# Deep Visual Models with Interpretable Features and Modularized Structures

Quanshi Zhang

**John Hopcroft Center**

Shanghai Jiao Tong University

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu, "Interpretable Convolutional Neural Networks" in CVPR (Spotlight) 2018
Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu, "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI, 2018
Quanshi Zhang, Yu Yang, Yuchen Liu, Ying Nian Wu, and Song-Chun Zhu, "Unsupervised Learning of Neural Networks to Explain Neural Networks" extended abstract in AAAI-19 Workshop on Network Interpretability for Deep Learning, 2019
Quanshi Zhang, Yu Yang, Qian Yu, Ying Nian Wu, and Song-Chun Zhu, "Network Transplanting" extended abstract in AAAI-19 Workshop on Network Interpretability for Deep Learning, 2019

# Explanations → Trustiness & diagnosis

- How to make human beings trust a computer?

**Computer:** We must make a surgery on your head?
**Human:** Why should I trust you and let you cut my head
**Computer:** It is because
1) Filter 1 detected a lesion in Organ A
2) Filter 2 detected a lesion in Organ B
…

## Quantitative explanation

An accident happed.

Human: tell me the reason for road planning before the traffic accident.

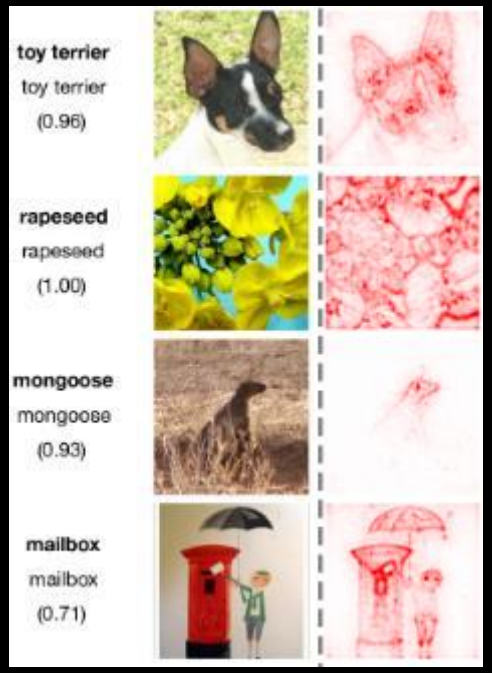
**Computer:** It is because
1) Filter 1 detected a tree
2) Filter 2 detected a person
3) Filter 3 detected the road
4) Filter 4 detected **another road**
…

Human: I find Filter 4 considers a river as a road.

Fix representation flaws in the CNN

# Network visualization & diagnosis

Visualization of appearance encoded by a filter




Pixels related to the final prediction output

Can only visualize salient information

The key problem is to explain most information (e.g. 70%--90%) in a network

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. Distill, 2017. https://distill.pub/2017/feature-visualization.

Pieter-Jan Kindermans, Kristof T. Sch¨utt, Maximilian Alber, Klaus-Robert M¨uller, Dumitru Erhan, Been Kim, and Sven D¨ahne. Learning how to explain neural networks: Patternnet and patternattribution. In arXiv: 1705.05598, 2017.

# Deep learning, a science or a technology?

Deep neural network → a piecewise linear model → unexplainable
→ We will never get accurate explanation for 100% information of a DNN

Alchemy?

- Explain features in intermediate layers
- Semantically
- Quantitatively
  - **What patterns are learned**
  - **Given an image, which patterns are triggered.**
  - **E.g. 90% information is interpretable**
    - **83% represents object parts**
    - **7% represents textures**
    - **10% cannot be interpreted**

# Outline

- How to represent CNNs using semantic graphical models
- How to learn disentangled, interpretable features in middle layers
- How to boost interpretability without hurting the discrimination power
- How to learn networks with functionally interpretable structures

# Outline

- **How to represent CNNs using semantic graphical models**
- How to learn disentangled, interpretable features in middle layers
- How to boost interpretability without hurting the discrimination power
- How to learn networks with functionally interpretable structures

# Background: Learning explanatory graphs for CNNs

- Given a CNN that is pre-trained for object classification
  - **How many types of visual patterns are memorized by a convolutional filter of the CNN?**

Distribution of activations in a feature map

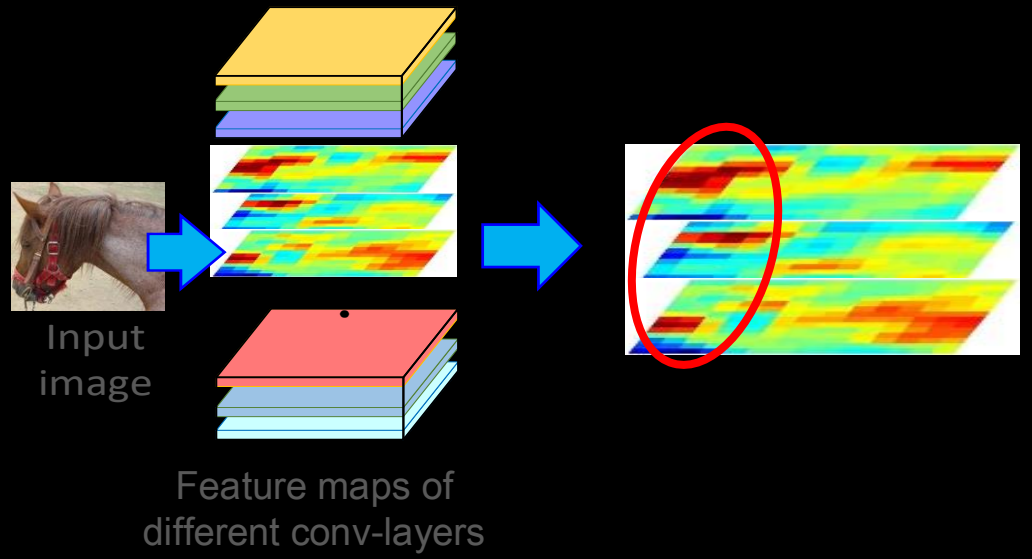
A head pattern

??? pattern


Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Background: Learning explanatory graphs for CNNs

