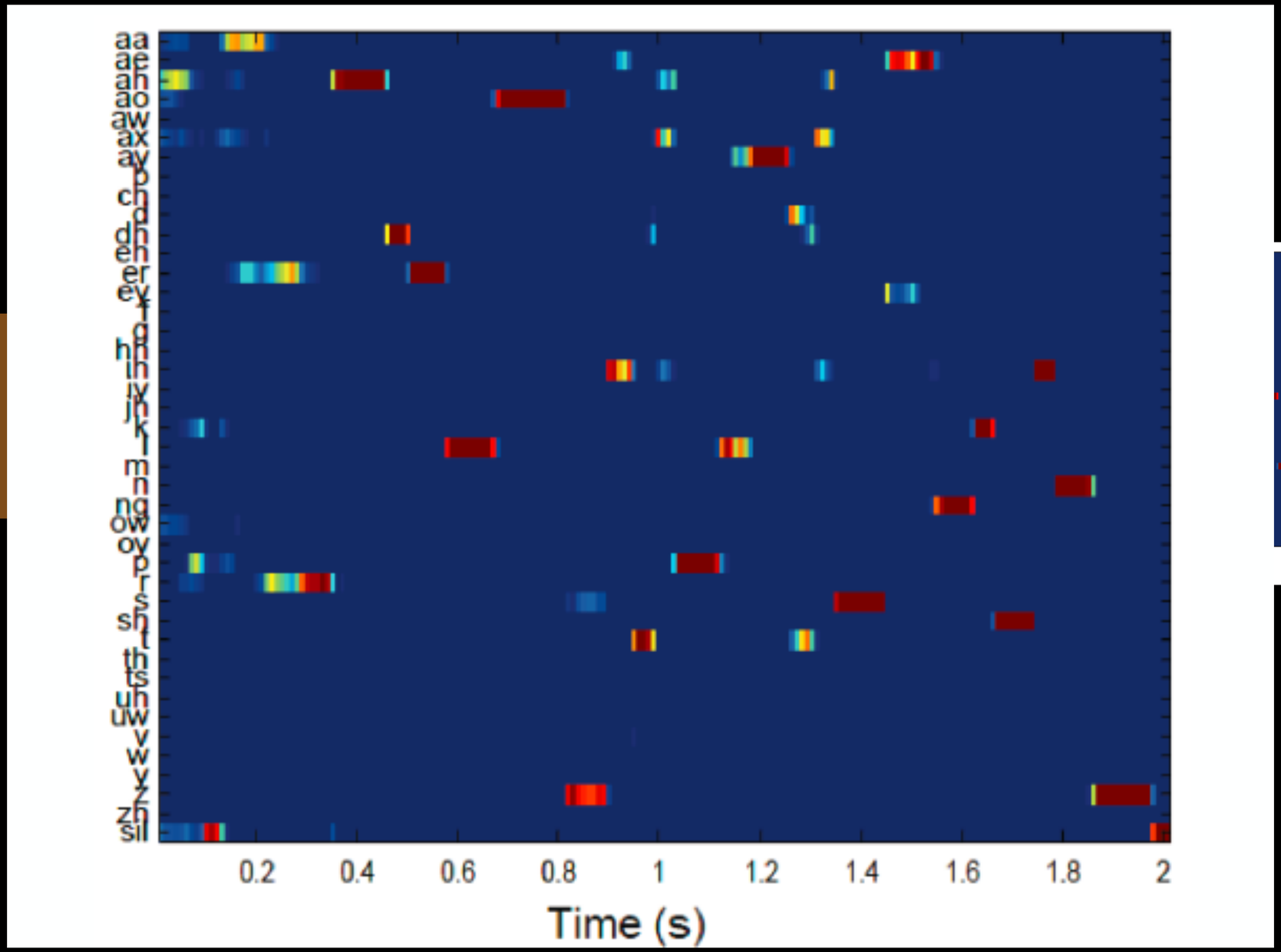
Department of Computer Science, University of Toronto

