Understanding Neural Networks

Speaker: Yi-Shan Wu

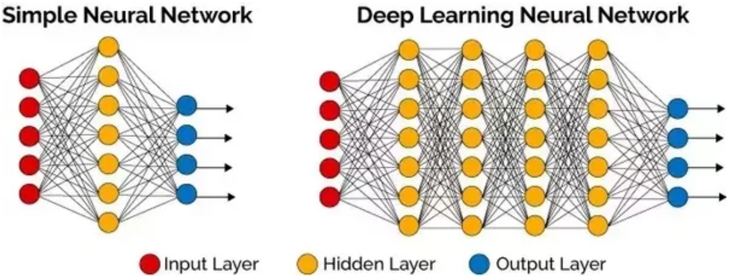
Institute of Information Science

Academia Sinica

Taiwan

Feb 22, 2019

What do neural networks learn ?



- such that they can usually generalize
- such that transfer learning is possible
- ...

1. How transferable are features in deep neural networks ? (NIPS 2014)
2. Convergent Learning : Do different neural networks learn the same representations ? (ICLR 2016)
3. Towards Understanding Learning Representations : To what extent do different neural networks learn the same representation ? (NIPS 2018)

Selected Authors

- Jason Yosinski : Ph.D. at Cornell, now at Uber AI Lab
  - understanding neural networks and figuring out how to make them better
- John E. Hopcroft : 79 years old, 1986 Turing Awards
- Liwei Wang : several papers related to generalization issues

Neural networks are found to learn general features on the first layer.

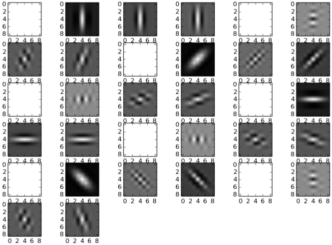


FIGURE – Gabor filter

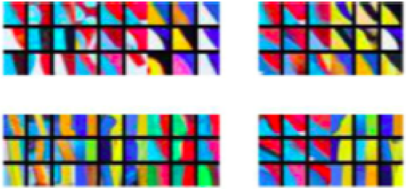


FIGURE – Color blobs

- When to change from **general** to **specific** for a particular dataset ?
- Can we use the **general** part of NN for transfer learning ?

Possible benefits of transfer learning

1. When the target dataset is significantly small, transfer learning can be a powerful tool to prevent overfitting.
2. We can learn a new task with just few training samples. (testing error of the new task is close to its training error)
3. Save time and memories

Can we quantify or even measure the degree to which a particular layer is general or specific ?

How general ?

1. Universally general ? (the feature can be applied to any task)
2. some particular features often appear (Use same network architecture, some particular features appear even with different initialization.)
3. Transferable to the same task ? (Use same network architecture, the learned feature can be applied to other samples in the same task.)
4. Transferable to other tasks ? (transferable to several/particular tasks such that it can perform at least as good as relearn the task)

Convergent learning

Use the same network architecture, would features learned by multiple networks converge to a set of particular features ?
(one-to-one mapping or span similar low-dimensional subspaces or neither of them)

Basic Settings

- same architecture (5 convolutional layers + 3 fully-connected)
- same training dataset (ImageNet)
- different random initializations (results in Net1, Net2)

Quantify similarity

Calculate correlation of two random variables, $X^n_{\ell,i}$ and $X^m_{\ell,j}$.

