# Learning Disentangled Representations that Generalize Across Tasks, Domains and Time

**Yoshua Bengio**

December 13, 2014

NIPS'2014 Workshop on Transfer and Multi-Task learning

Université de Montréal

PLUG: **Deep Learning**, MIT Press book in preparation, draft chapters online for feedback

LISA

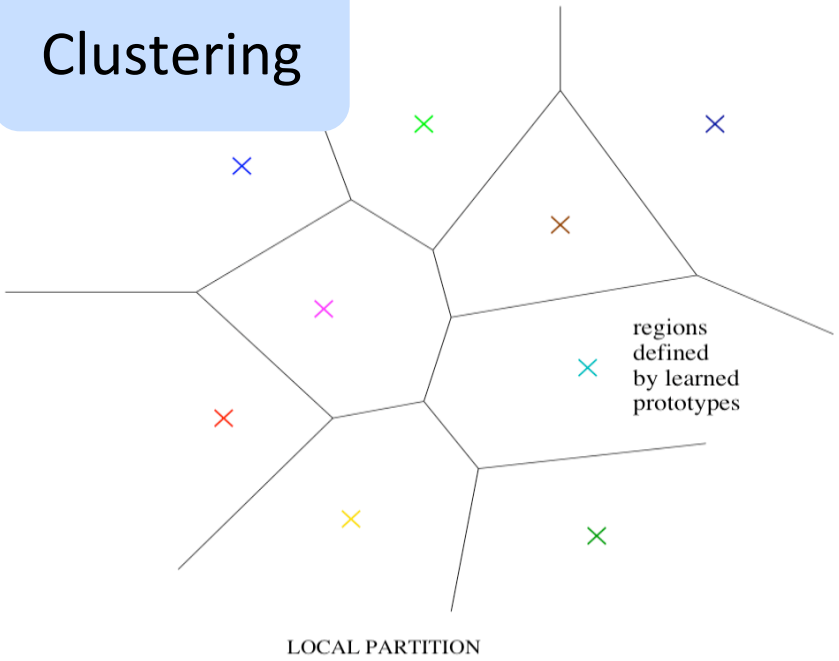
# Technical Goals Hierarchy

## To reach AI:

- Needs **knowledge**
- Needs **learning**
  (involves priors + ***optimization***/*search* ***+ efficient computation***)
- Needs **generalization**
  (guessing where probability mass concentrates)
- Needs ways to fight the curse of dimensionality
  (exponentially many configurations of the variables to consider)
- Needs disentangling the underlying explanatory factors
  (making sense of the data)

# Non-distributed representations

Clustering




- Clustering, Nearest-Neighbors, RBF SVMs, local non-parametric density estimation & prediction, decision trees, etc.

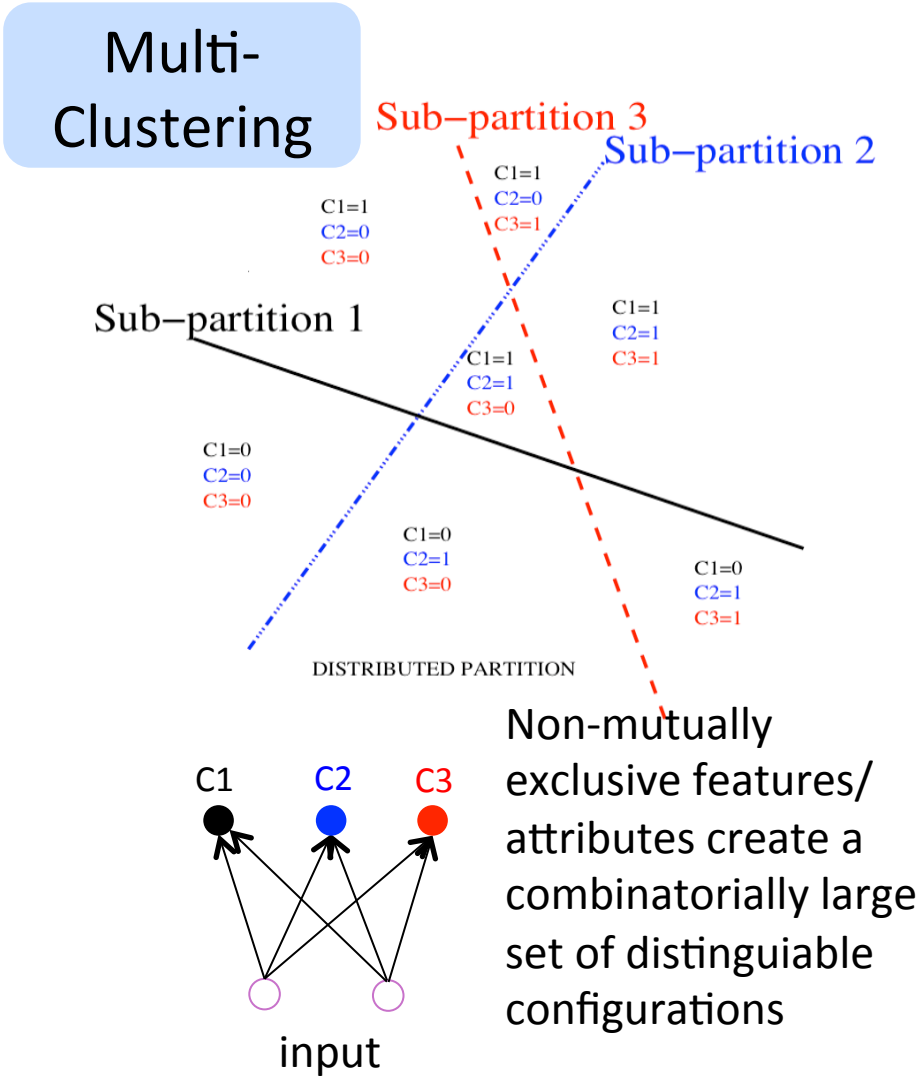
- Parameters for each distinguishable region
- **# of distinguishable regions is linear in # of parameters**

→ No non-trivial generalization to regions without examples

# The need for distributed representations

- Factor models, PCA, RBMs, Neural Nets, Sparse Coding, Deep Learning, etc.

- Each parameter influences many regions, not just local neighbors
- **# of distinguishable regions grows almost exponentially with # of parameters**
- **GENERALIZE NON-LOCALLY TO NEVER-SEEN REGIONS**



# Learning multiple levels of representation

There is theoretical and empirical evidence in favor of multiple levels of representation

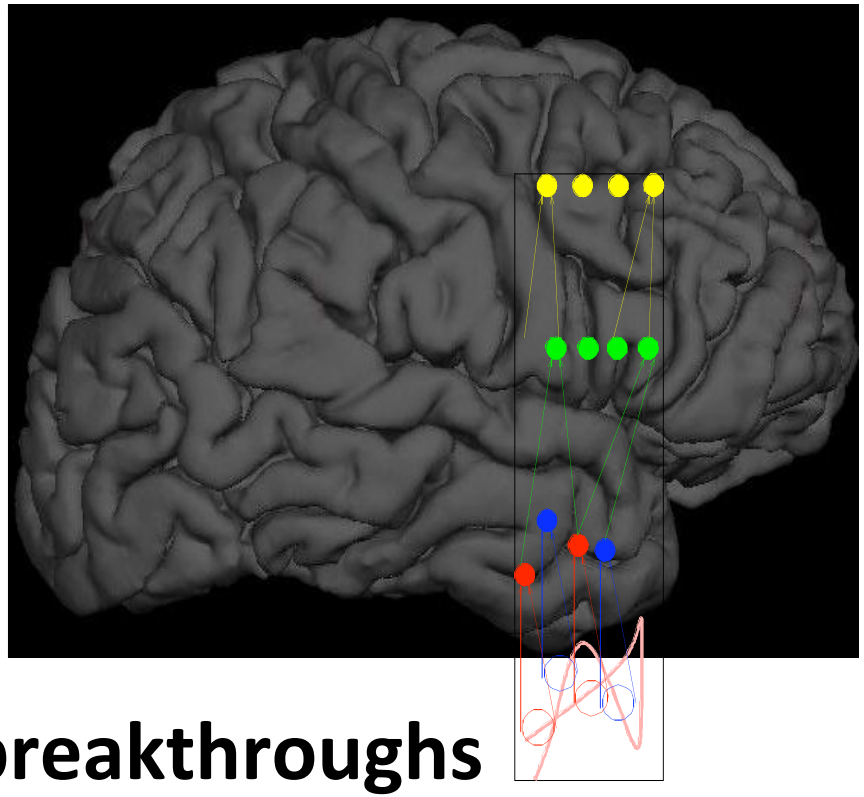
**Exponential gain for some families of functions**