Speech Recognition

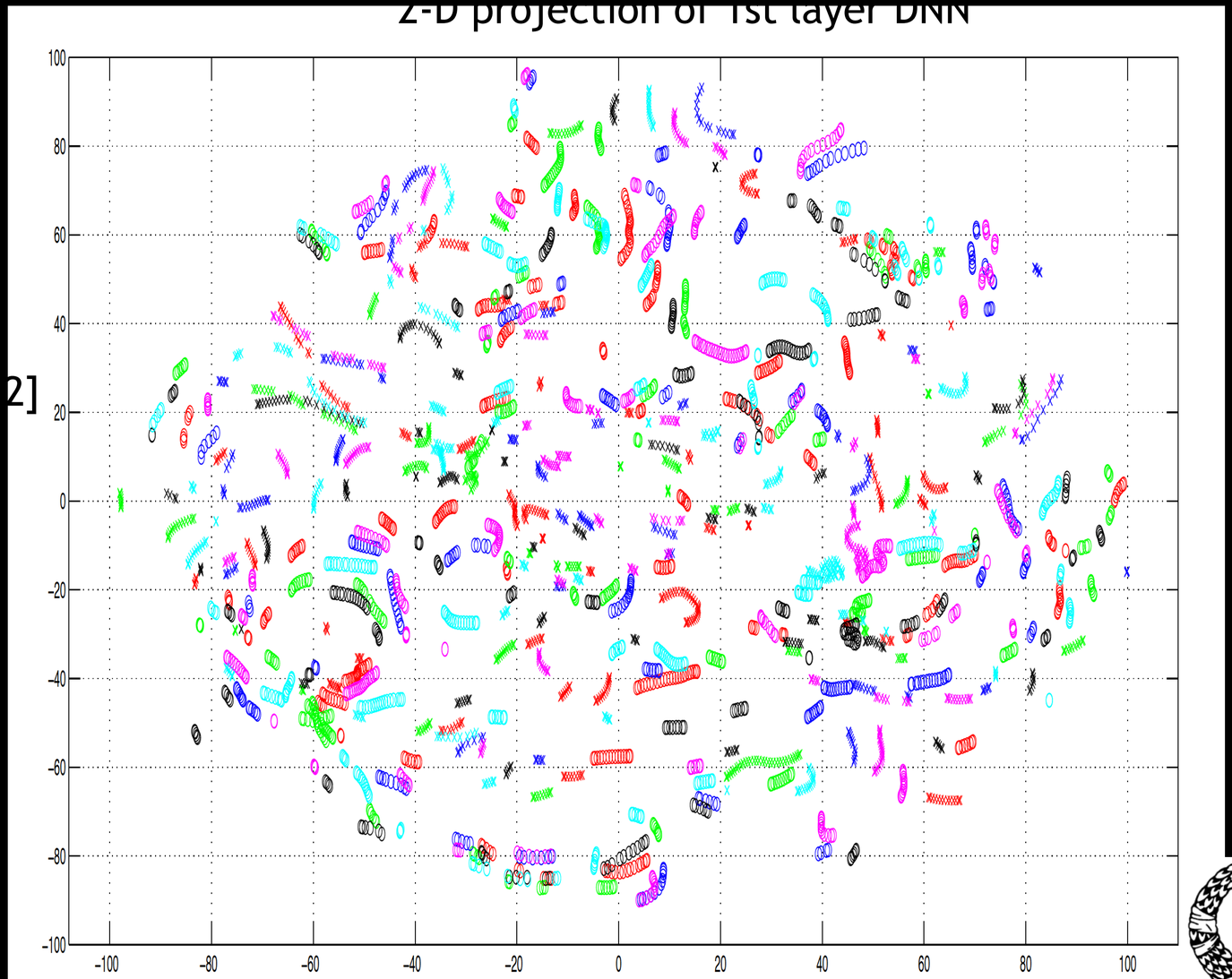


- Classical machine learning - train a classifier on speech training data that maps to the target phoneme class.