- Given a CNN that is pre-trained for object classification
  - How many types of visual patterns are memorized by a convolutional filter of the CNN?
  - **Which patterns are co-activated to describe a part?**




These filters are co-activated in certain area to represent the head of a horse.

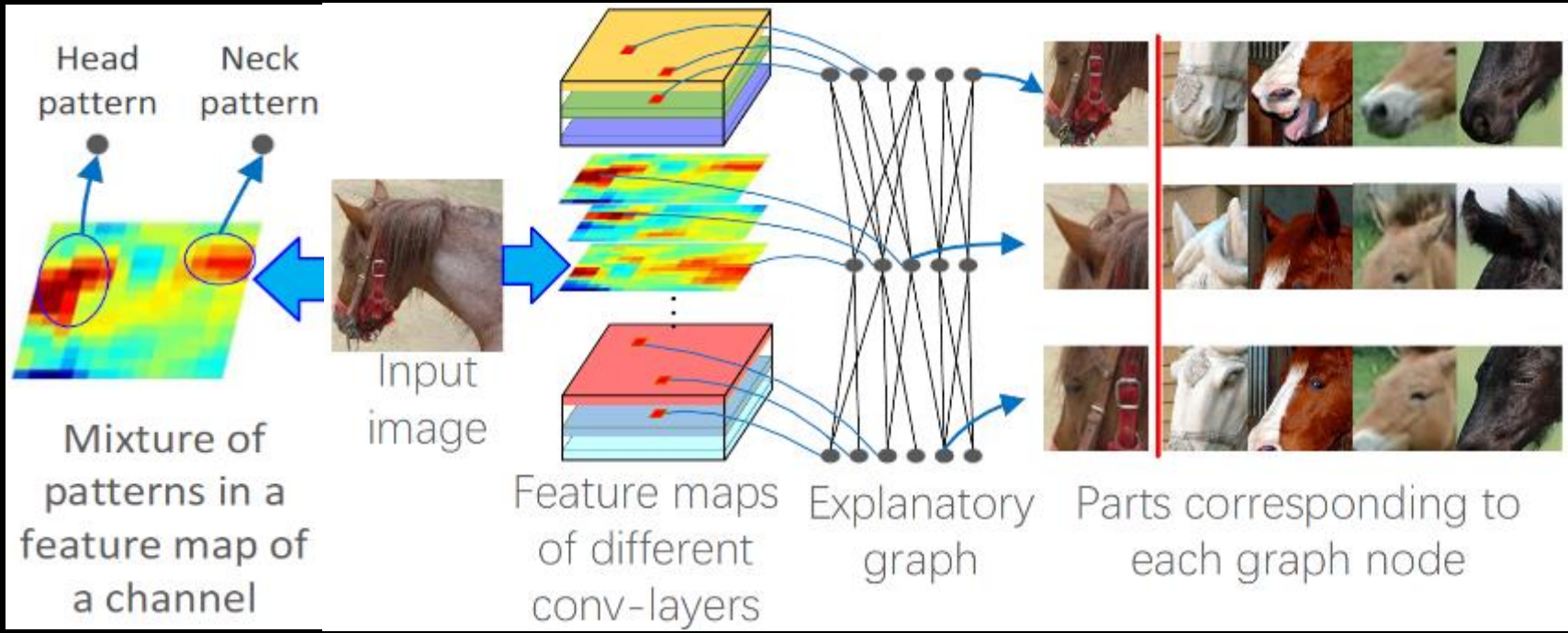
Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Background: Learning explanatory graphs for CNNs

- Given a CNN that is pre-trained for object classification
  - How many types of visual patterns are memorized by a convolutional filter of the CNN?
  - Which patterns are co-activated to describe a part?
  - **What is the spatial relationship between two patterns?**

Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Objective: Summarize knowledge in a CNN into a semantic graph



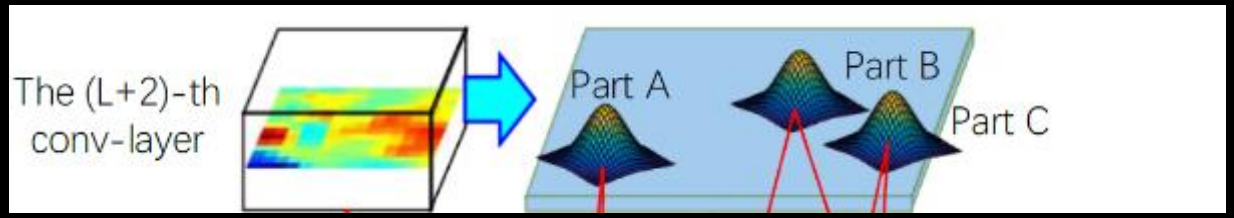

- The graph has multiple layers → multiple conv-layers of the CNN
- Each node → a pattern of an object part
- A filter may encode multiple patterns (nodes)→ disentangle a mixture of patterns from the feature map of a filter
- Each edge → co-activation relationships and spatial relationships between two patterns