Biologically inspired learning

Brain has a deep architecture

Cortex seems to have a generic learning algorithm

**Humans first learn simpler concepts and compose them**



**It works! Speech + vision breakthroughs**

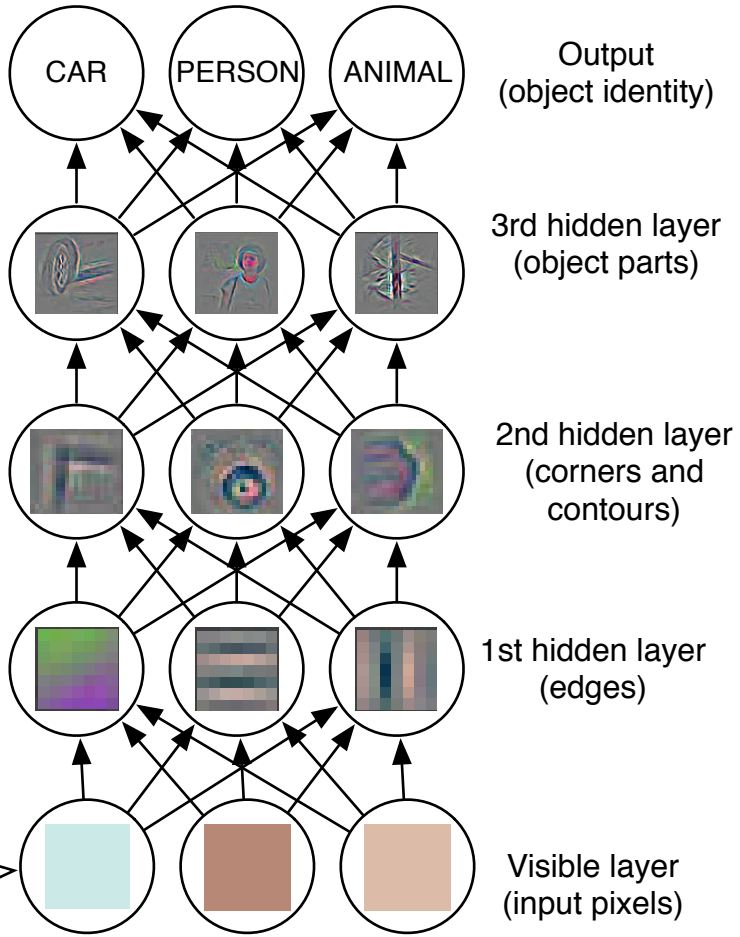
# Composing Features on Features

Higher-level features

are defined in terms of

lower-level

features



# Why Unsupervised Representation Learning? Because of Causality.

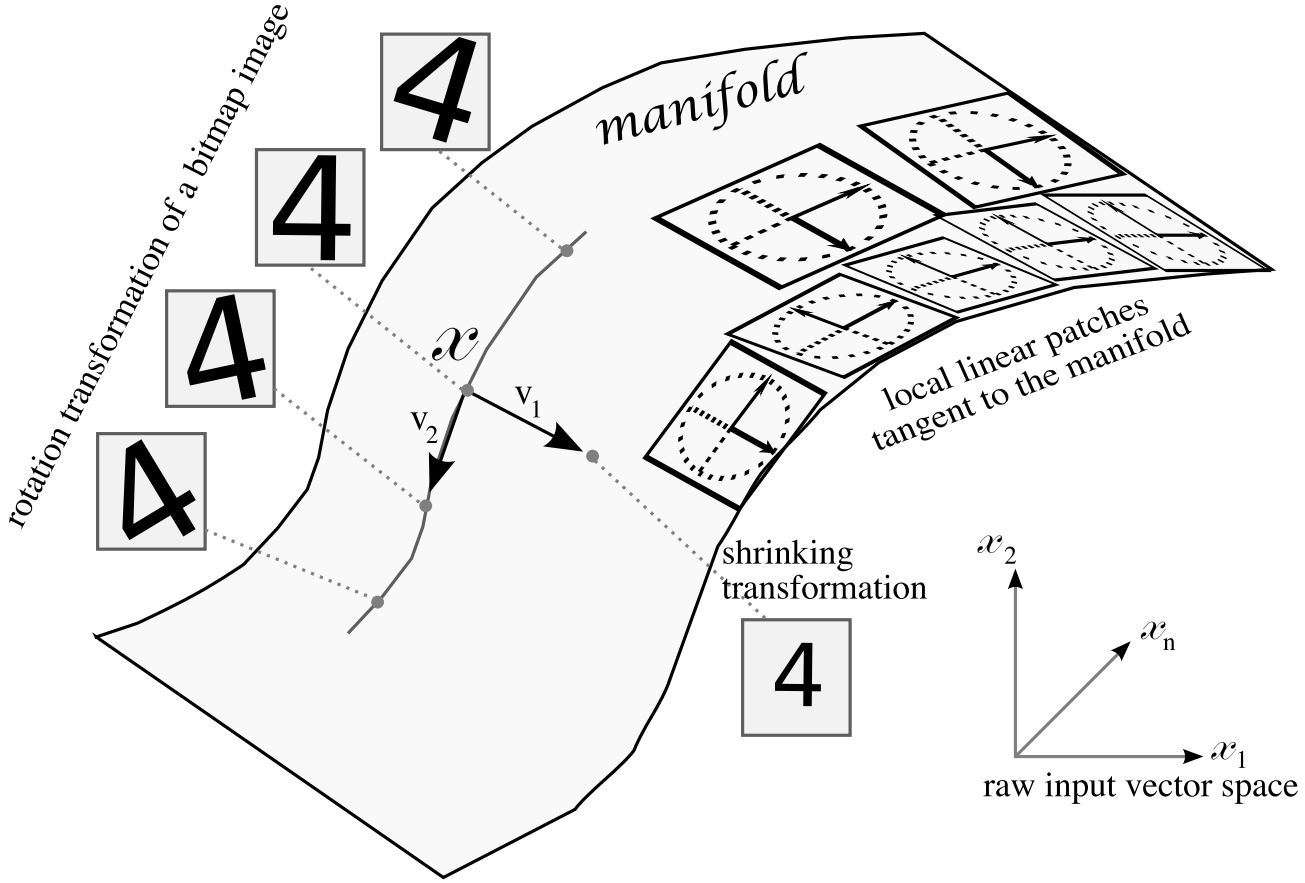
- If Ys of interest are among the causal factors of X, then

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

is tied to P(X) and P(X|Y), and P(X) is defined in terms of P(X|Y), i.e.

- The best possible model of X (unsupervised learning) MUST involve Y as a latent factor, implicitly or explicitly.
- Representation learning SEEKS the latent variables H that explain the variations of X, making it likely to also uncover Y.
- We need 3 pieces:
  - latent variable model P(H),
  - generative decoder P(X|H), and
  - approximate inference encoder Q(H|X).

# Real Data Are on Highly Curved Manifolds

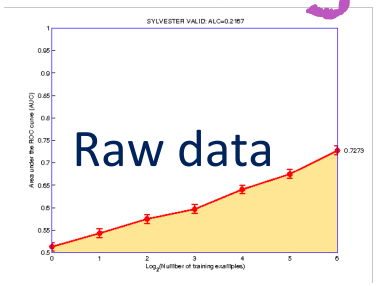


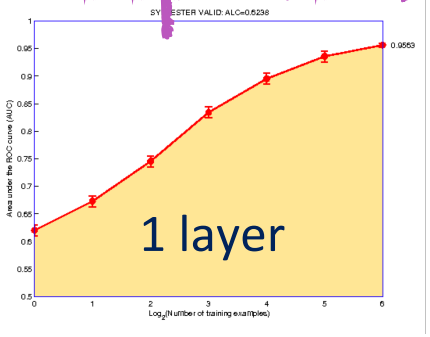
# How do humans generalize from very few examples?

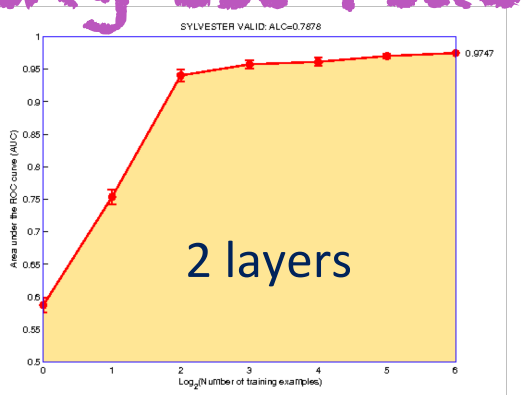
- They **transfer** knowledge from previous learning:
  - Representations
  - Explanatory factors
- Previous learning from: unlabeled data

  + labels for other tasks
- **Prior: shared underlying explanatory factors, in particular between P(x) and P(Y|x)**

# Unsupervised and Transfer Learning Challenge + Transfer Learning Challenge: Deep Learning 1st Place

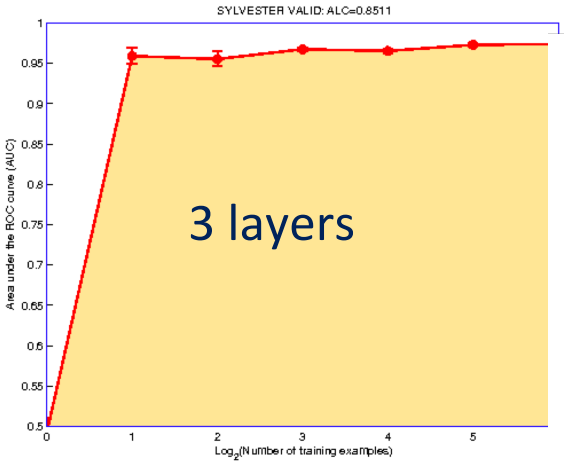


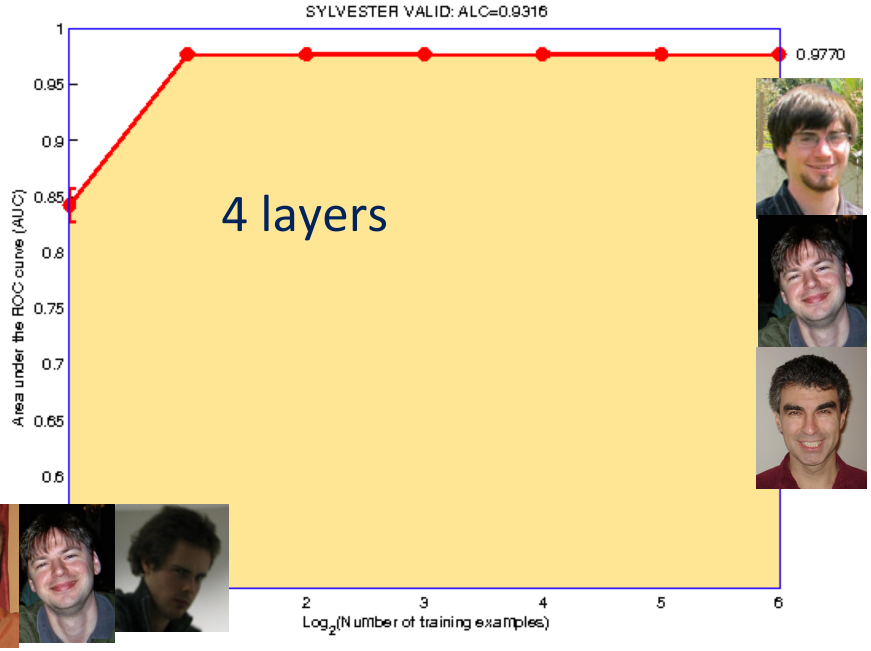




NIPS'2011 Transfer Learning Challenge Paper: ICML'2012

ICML'2011 workshop on Unsup. & Transfer Learning



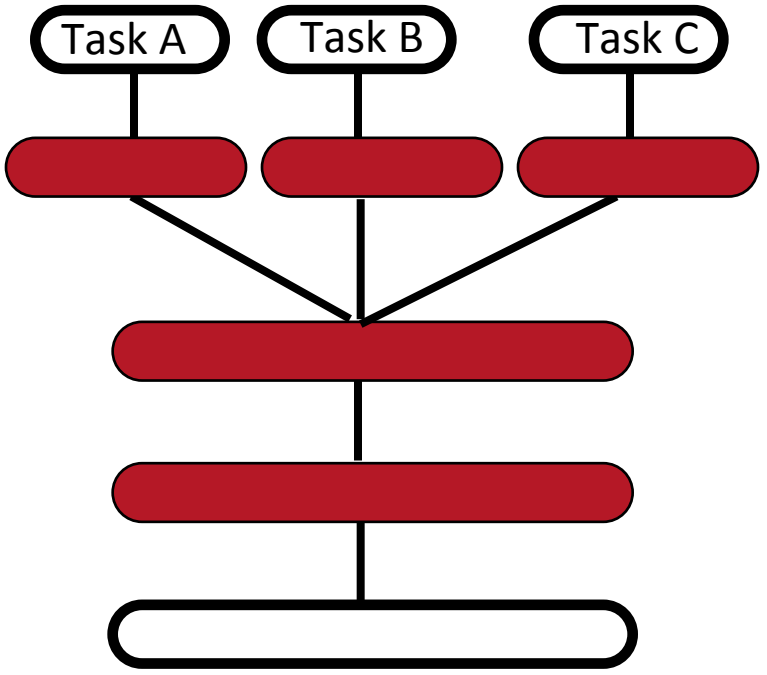






# Multi-Task Learning: Sharing Statistical Strength Across Tasks

- Generalizing better to new tasks (tens of thousands!) is crucial to approach AI
- Deep architectures learn good intermediate representations that can be shared across tasks

  (Collobert & Weston ICML 2008, Bengio et al AISTATS 2011)
- Good representations that disentangle underlying factors of variation make sense for many tasks because **each task concerns a subset of the factors**