Speech recognition



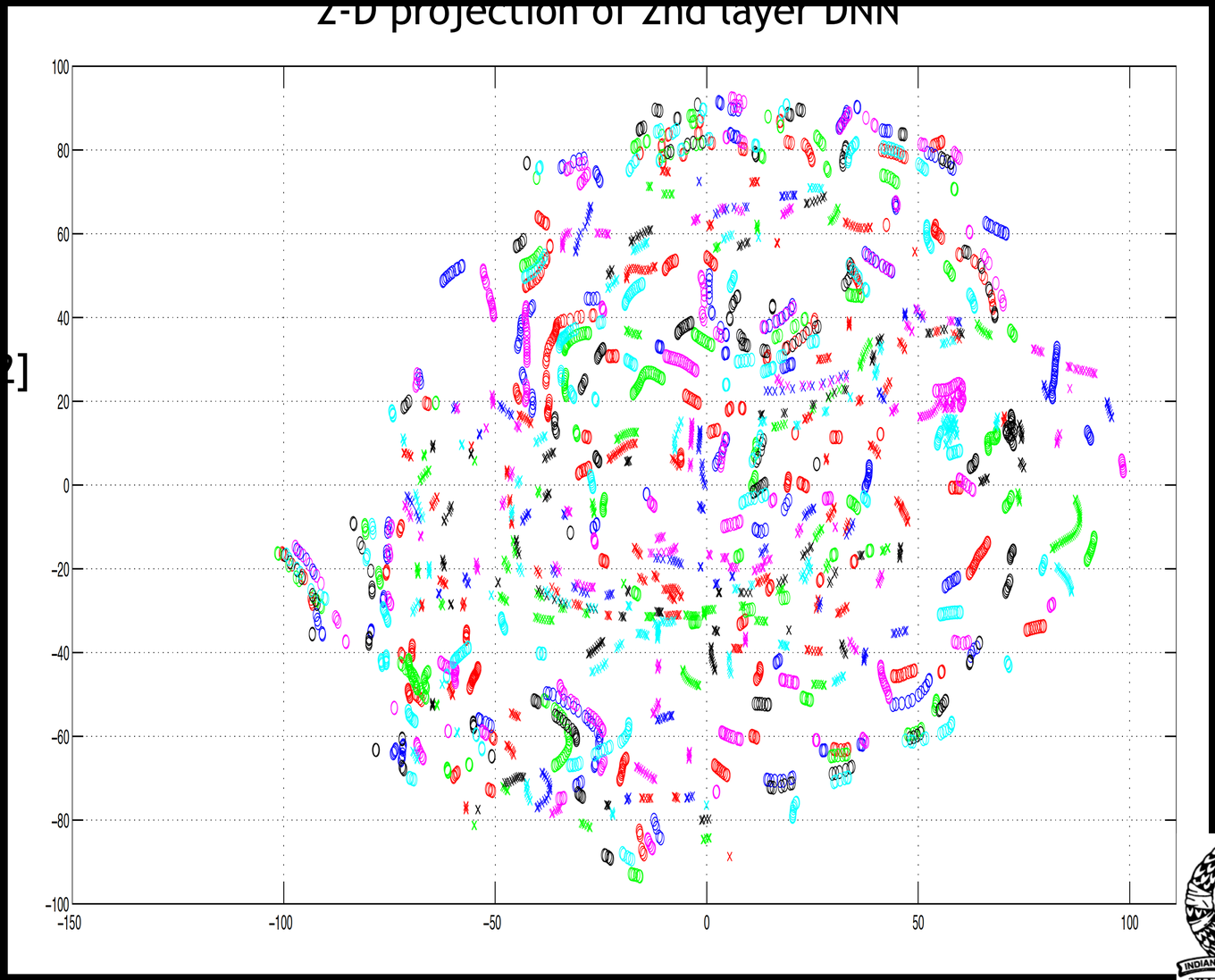
Understanding DNNs for Speech