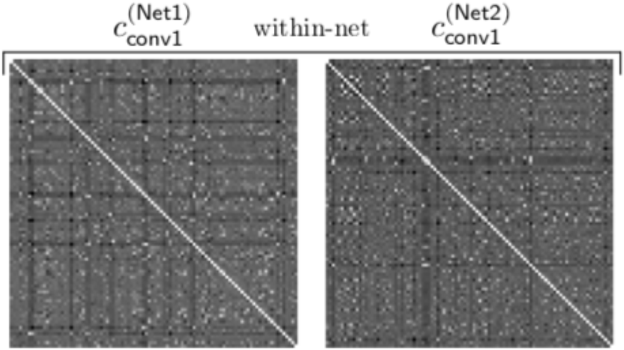
- n, m : Net$n$ and Net$m$ (n, m can be the same)
- ℓ : layer $\ell \in$ {conv1, conv2, conv3, conv4, conv5, fc6, fc7}
- i, j : i's activation unit in layer ℓ v.s. j's activation unit in layer ℓ (i, j can be the same)
- unit : hidden nodes for fully-connected layers while sum over values of all spatial positions of a filter/channel.

ex. Mean : $\mu^n_{\ell,i} = \mathbb{E}[X^n_{\ell,i}]$ (aggregating over the validation set)

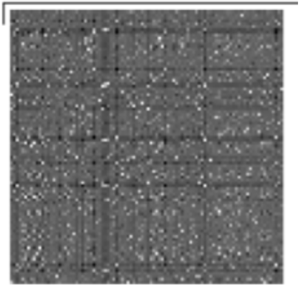
- within-net correlation : $\mathbb{E}[(X^n_{\ell,i} - \mu^n_{\ell,i})(X^n_{\ell,j} - \mu^n_{\ell,j})]/\sigma^n_{\ell,i}\sigma^n_{\ell,j}$
- between-net correlation : $\mathbb{E}[(X^n_{\ell,i} - \mu^n_{\ell,i})(X^m_{\ell,j} - \mu^m_{\ell,j})]/\sigma^n_{\ell,i}\sigma^m_{\ell,j}$

Results for within-net correlation

Correlation value serves as a way of measuring how related the activations of one unit are to another unit, either within the network or between networks.



Results for between-net correlation



natural (unaligned) order

Is there a one-to-one alignment between features learned by different neural networks ?

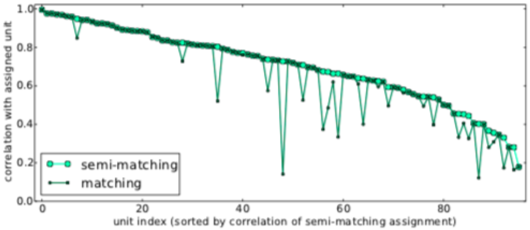
If we allow ourselves to permute the units of one network, to what extent can we find equivalent or nearly-equivalent units across networks ?

between-net correlation

bipartite semi-matching

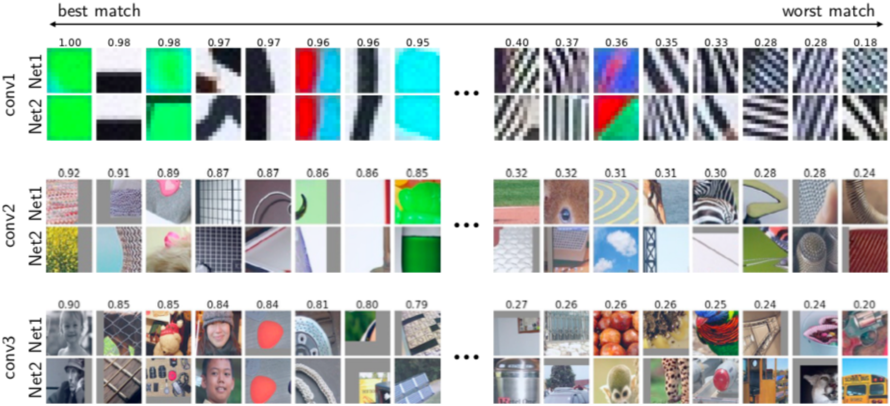
For each unit in Net1, we find the unit in Net2 with maximum correlation to it.