# Input & Output

- Input:
  - A pre-trained CNN
    - trained for classification, segmentation, or ...
    - AlexNet, VGG-16, ResNet-50, ResNet-152, and etc.
  - **Without any annotations of parts or textures**
- Output: an explanatory graph

Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Mining an explanatory graph



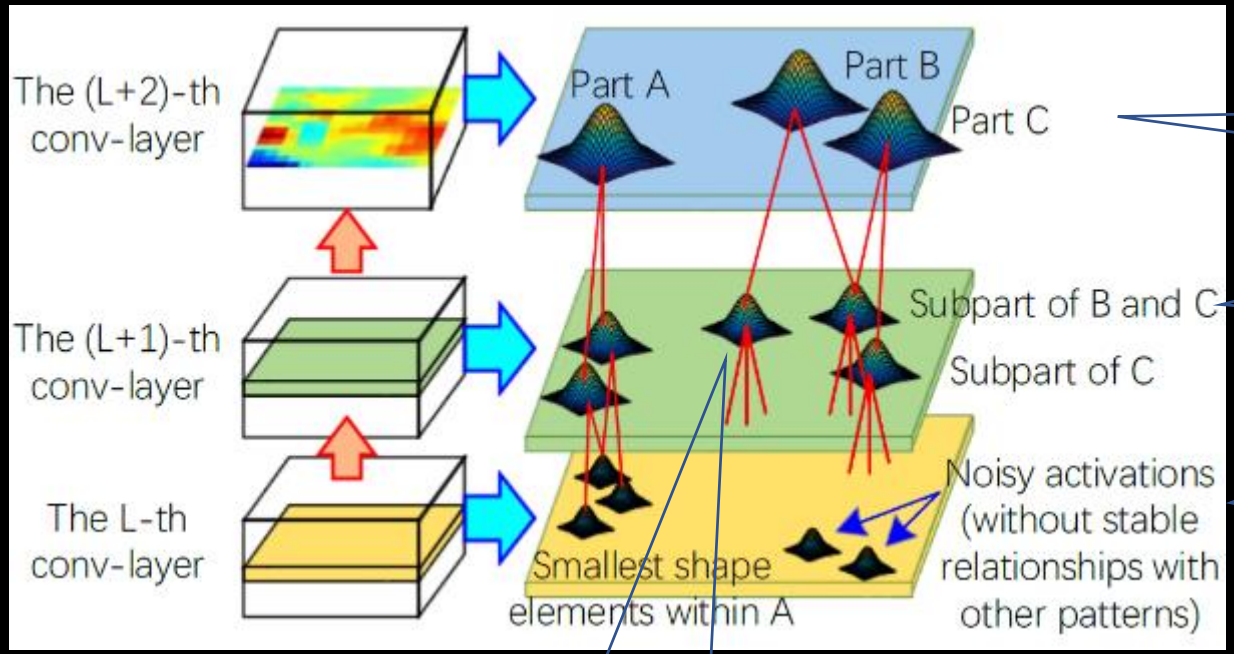

Just like GMM, we use a mixture of patterns to fit activation distributions of a feature map.

a feature map of a filter

→ a distribution of “activation entities”

Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Mining an explanatory graph




Patterns for large parts

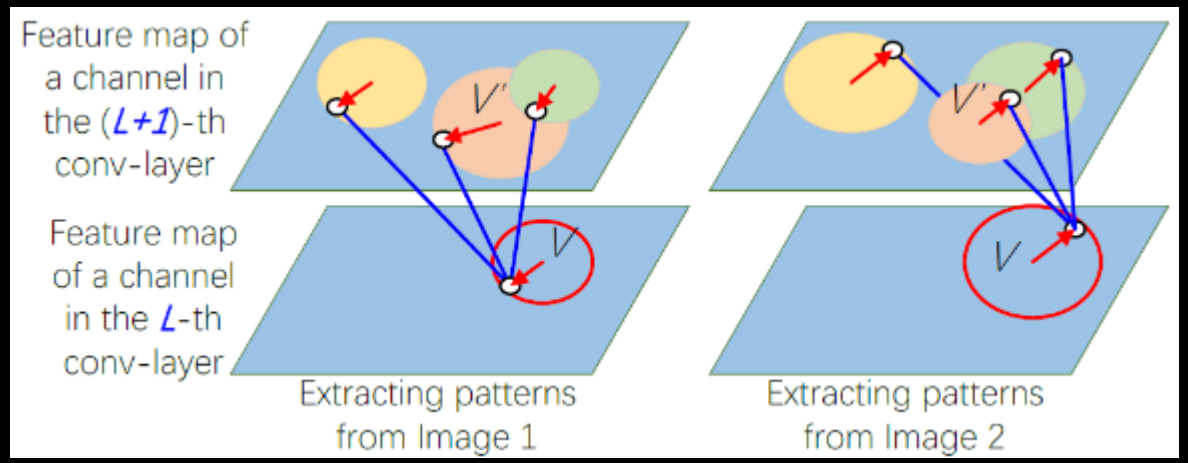
Patterns for subparts

Patterns for even smaller parts

Edges: spatial relationships between co-activated patterns

Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Mining an explanatory graph




Learning node connections
Learning spatial relationship between nodes

Mining a number of cliques: a node V with multiple parents, which keep certain spatial relationships among different images.

Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Using each node in the explanatory graph for part localization

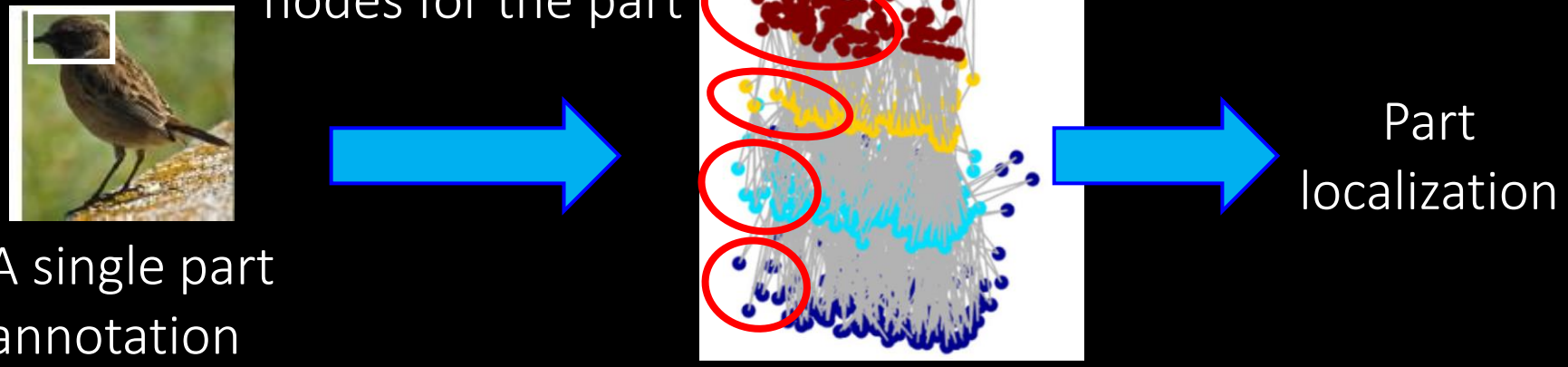

Nodes in the explanatory graph


Raw filters in the CNN

We disentangle each pattern component
from each filter's feature map.

# Knowledge transferring → One/multi-shot part localization

- The part pattern in each node is sophisticatedly learned using numerous images.
  - The retrieved nodes are not overfitted to the labeled part, but represent the common shape among all images




Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018

# Building And-Or graph for semantic hierarchy

Input:

1) An explanatory graph
2) Very few (1—3) annotations for each semantic part

Output:

An AOG as an interpretable model for semantic part localization

Associating the mined patterns with semantic parts of objects

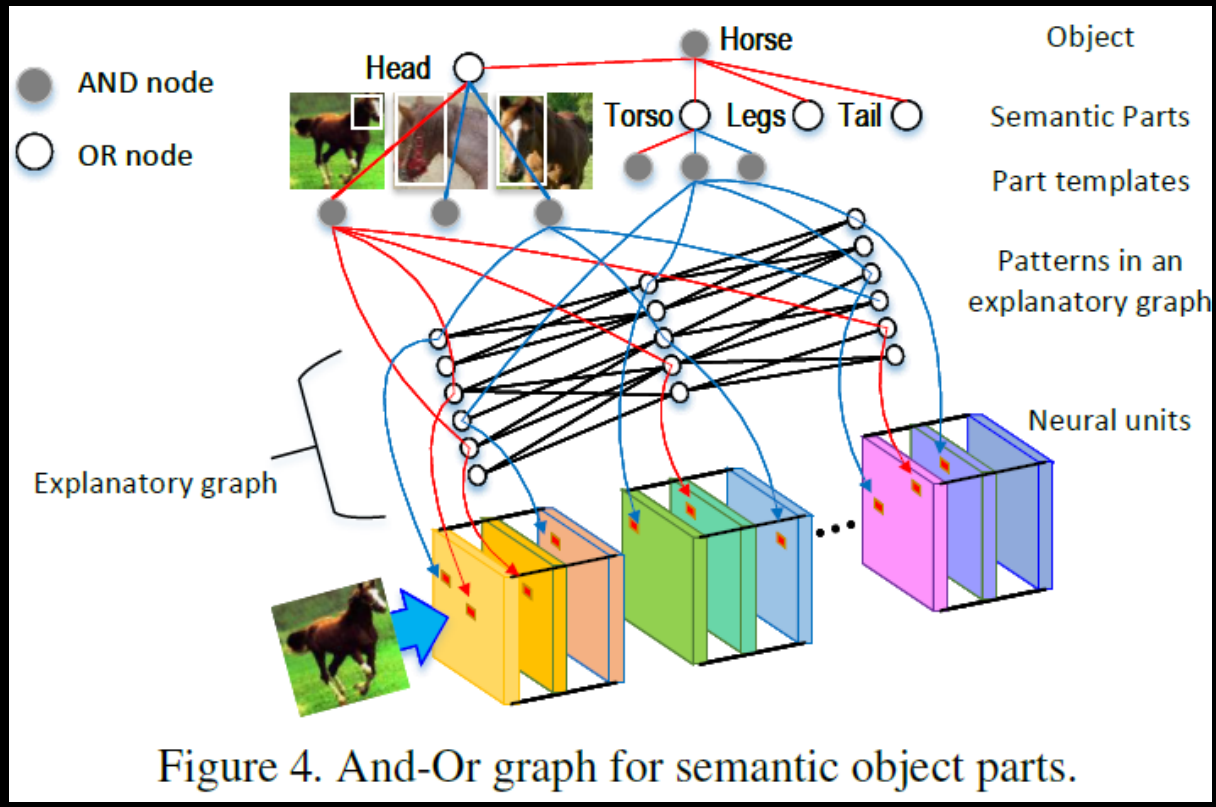
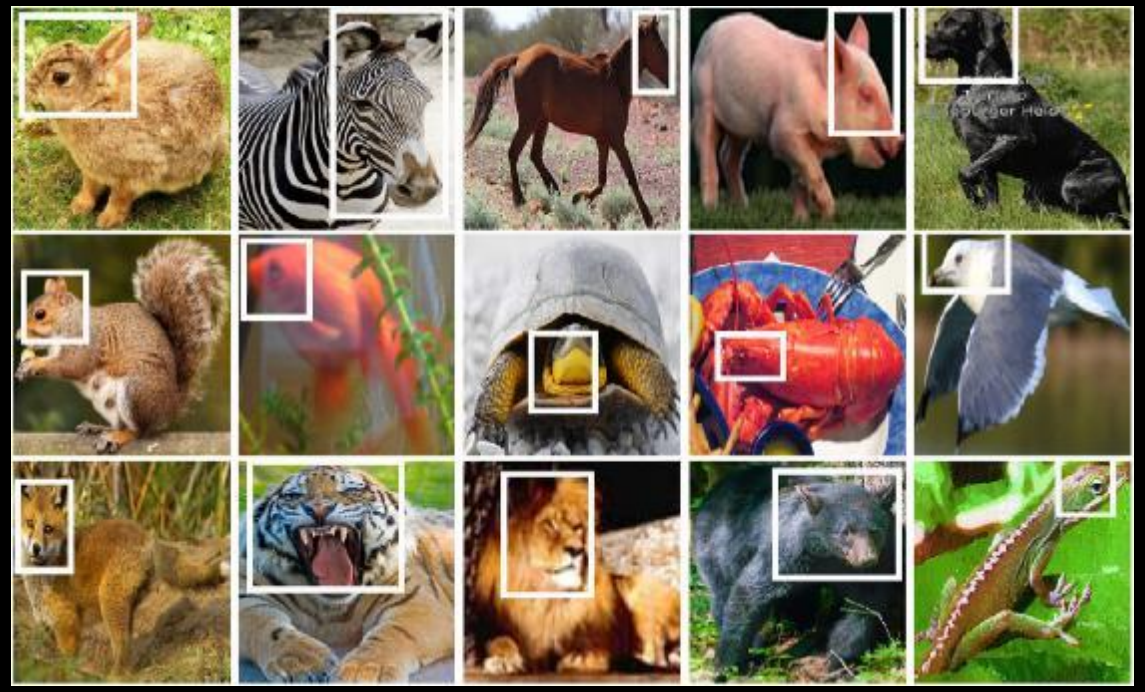



Figure 4. And-Or graph for semantic object parts.

# Performance of few (3)-shot semantic part localization



Decrease 1/3—2/3 localization errors

| | Method | Finetune | normalized distance |
|---|---|---|---|
| no-RL | SS-DPM-Part [2] | N | 0.3469 |
| | PL-DPM-Part [18] | N | 0.3412 |
| | Part-Graph [3] | N | 0.4889 |
| unsup-RL | fc7+linearSVM | Y | 0.3120 |
| | fc7+sp+linearSVM | Y | 0.3120 |
| | Ours | Y | **0.0862** |
| sup-RL | CNN-PDD [26] | N | 0.2333 |
| | CNN-PDD-ft [26] | Y | 0.3269 |
| | Fast-RCNN (1 ft) [9] | N | 0.4517 |
| | Fast-RCNN (2 fts) [9] | Y | 0.4131 |

Table 2. Normalized distance of part localization on the CUB200-

| | | bird | cat | cow | dog | horse | sheep | **Avg** |
|---|---|---|---|---|---|---|---|---|
| no-RL | SS-DPM-Part [2] | 0.356 | 0.270 | 0.264 | 0.242 | 0.262 | 0.286 | 0.280 |
| | PL-DPM-Part [18] | 0.294 | 0.328 | 0.282 | 0.312 | 0.321 | 0.840 | 0.396 |
| | Part-Graph [3] | 0.360 | 0.208 | 0.263 | 0.205 | 0.386 | 0.500 | 0.320 |
| unsup-RL | fc7+linearSVM | 0.247 | 0.174 | 0.251 | 0.217 | 0.261 | 0.317 | 0.244 |
| | fc7+sp+linearSVM | 0.247 | 0.174 | **0.249** | 0.217 | 0.261 | 0.317 | 0.244 |
| | Ours | **0.162** | **0.130** | 0.258 | **0.137** | **0.181** | **0.192** | **0.177** |
| sup-RL | CNN-PDD [26] | 0.301 | 0.246 | 0.220 | 0.248 | 0.292 | 0.254 | 0.260 |
| | CNN-PDD-ft [26] | 0.358 | 0.268 | 0.220 | 0.200 | 0.302 | 0.269 | 0.269 |
| | Fast-RCNN (1 ft) [9] | 0.324 | 0.324 | 0.325 | 0.272 | 0.347 | 0.314 | 0.318 |
| | Fast-RCNN (2 fts) [9] | 0.350 | 0.295 | 0.255 | 0.293 | 0.367 | 0.260 | 0.303 |

Table 3. Normalized distance of part localization on the Pascal VOC Part dataset.