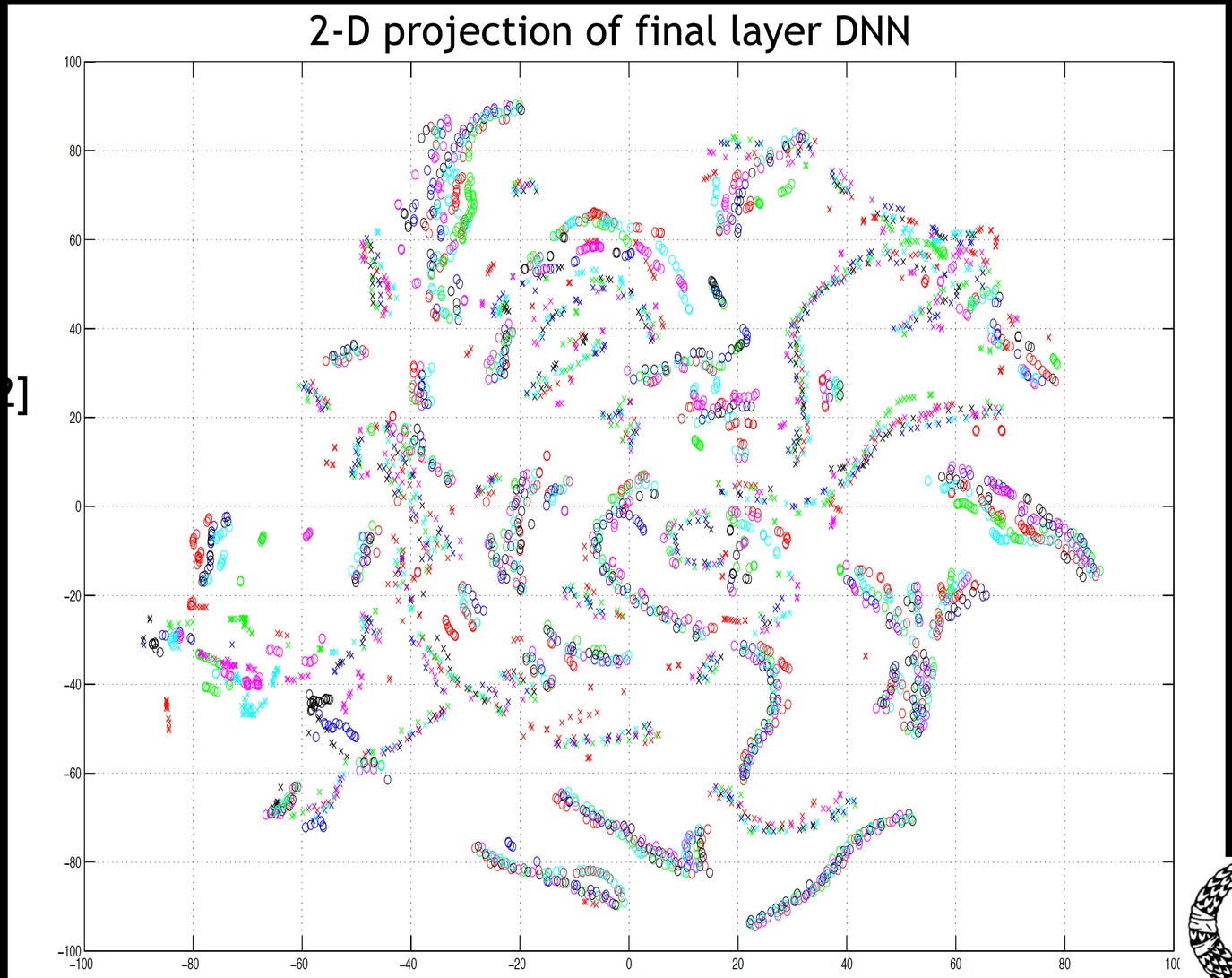
Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