Net1	1	2	3	4	5	6	7	8	9	10
Net2	5	2	6	1	8	4	2	9	6	7
	0.9	0.82	0.7	0.8	1	0.95	0.82	0.7	0.97	0.78



bipartite semi-matching

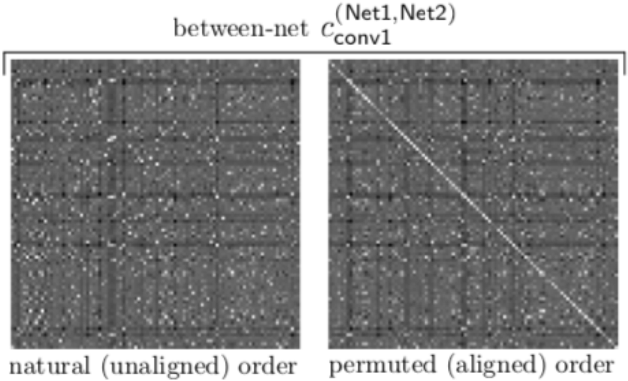
Plot the image patch from the validation set that causes the highest activation for that unit.



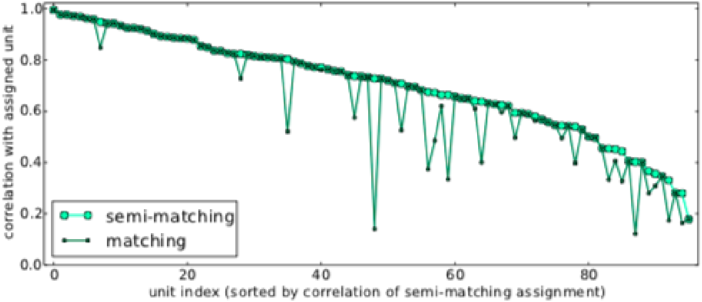
between-net correlation

bipartite matching

Find a one-to-one assignment between Net1 and Net2 such that the correlation is maximized.

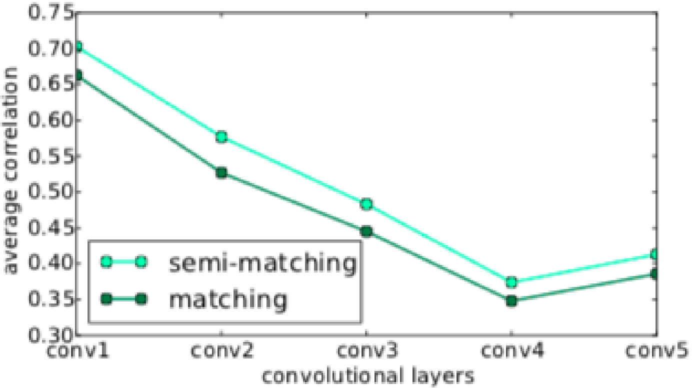


between-net correlation



Correlations between paired conv1 units in Net1 and Net2.

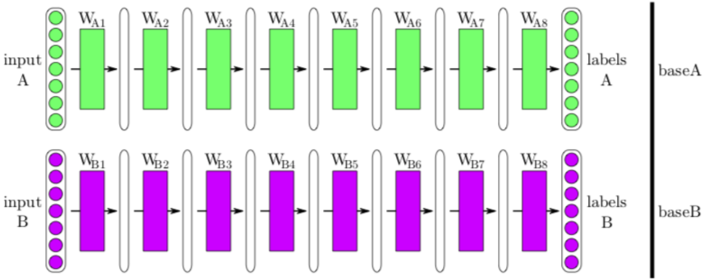
between-net correlation



Average correlations between paired conv1 units in Net1 and Net2.

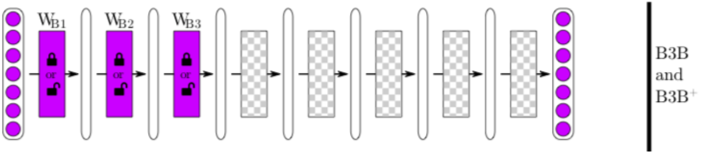
Experiments

- Prepare non-overlapping task A and task B (can be either similar or not, which will be specified later)
- Two eight-layer CNN on A and B respectively.