# Outline

- How to represent CNNs using semantic graphical models
- **How to learn disentangled, interpretable features in middle layers**
- How to boost interpretability without hurting the discrimination power
- How to learn networks with functionally interpretable structures

# Background

In traditional CNNs, feature maps of a filter are usually chaotic.

Feature maps of Filter 1

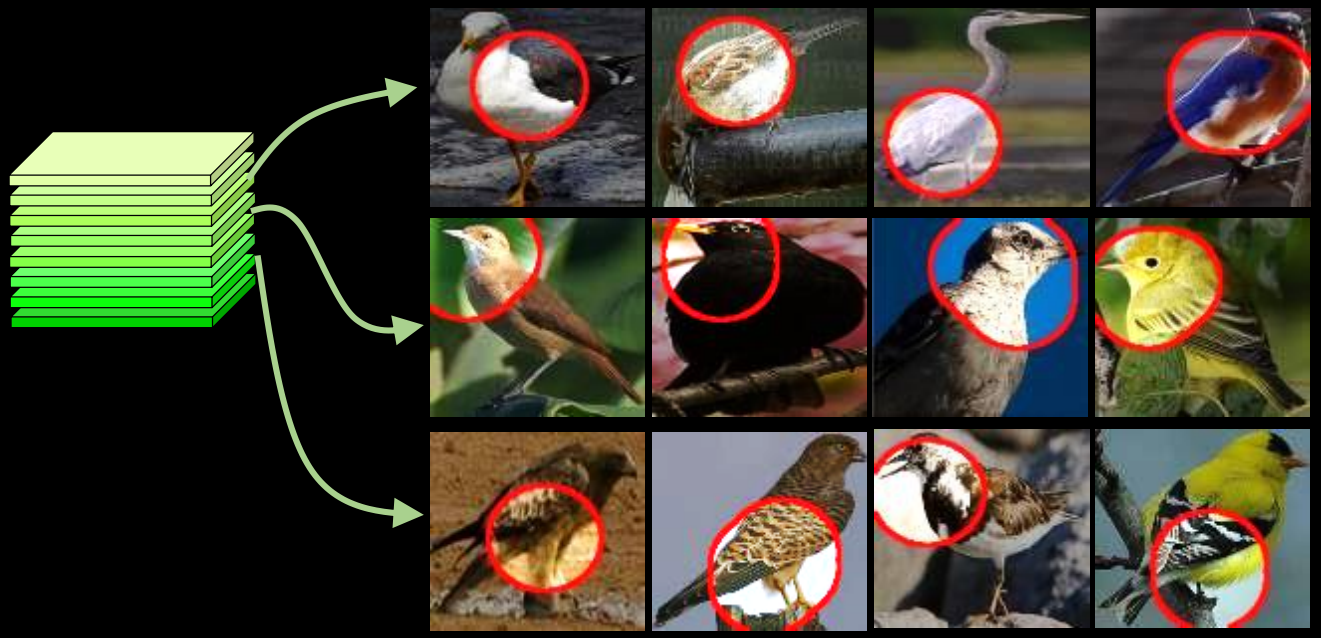
Feature maps of Filter 2

Feature maps of Filter 3

Quanshi Zhang et al. “Interpretable Convolutional Neural Networks” in CVPR 2018

# Objective

**Without** additional part annotations, learn a CNN, where each filter represents a specific part through different objects.



Neural activations of 3 interpretable filters

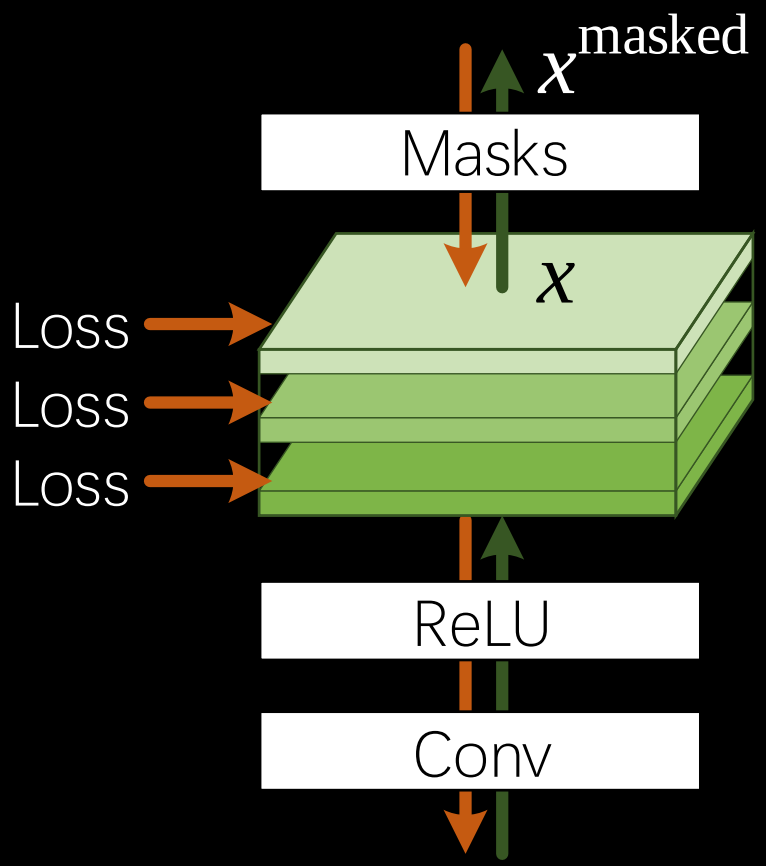
Quanshi Zhang et al. “Interpretable Convolutional Neural Networks” in CVPR 2018