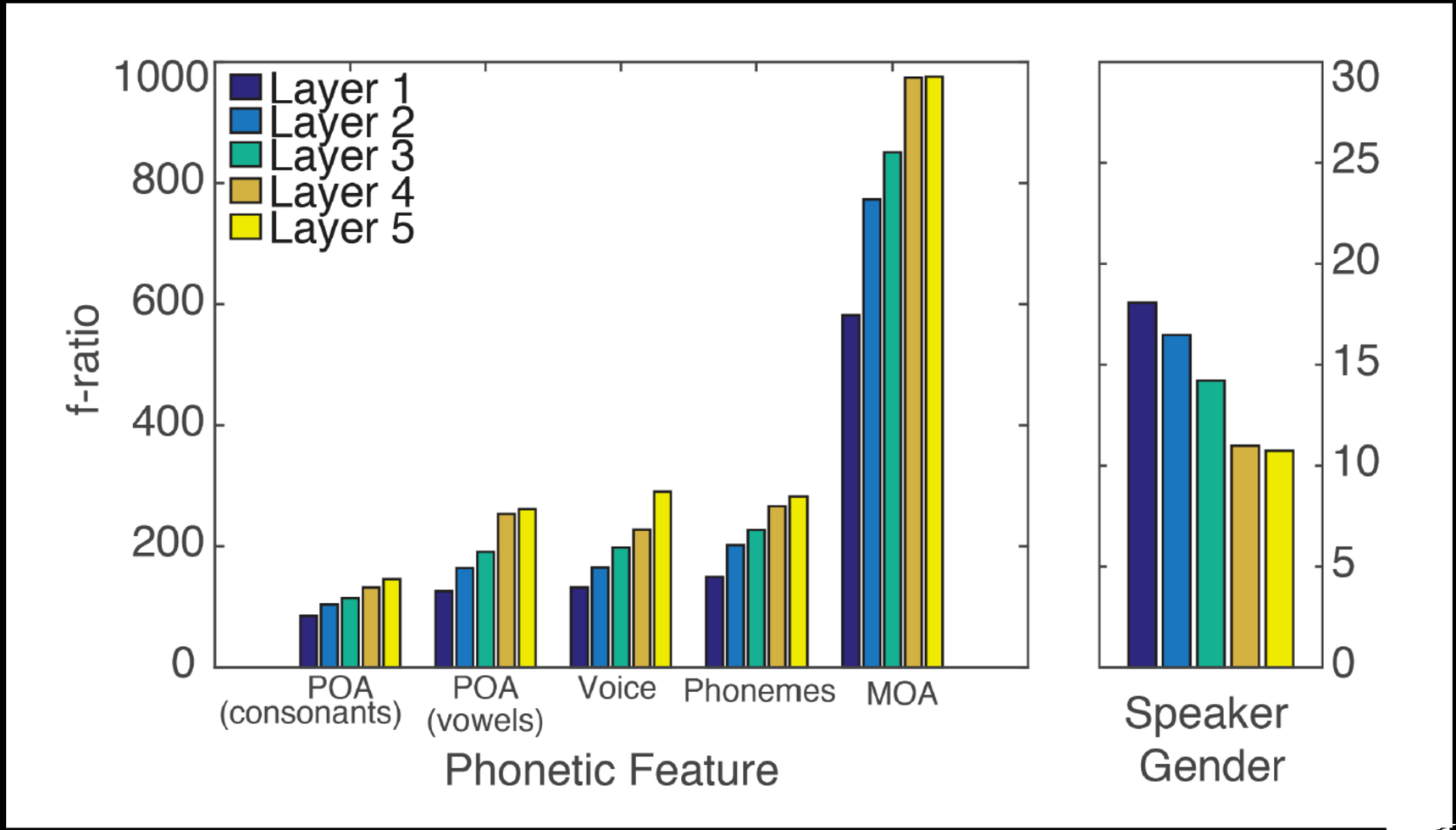
Understanding DNNs for Speech



Understanding DNNs for Speech



Understanding DNNs for Speech



Summary thus far

★ Deep neural networks perform hierarchical data abstractions

- ✓ Early layers form representations that are less oriented towards the task.
- ✓ Later layers form representations that more oriented to the task.

★ Connections with biological processing of audio/images.

Questions about representations

- Can we quantify the degree to which a particular layer is general or specific?
- Does the transition occur suddenly at a single layer, or is it spread out over several layers?
- Where does this transition take place: near the first, middle, or last layer of the network?

Questions about representations

How transferable are features in deep neural networks?