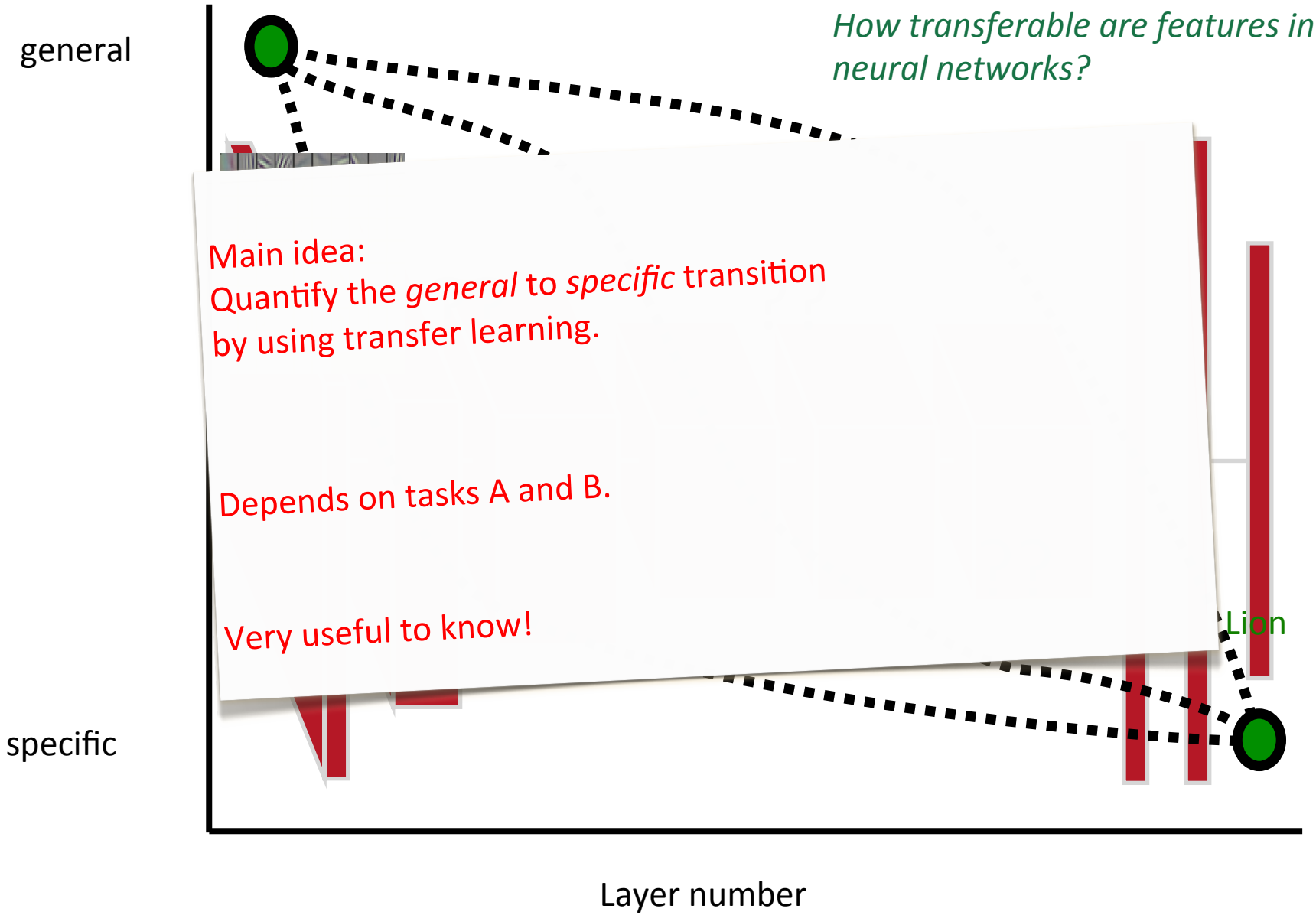


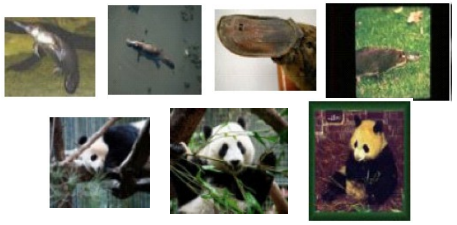

E.g. dictionary, with intermediate concepts re-used across many definitions

**Prior: shared underlying explanatory factors between tasks**

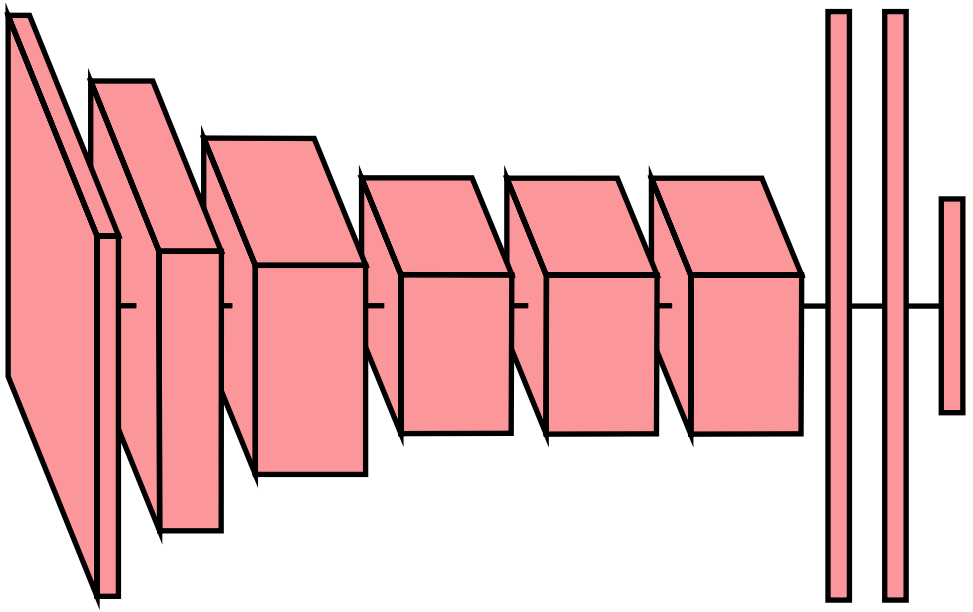
# Yosinski et al NIPS'2014

*How transferable are features in deep neural networks?*