# Input & Output: Interpretable CNNs

- Input
  - Training samples $(X_i, Y_i)$ for a certain task
    - Applicable to different tasks, e.g., classification & segmentations
    - Applicable to different CNNs, e.g., AlexNet, VGG-16, VGG-M, VGG-S
  - **No** annotations of parts or textures are used.
- Output
  - An interpretable CNN with disentangled filters

# Network structure

We add a loss to each channel to construct an interpretable layer

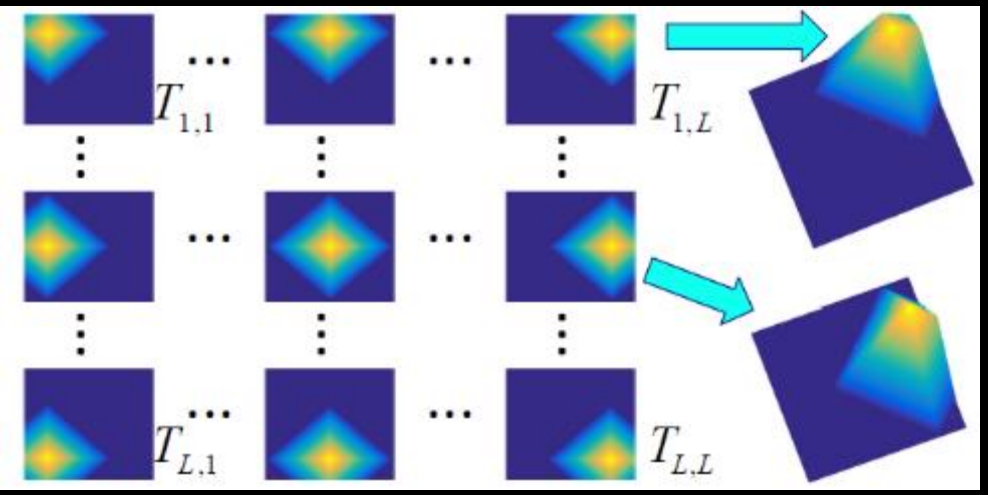



$$Loss = \underbrace{Loss(\hat{y}, y^*)}_{\text{task loss}} + \sum_f \underbrace{Loss_f(x)}_{\text{filter loss}}$$

The filter loss boosts the mutual information between feature maps $\boldsymbol{X}$ and a set of pre-defined part locations $\boldsymbol{T}$.

$$\mathbf{Loss}_f = -\,MI(\mathbf{X};\mathbf{T}) \quad \text{for filter } f$$

# Network structure




$$Loss = \underbrace{Loss(\hat{y}, y^*)}_{\text{task loss}} + \sum_f \underbrace{Loss_f(x)}_{\text{filter loss}}$$

$$\mathbf{Loss}_f = -\,MI(\mathbf{X};\mathbf{T}) \quad \text{for filter } f$$

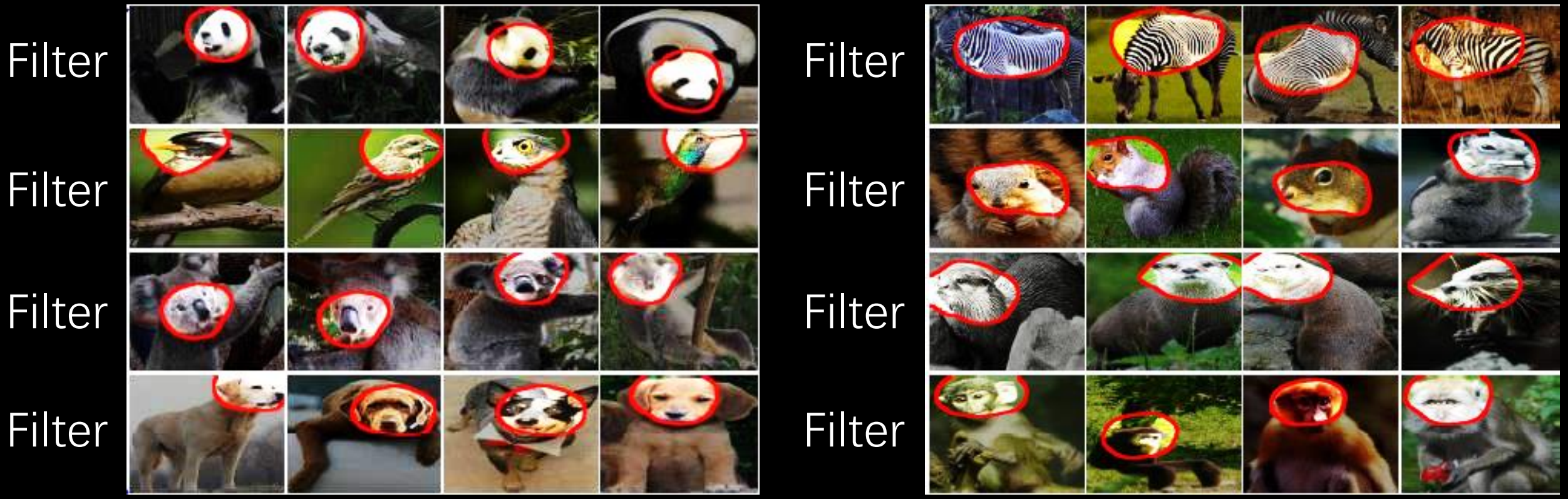
$$Loss = -\,H(\mathbf{T}) + H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X}) + \sum_x p(\mathbf{T}^+, x) H(\mathbf{T}^+ = \{T_\mu\}|X = x)$$