Jason Yosinski,[1] **Jeff Clune,**[2] **Yoshua Bengio,**[3] **and Hod Lipson**[4]
[1] Dept. Computer Science, Cornell University
[2] Dept. Computer Science, University of Wyoming
[3] Dept. Computer Science & Operations Research, University of Montreal
[4] Dept. Mechanical & Aerospace Engineering, Cornell University

Questions about representations

A

B

Imagenet Dataset

Questions about representations

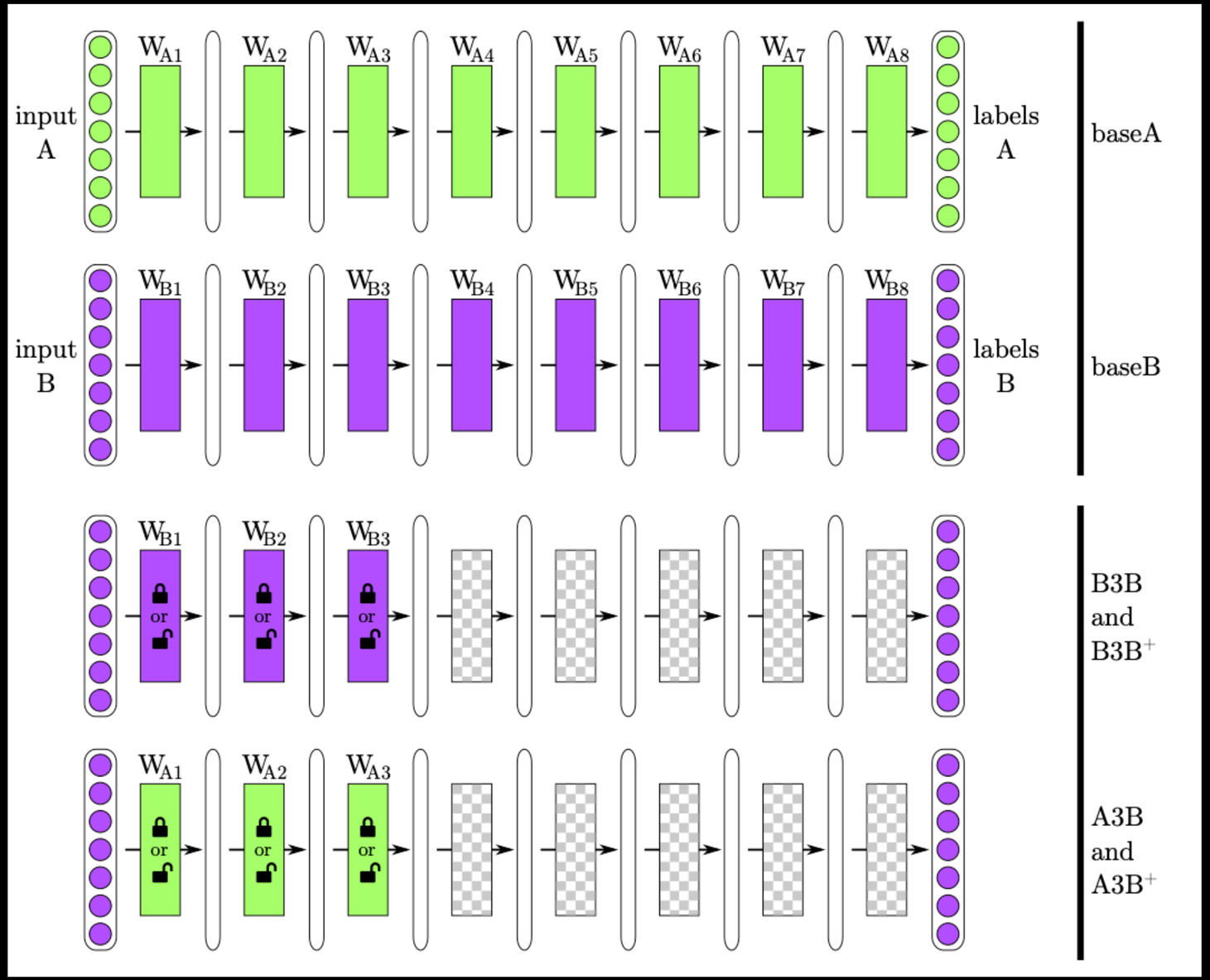
- A selffer network B3B: the first 3 layers are copied from baseB and frozen. The five higher layers (4–8) are initialized randomly and trained on dataset B. This network is a control for the next transfer network.