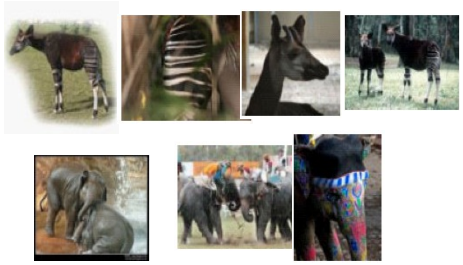




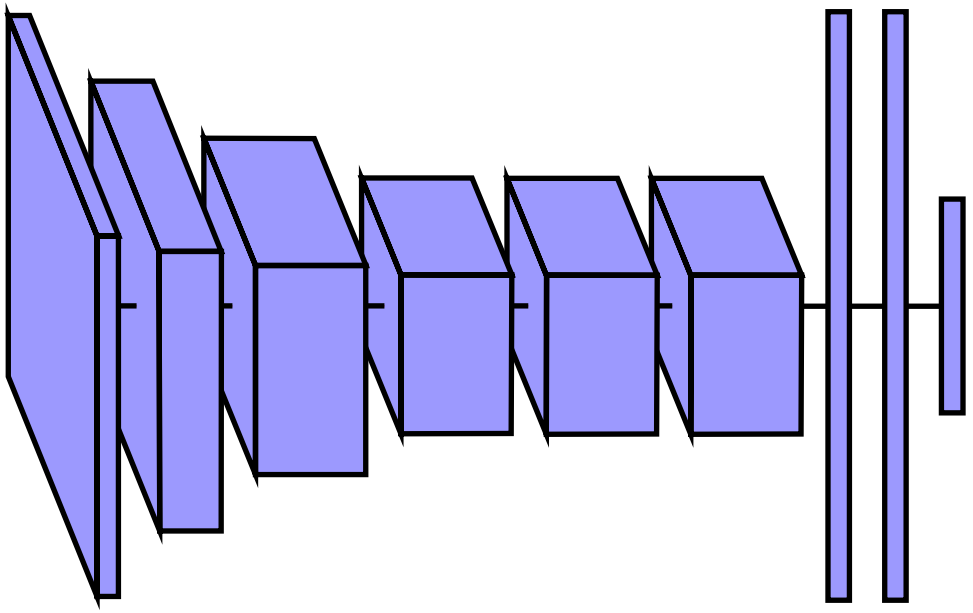
A Images



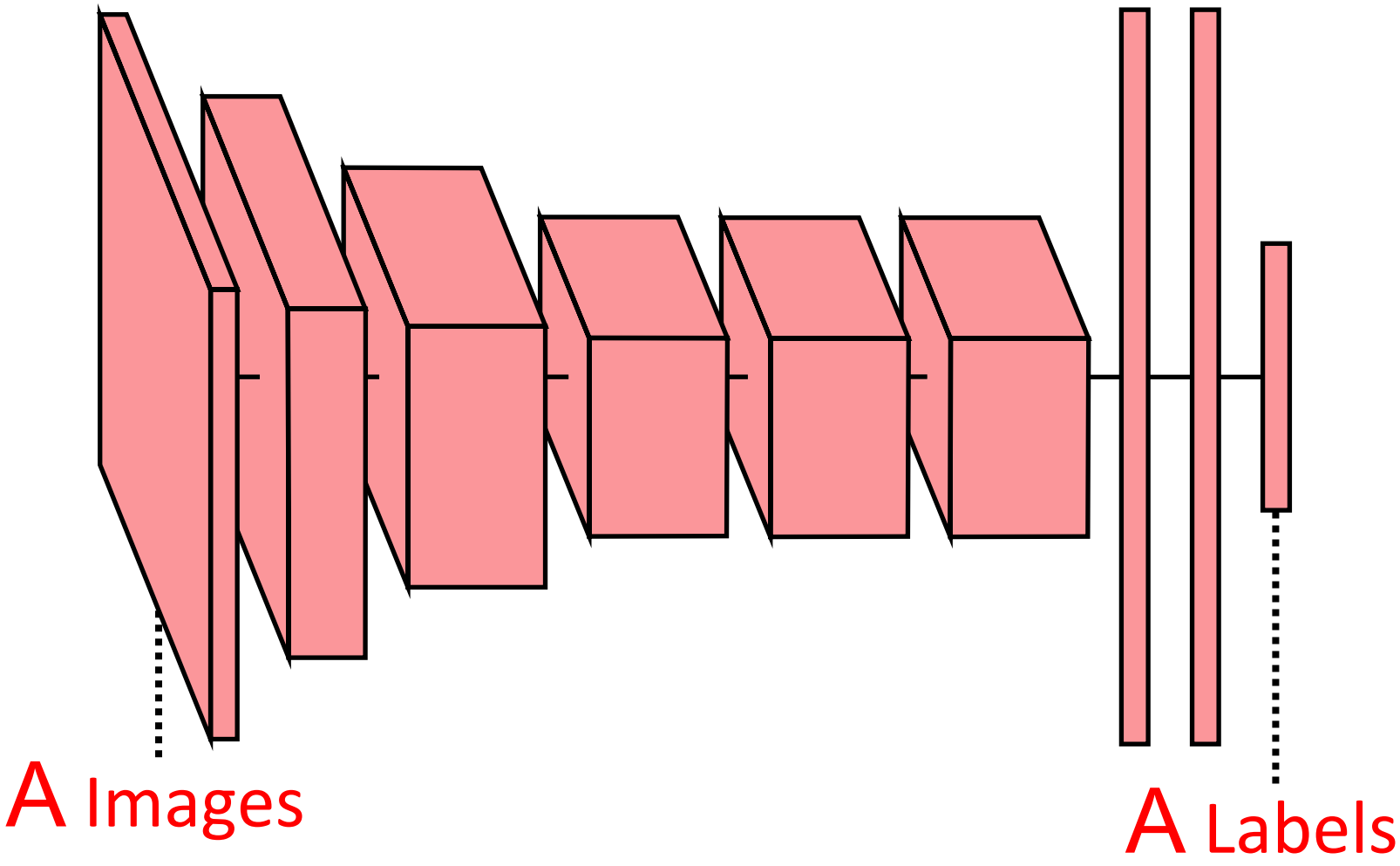
baseA

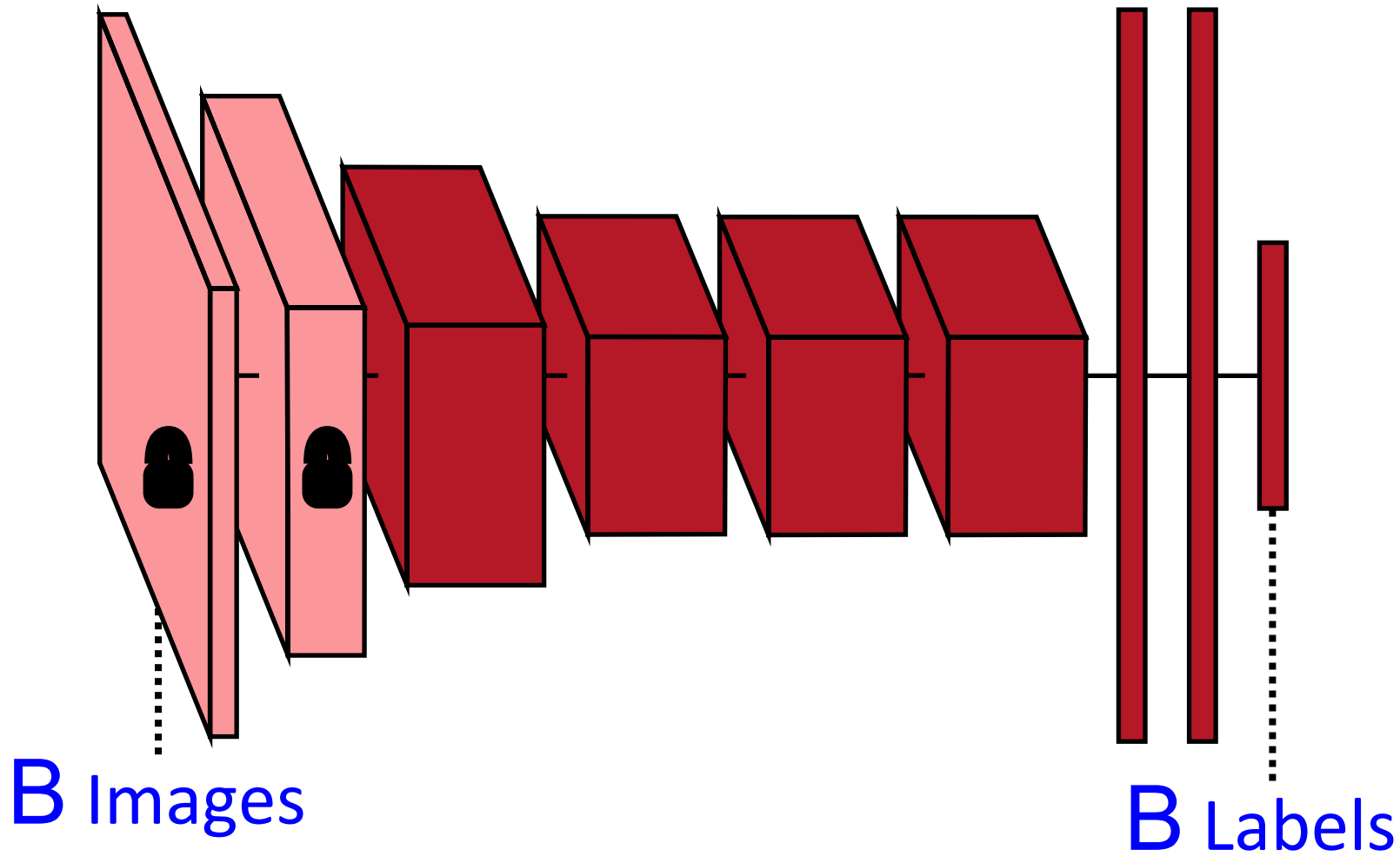


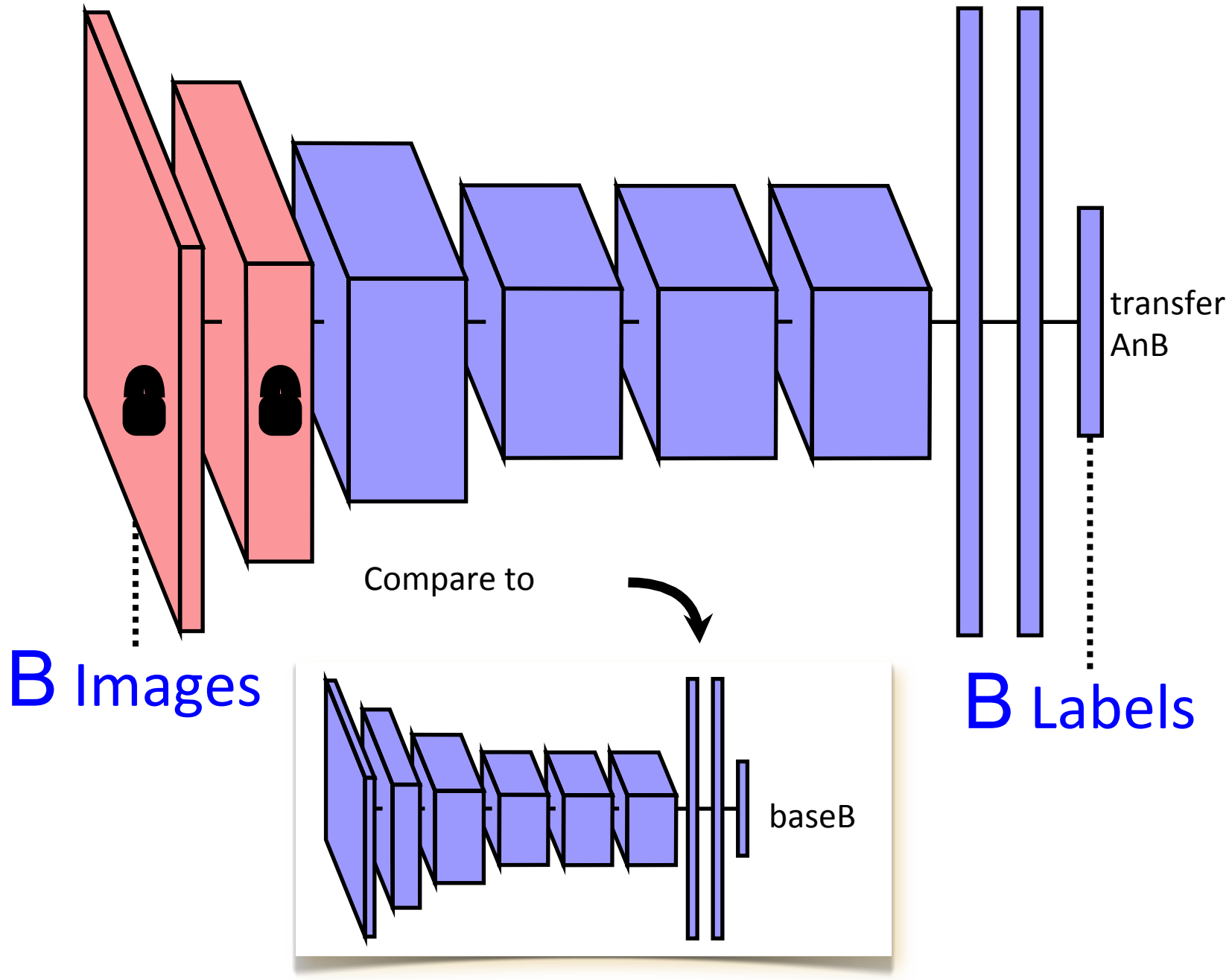
B Images

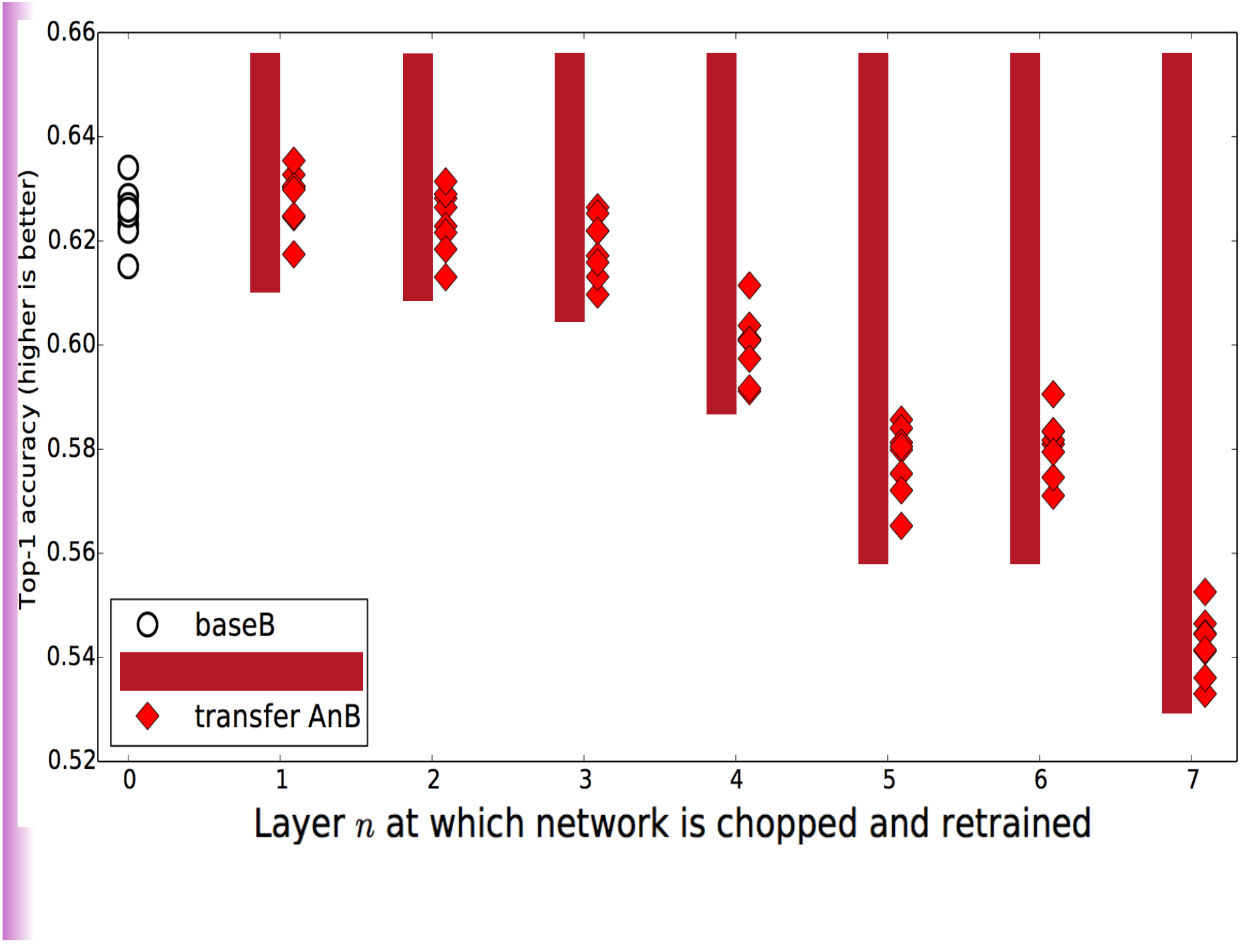


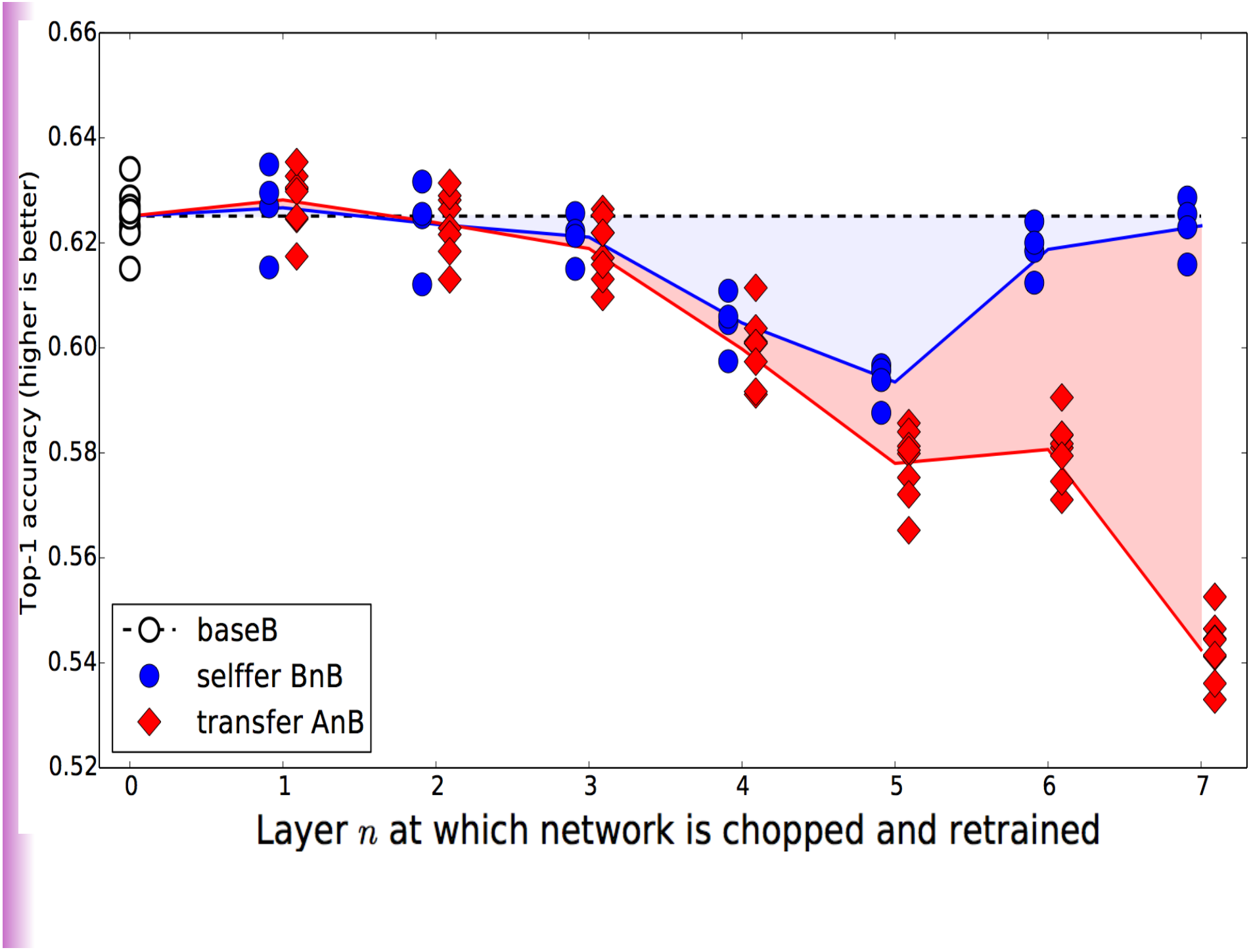