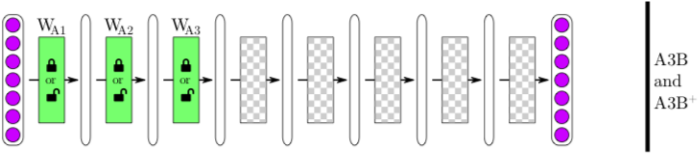


Transfer networks example

- **selffer** network BnB : the first *n* layers are copied from baseB and frozen. The higher layers are initialized randomly and trained on dataset B.



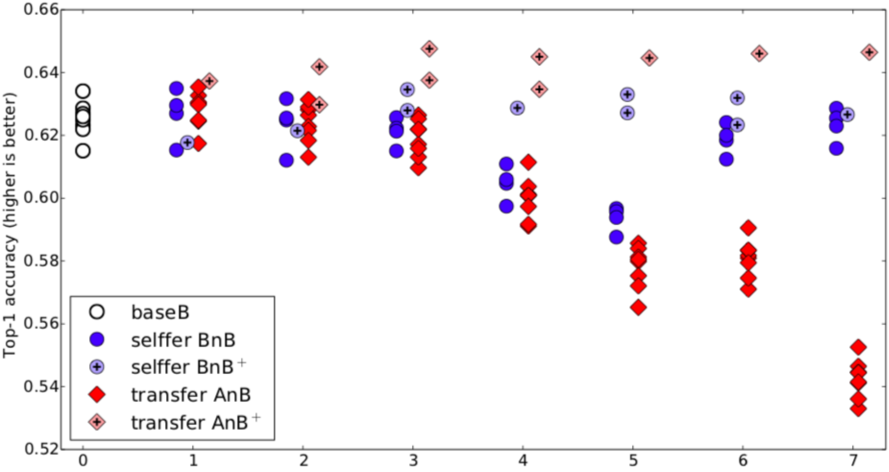
- **transfer** network AnB : the first *n* layers are copied from baseA and frozen. The higher layers are initialized randomly and trained on dataset B.



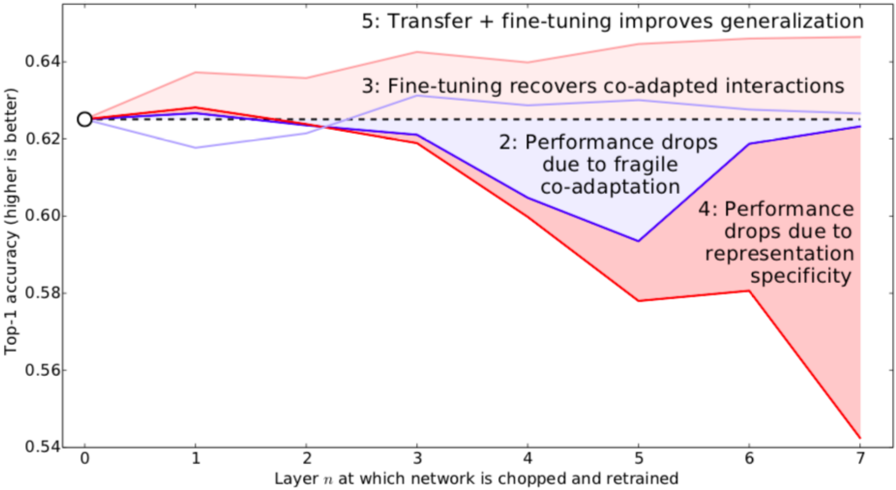
Datasets

1. for transferable to the similar task :
   Randomly assign half of the 1000 ImageNet classes to A and half to B. (ImageNet contains clusters of similar classes, ex : several kinds of dogs and cats)
2. for transferable to different tasks :
   551 classes in the man-made group and 449 in the natural group