- A transfer network A3B: the first 3 layers are copied from baseA and frozen. The five higher layers (4–8) are initialized randomly and trained toward dataset B. Intuitively, here we copy the first 3 layers from a network trained on dataset A and then learn higher layer features on top of them to classify a new target dataset B.

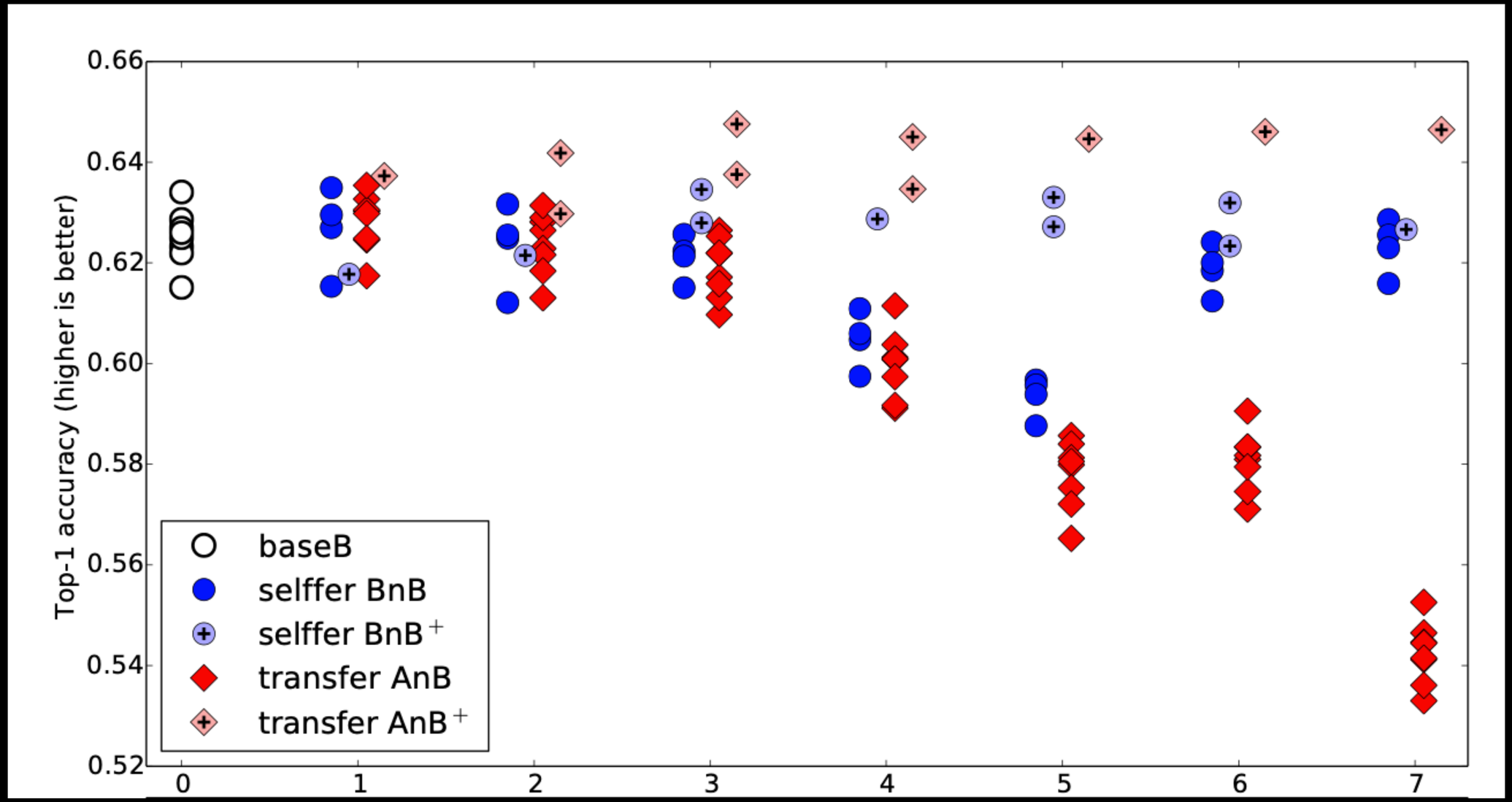
Questions about representations

- Can we quantify the degree to which a particular layer is general or specific?
- Does the transition occur suddenly at a single layer, or is it spread out over several layers?
- Where does this transition take place: near the first, middle, or last layer of the network?