baseB

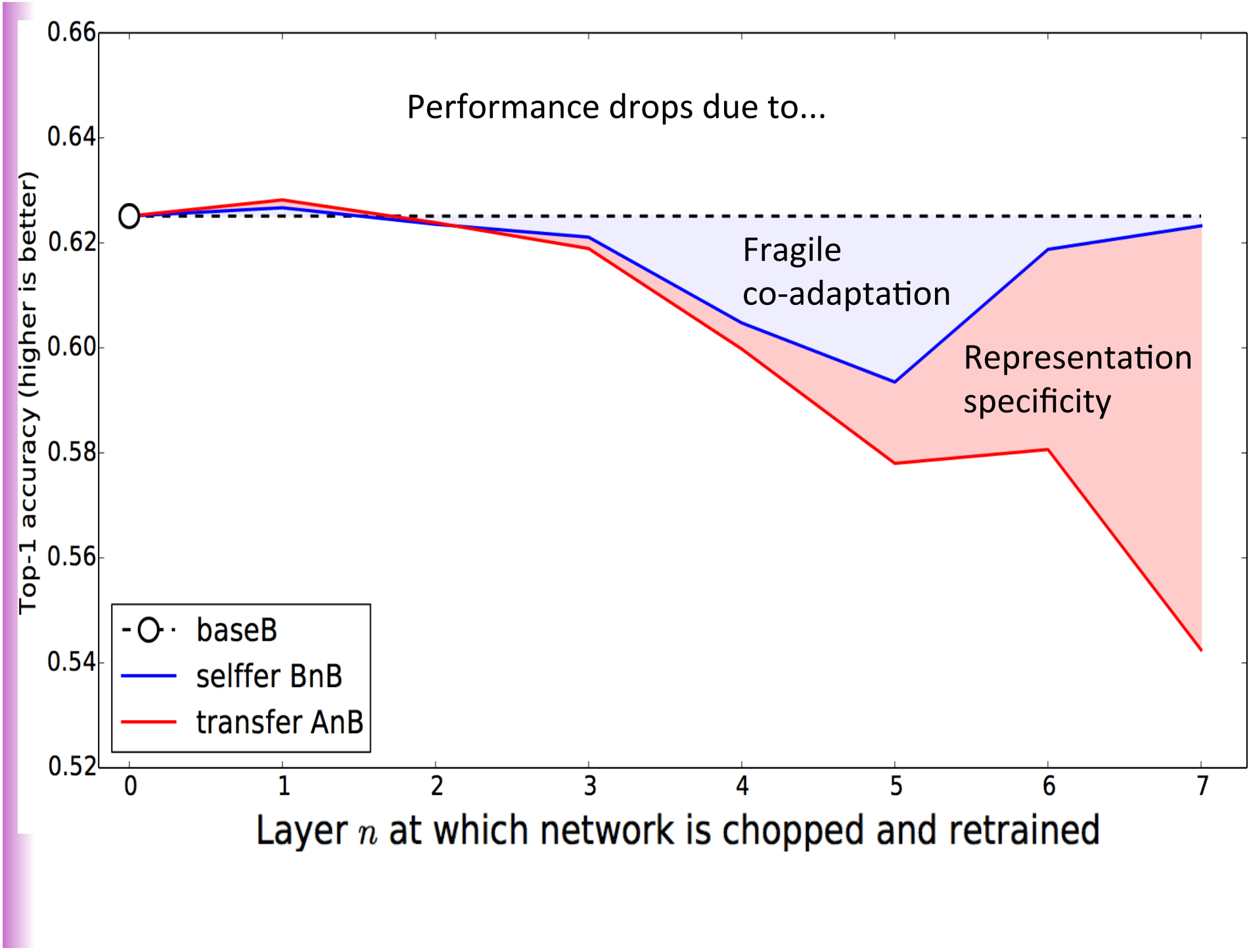

A Images
A Labels

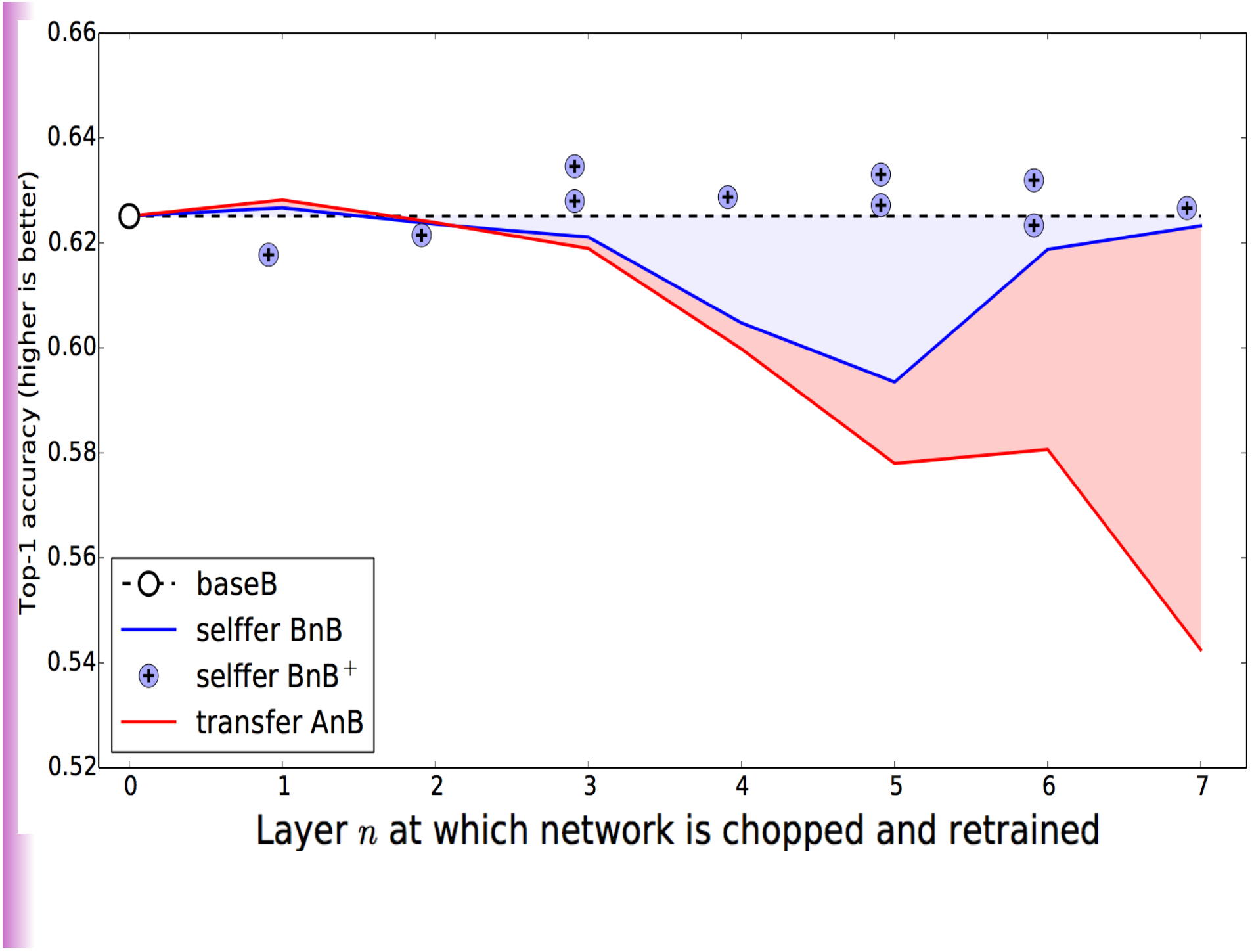

B Images
B Labels

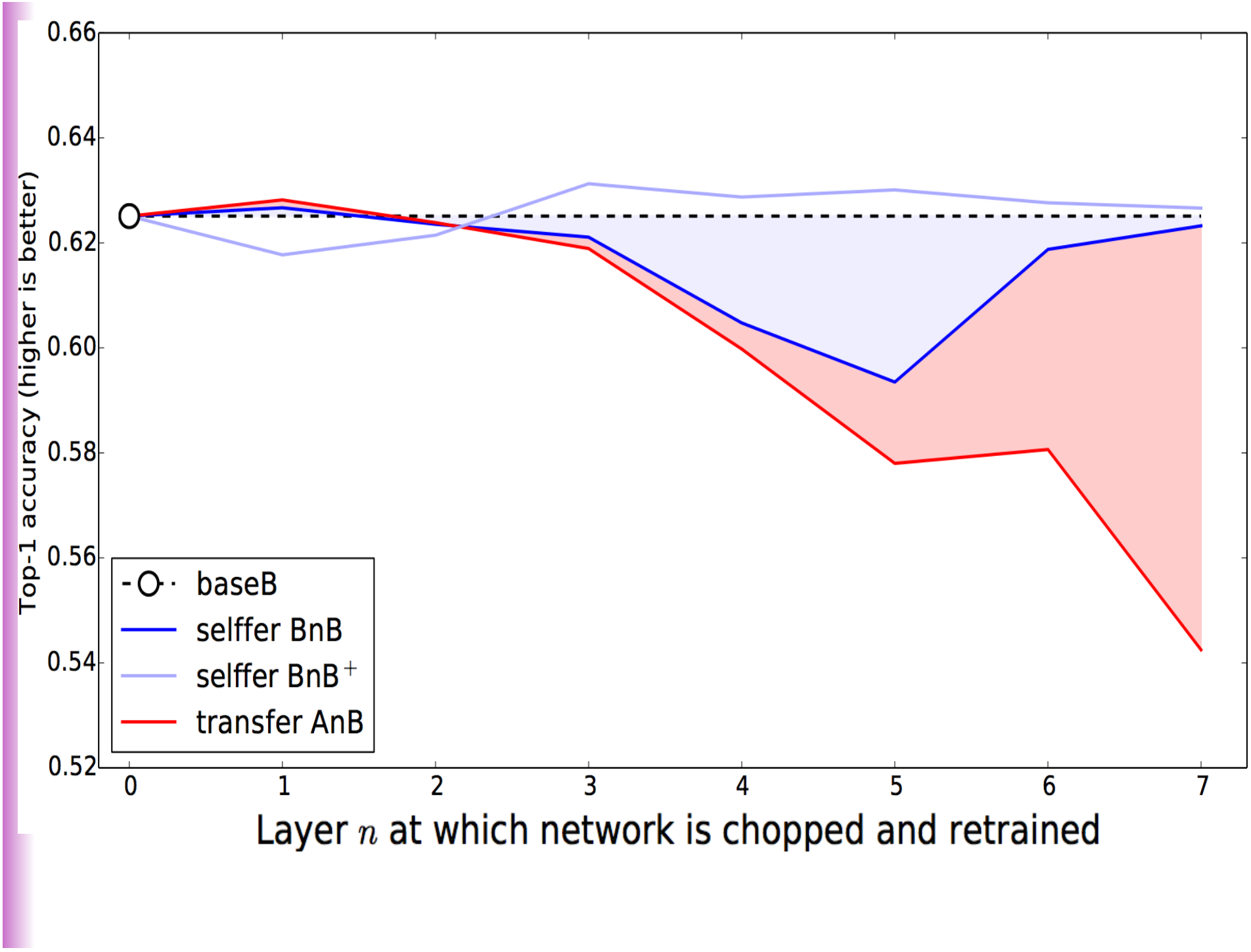

transfer
AnB
Compare to
B Images
B Labels
baseB

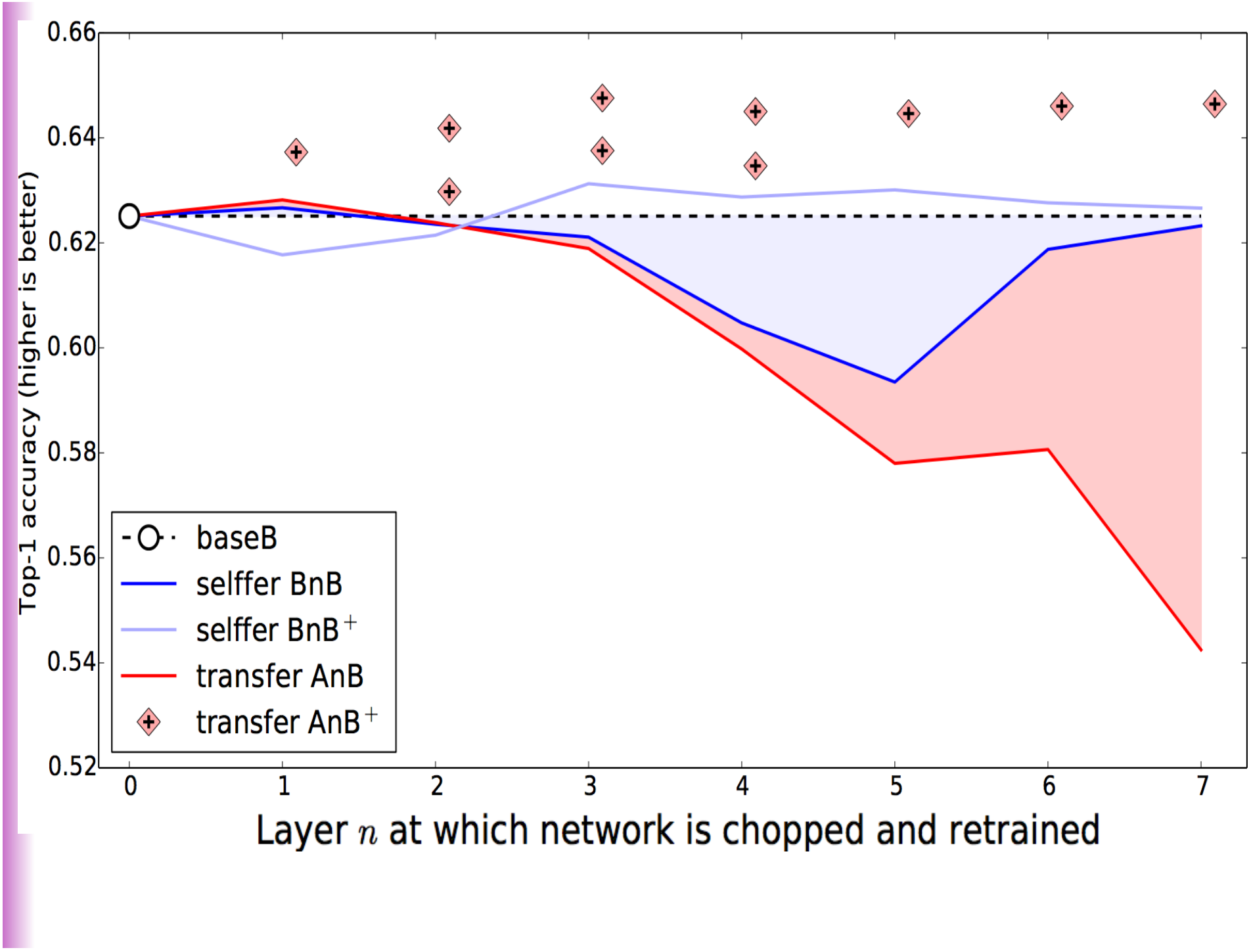

0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Top-1 accuracy (higher is better)