Results for transferable to the similar task



Results for transferable to the similar task



Discussion

- The first two layers are general for they can be transfered almost perfectly from A to B.
- AnB^{+} performs extraordinary well.
- B4B, B5B exhibit worse performance. (the reason is unknown) (Authors : The original network contained fragile co-adapted features on successive layers, that is, features that interact with each other in a complex or fragile way such that this co-adaptation could not be relearned by the upper layers alone.)
- AnB might caused by two effects : the drop from lost co-adaptation and the drop from features that are less and less general

Results for transferable to a different task

- task A : man-made
- task B : natural

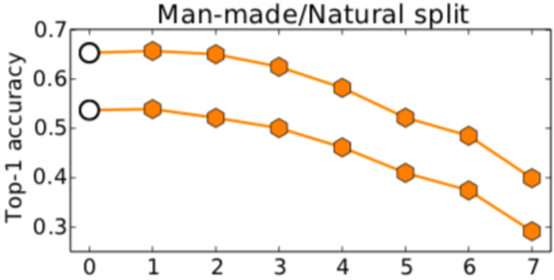
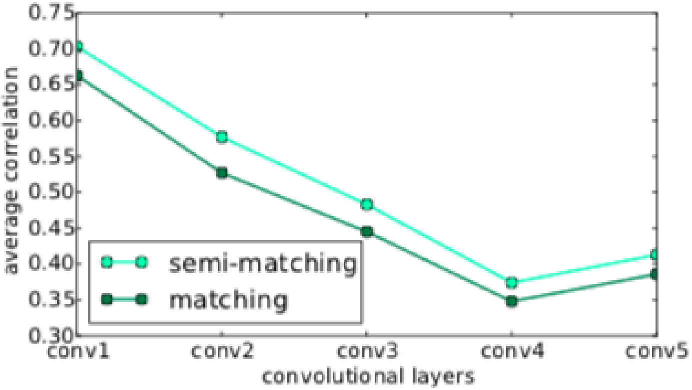


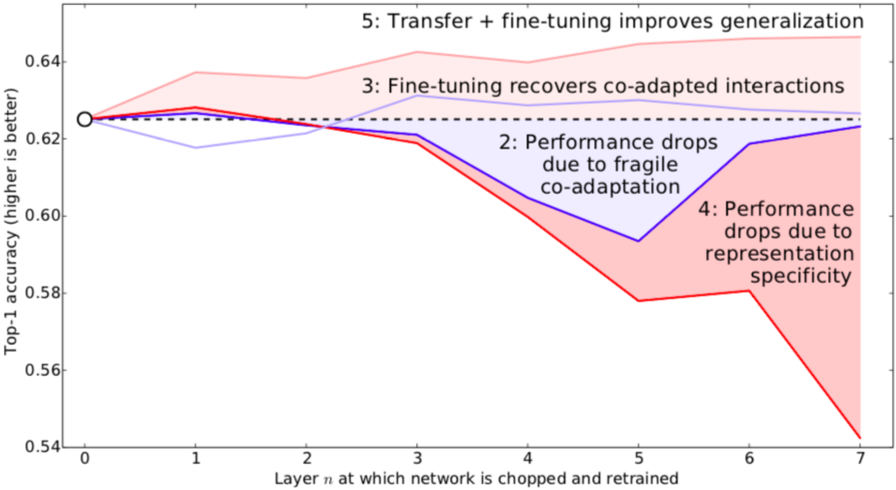
FIGURE – upper : baseB and AnB

Although the two tasks now are less similar, the first two layers are still quite general.

Recap two results



Recap two results



References

1. How transferable are features in deep neural networks ? (NIPS 2014)
2. Convergent Learning : Do different neural networks learn the same representations ? (ICLR 2016)
3. SVCCA : Singular Vector Canonical Correlation Analysis for Deep Understanding and Improvement
4. Towards Understanding Learning Representations : To what extent do different neural networks learn the same representation ? (NIPS 2018)