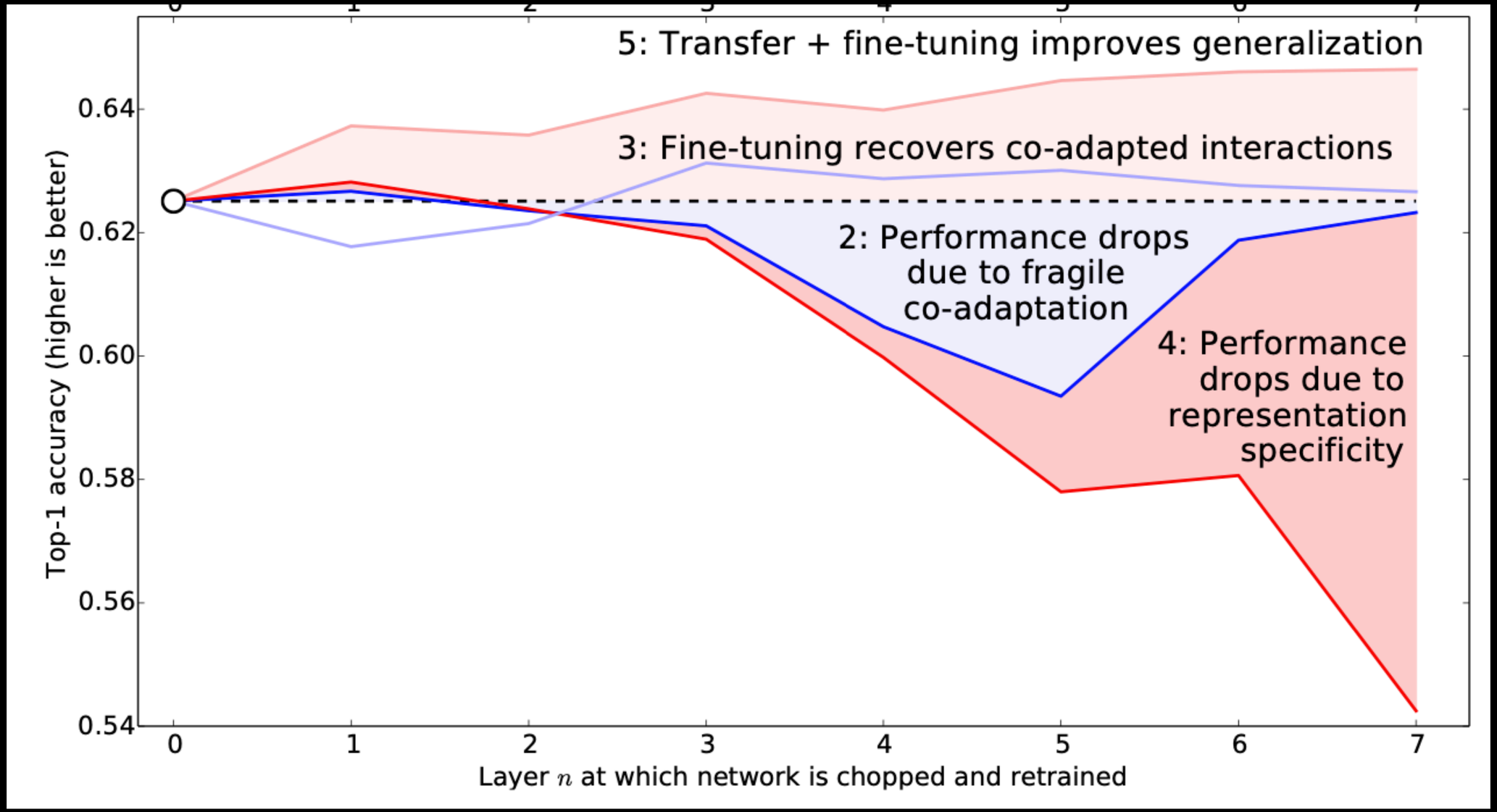
Questions about representations



Questions about representations



Questions about representations



Questions about representations