baseB
transfer AnB
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained

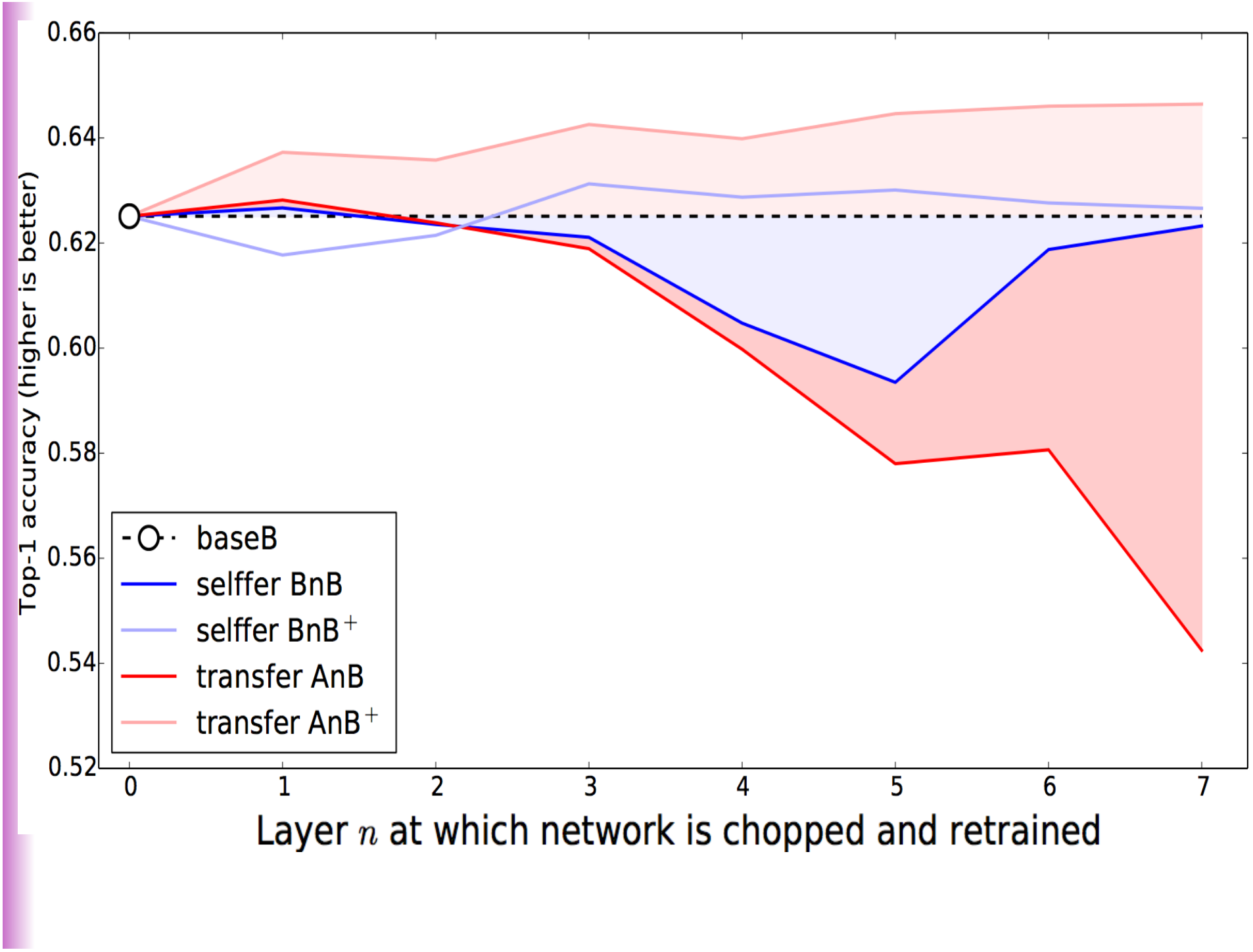

0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Top-1 accuracy (higher is better)
baseB
selffer BnB
transfer AnB
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained

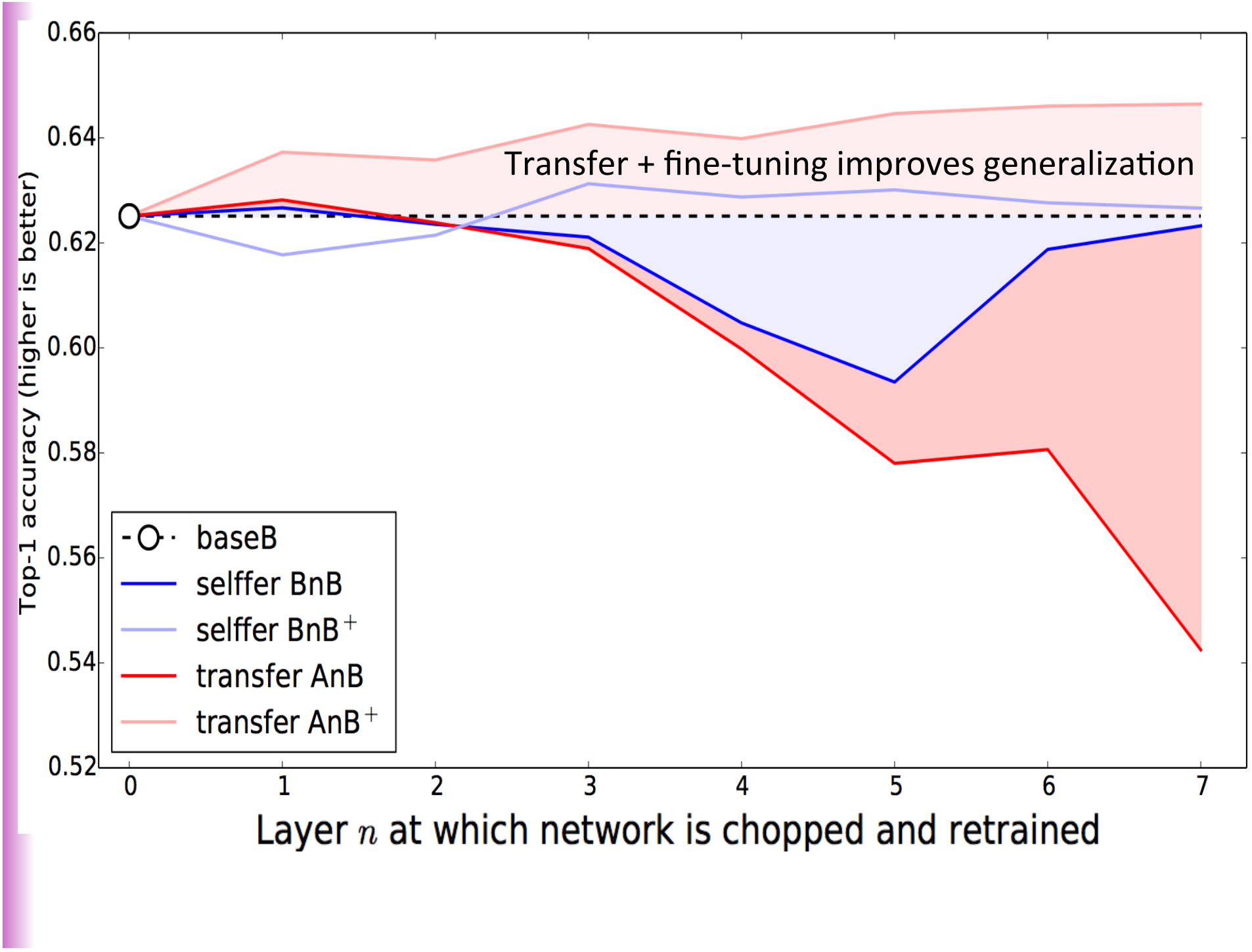

Performance drops due to...
Fragile
co-adaptation
Representation
specificity
baseB
selffer BnB
transfer AnB
Top-1 accuracy (higher is better)
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained


0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Top-1 accuracy (higher is better)
baseB
selffer BnB
selffer BnB+
transfer AnB
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained

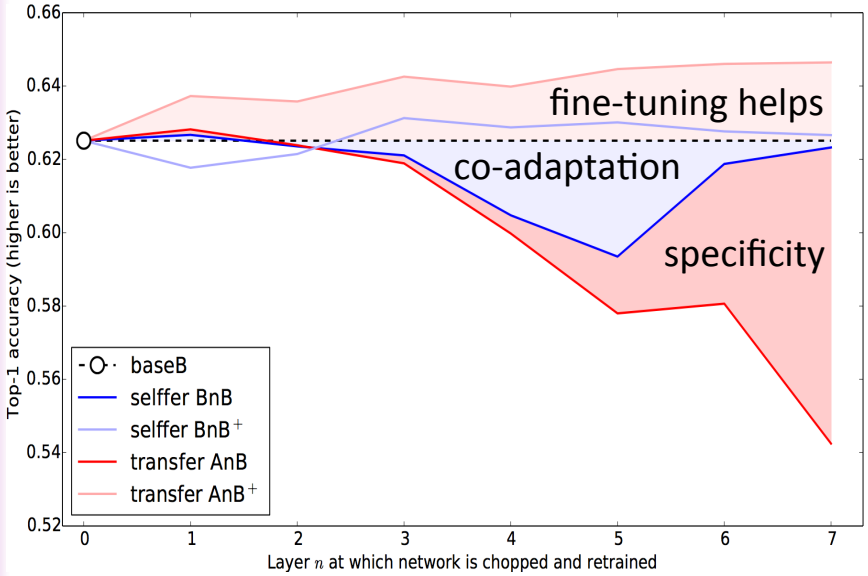

0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Top-1 accuracy (higher is better)
baseB
selffer BnB
selffer BnB+
transfer AnB
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained

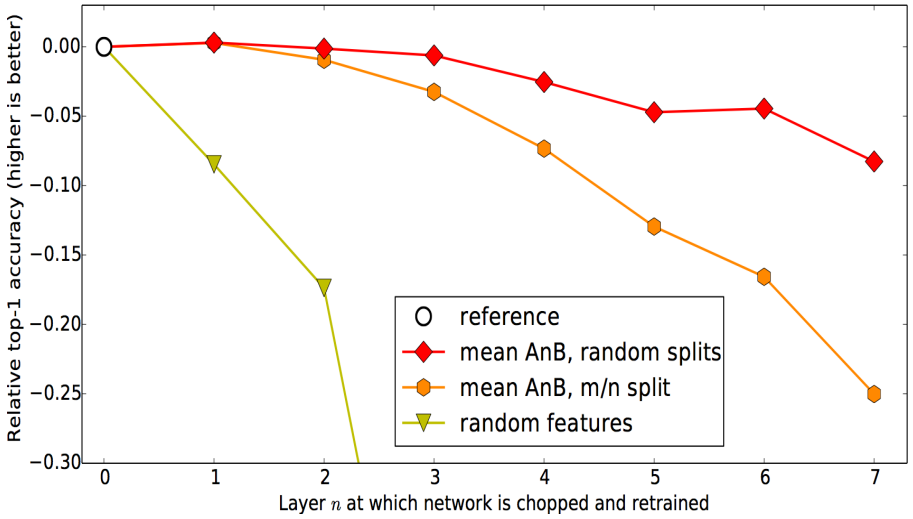

0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Top-1 accuracy (higher is better)
baseB
selffer BnB
selffer BnB+
transfer AnB
transfer AnB+
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained


0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Top-1 accuracy (higher is better)
baseB
selffer BnB
selffer BnB+
transfer AnB
transfer AnB+
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained


Transfer + fine-tuning improves generalization
Top-1 accuracy (higher is better)
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
baseB
selffer BnB
selffer BnB+
transfer AnB
transfer AnB+
0
1
2
3
4
5
6
7
Layer $n$ at which network is chopped and retrained

## Conclusions




- Measure *general* to *specific* transition layer by layer
- Transferability governed by:
  - lost co-adaptations
  - specificity
  - difference between base and target dataset
- Fine-tuning helps even on large target dataset

# Better Representations → Better Transfer → Better Domain Adaptation

- What is a good representation?

- Separate the « noise » from the « signal »
- Disentangle the underlying causal factors from each other

# Invariance and Disentangling

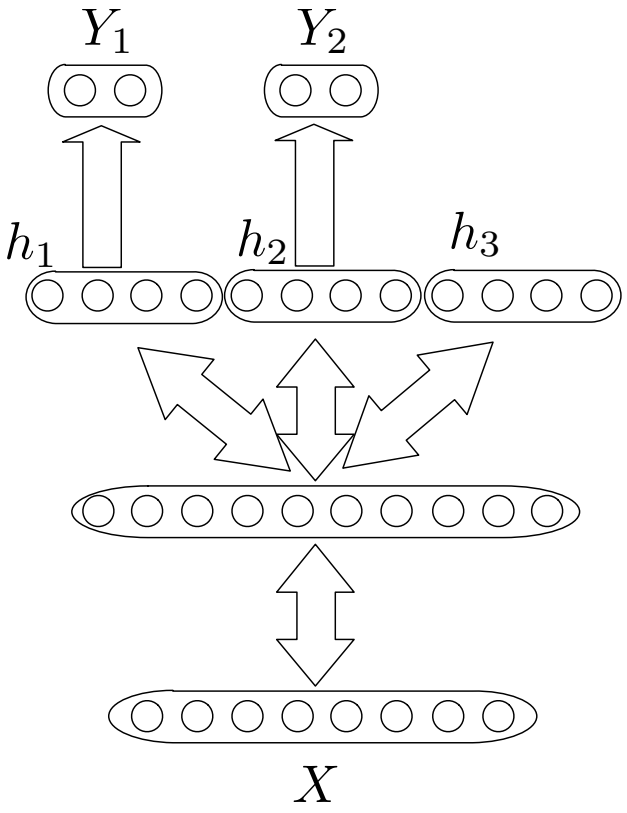
- Invariant features

- Which invariances?
- Alternative: learning to disentangle factors
- Good disentangling →

  avoid the curse of dimensionality

# Hints to Help Disentangling

- (Rifai et al, ECCV 2012, *Disentangling factors of variation for facial expression recognition*)
- (Kingma & Welling, NIPS 2014, *Semi-Supervised Learning with Deep Generative Models*)
- Some hidden units predict some of the factors, others are free to be used to reconstruct the input. Different groups of hidden units assigned to different factors. Orthogonality or penalty or independence prior between hidden units of different groups



# Broad Priors as Hints to Disentangle the Factors of Variation

- *Multiple factors*: distributed representations
- Multiple levels of abstraction: *depth*
- *Semi-supervised* learning: Y is one of the factors explaining X
- *Multi-task* learning: different tasks share some factors
- *Manifold* hypothesis: probability mass concentration
- Natural *clustering*: class = manifold, well-separated manifolds
- Temporal and spatial *coherence*
- *Sparsity*: most factors irrelevant for particular X
- *Simplicity* of factor dependencies (in the right representation)

# Emergence of Disentangling

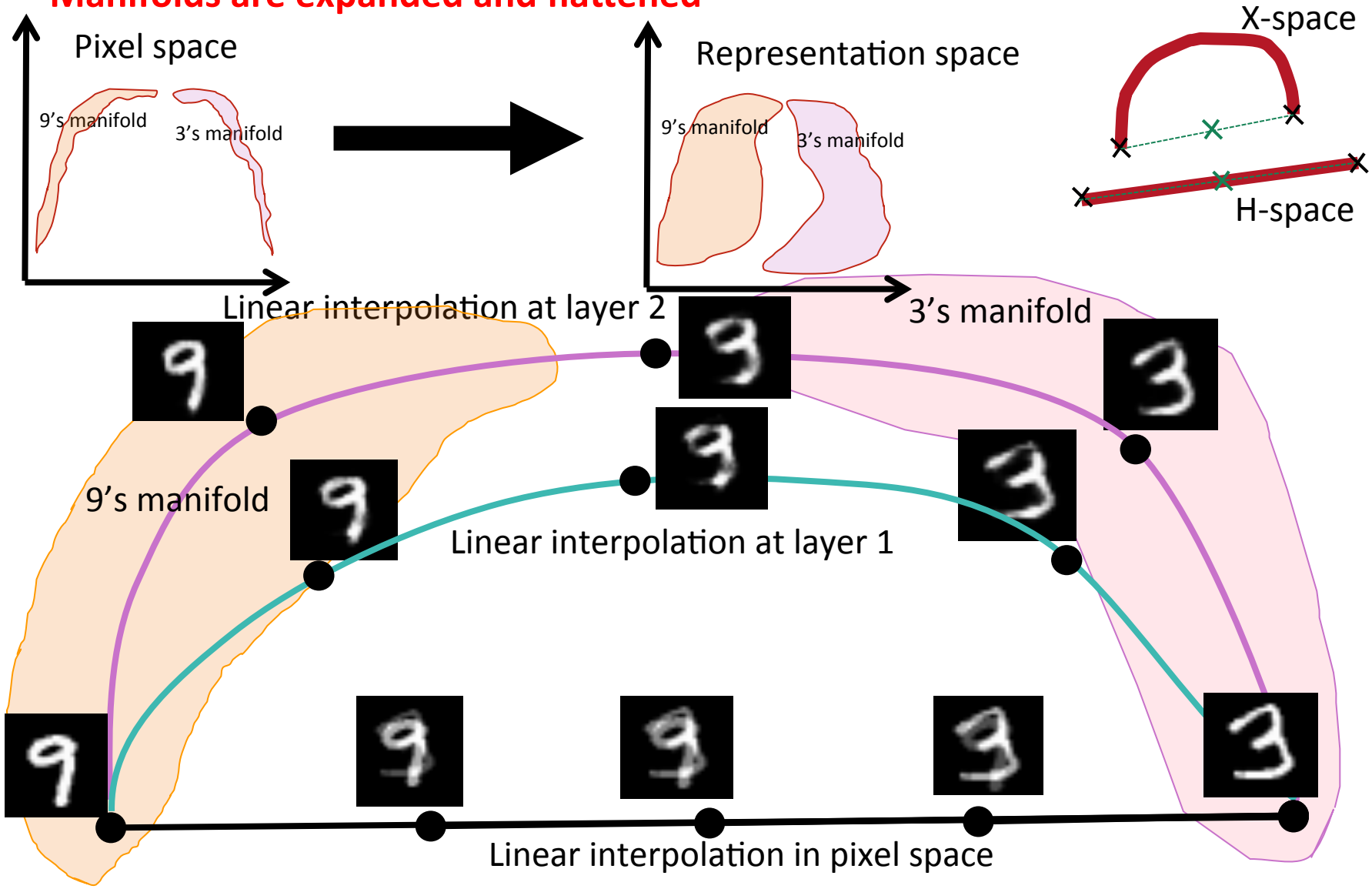
- (Goodfellow et al. 2009): sparse auto-encoders trained on images
  - some higher-level features more invariant to geometric factors of variation

- (Glorot et al. 2011): sparse rectified denoising auto-encoders trained on bags of words for sentiment analysis
  - different features specialize on different aspects (domain, sentiment)

**WHY?**

# Space-Filling in Representation-Space

- **Deeper representations → abstractions → disentangling**
- **Manifolds are expanded and flattened**

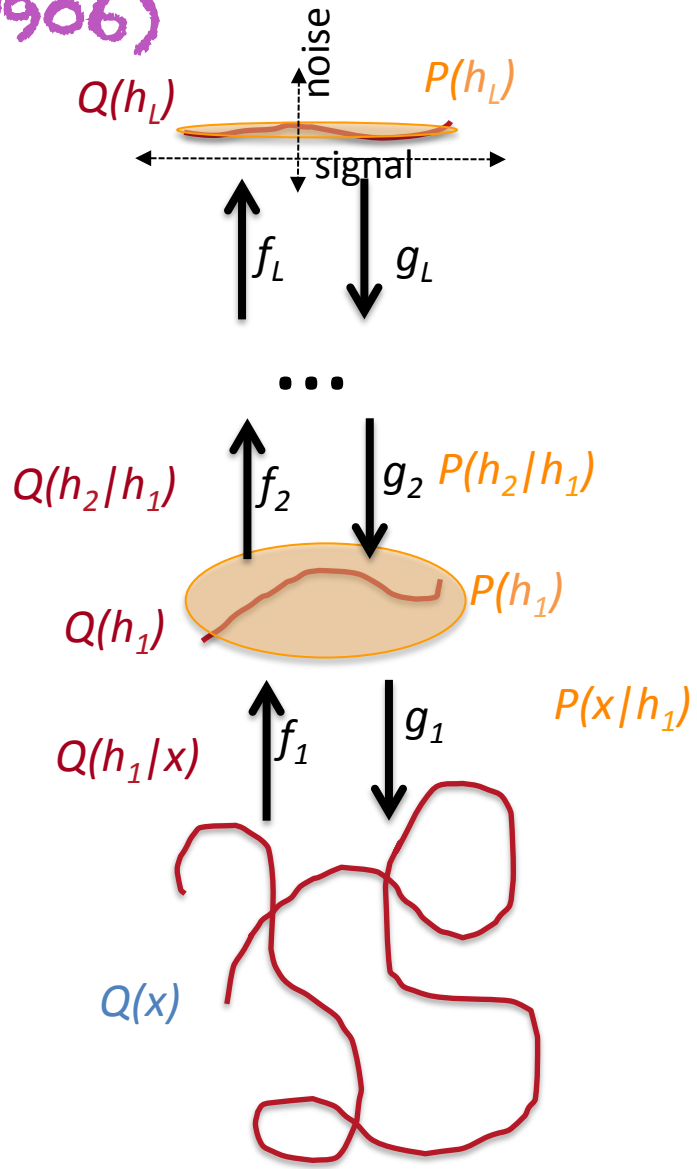


# Extracting Structure By Gradual Disentangling and Manifold Unfolding (Bengio 2014, arXiv 1407.7906)

Each level transforms the data into a representation in which it is easier to model, unfolding it more, contracting the noise dimensions and mapping the signal dimensions to a factorized (uniform-like) distribution.

$$\min KL(Q(x,h)||P(x,h))$$

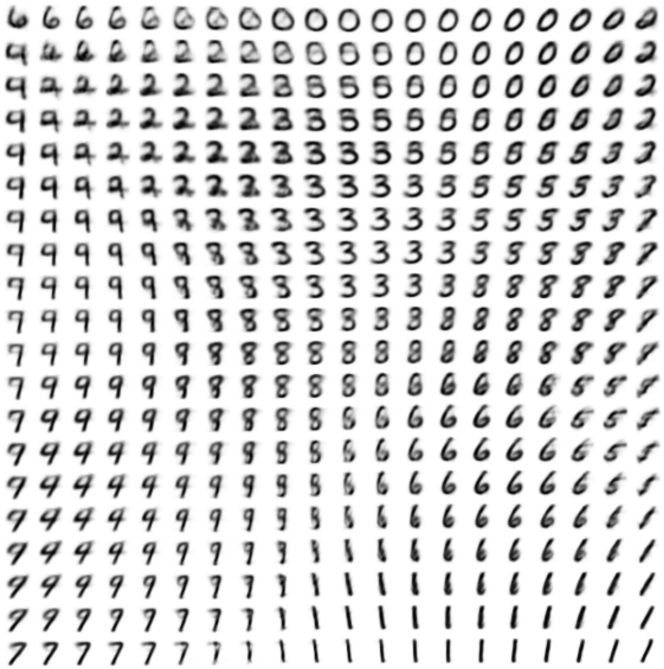
for each intermediate level h



# Variational Auto-Encoder: Random Sampling at Top Level

- Models trained with the KL(Q||P) or VAE training objective
- Randomly sample from 2-D top-level h (Gaussian), project down:

(from Kingma & Welling ICLR 2014)



# Deep Directed Generative AEs

(Ozair & Bengio 2014, arXiv 1410.0630)

- log P(x) >= log P(x|h=f(x)) + log P(h=f(x))

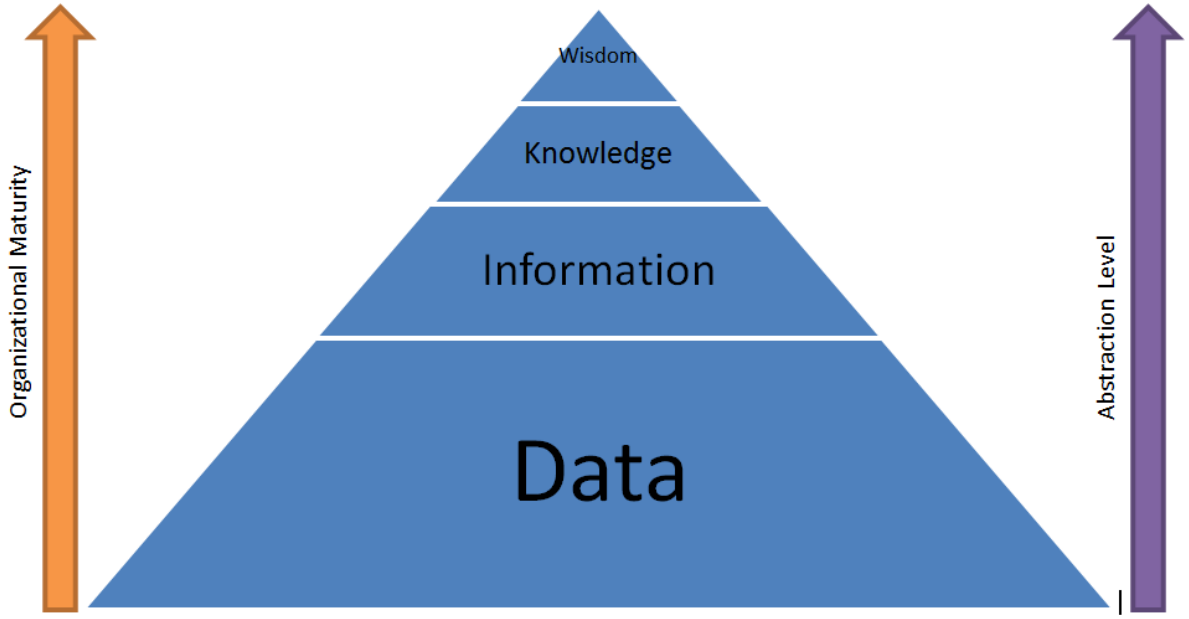
= bound that is maximized and becomes tight as training progresses

- Stacking such auto-encoders yields representations that become sparser and with less correlation between features

| Samples | Entropy | Avg # active bits | $\|\|Corr - diag(Corr)\|\|_F$ |
|---|---|---|---|
| Data $(X)$ | 297.6 | 102.1 | 63.5 |
| Output of $1^{st}$ encoder $(f_1(X))$ | 56.9 | 20.1 | 11.2 |
| Output of $2^{nd}$ encoder $(f_2(f_1(X)))$ | 47.6 | 17.4 | 9.4 |

# Conclusion: Learning Multiple Levels of Abstraction

- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions disentangle the factors of variation, which allows much easier generalization and transfer



# MILA: Montreal Institute for Learning Algorithms