A constant

Entropy of Inter-category activations

Entropy of the spatial distribution of activations

# Activation regions of interpretable filters



# Our method learns filters with much higher interpretability

| | gold. | bird | frog | turt. | liza. | koala | lobs. | dog | fox | cat | lion | tiger | bear | rabb. | hams. | squi. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AlexNet | 0.161 | 0.167 | 0.152 | 0.153 | 0.175 | 0.128 | 0.123 | 0.144 | 0.143 | 0.148 | 0.137 | 0.142 | 0.144 | 0.148 | 0.128 | 0.149 |
| AlexNet, interpretable | **0.084** | **0.095** | **0.090** | **0.107** | **0.097** | **0.079** | **0.077** | **0.093** | **0.087** | **0.095** | **0.084** | **0.090** | **0.095** | **0.095** | **0.077** | **0.095** |
| VGG-16 | 0.153 | 0.156 | 0.144 | 0.150 | 0.170 | 0.127 | 0.126 | 0.143 | 0.137 | 0.148 | 0.139 | 0.144 | 0.143 | 0.146 | 0.125 | 0.150 |
| VGG-16, interpretable | **0.076** | **0.099** | **0.086** | **0.115** | **0.113** | **0.070** | **0.084** | **0.077** | **0.069** | **0.086** | **0.067** | **0.097** | **0.081** | **0.079** | **0.066** | **0.065** |
| VGG-M | 0.161 | 0.166 | 0.151 | 0.153 | 0.176 | 0.128 | 0.125 | 0.145 | 0.145 | 0.150 | 0.140 | 0.145 | 0.144 | 0.150 | 0.128 | 0.150 |
| VGG-M, interpretable | **0.088** | **0.088** | **0.089** | **0.108** | **0.099** | **0.080** | **0.074** | **0.090** | **0.082** | **0.103** | **0.079** | **0.089** | **0.101** | **0.097** | **0.082** | **0.095** |
| VGG-S | 0.158 | 0.166 | 0.149 | 0.151 | 0.173 | 0.127 | 0.124 | 0.143 | 0.142 | 0.148 | 0.138 | 0.142 | 0.143 | 0.148 | 0.128 | 0.146 |
| VGG-S, interpretable | **0.087** | **0.101** | **0.093** | **0.107** | **0.096** | **0.084** | **0.078** | **0.091** | **0.082** | **0.101** | **0.082** | **0.089** | **0.097** | **0.091** | **0.076** | **0.098** |

| | horse | zebra | swine | hippo | catt. | sheep | ante. | camel | otter | arma. | monk. | elep. | red pa. | gia.pa. | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AlexNet | 0.152 | 0.154 | 0.141 | 0.141 | 0.144 | 0.155 | 0.147 | 0.153 | 0.159 | 0.160 | 0.139 | 0.125 | 0.140 | 0.125 | 0.146 |
| AlexNet, interpretable | **0.098** | **0.084** | **0.091** | **0.089** | **0.097** | **0.101** | **0.085** | **0.102** | **0.104** | **0.095** | **0.090** | **0.085** | **0.084** | **0.073** | **0.091** |
| VGG-16 | 0.150 | 0.153 | 0.141 | 0.140 | 0.140 | 0.150 | 0.144 | 0.149 | 0.154 | 0.163 | 0.136 | 0.129 | 0.143 | 0.125 | 0.144 |
| VGG-16, interpretable | **0.106** | **0.077** | **0.094** | **0.083** | **0.102** | **0.097** | **0.091** | **0.105** | **0.093** | **0.100** | **0.074** | **0.084** | **0.067** | **0.063** | **0.085** |
| VGG-M | 0.151 | 0.158 | 0.140 | 0.140 | 0.143 | 0.155 | 0.146 | 0.154 | 0.160 | 0.161 | 0.140 | 0.126 | 0.142 | 0.127 | 0.147 |
| VGG-M, interpretable | **0.095** | **0.080** | **0.095** | **0.084** | **0.092** | **0.094** | **0.077** | **0.104** | **0.102** | **0.093** | **0.086** | **0.087** | **0.089** | **0.068** | **0.090** |
| VGG-S | 0.149 | 0.155 | 0.139 | 0.140 | 0.141 | 0.155 | 0.143 | 0.154 | 0.158 | 0.157 | 0.140 | 0.125 | 0.139 | 0.125 | 0.145 |
| VGG-S, interpretable | **0.096** | **0.080** | **0.092** | **0.088** | **0.094** | **0.101** | **0.077** | **0.102** | **0.105** | **0.094** | **0.090** | **0.086** | **0.078** | **0.072** | **0.090** |

Table 3. Location instability of filters ($\mathbf{E}_{f,k}[D_{f,k}]$) in CNNs that are trained for single-category classification using the ILSVRC 2013 DET Animal-Part dataset [36]. Filters in our interpretable CNNs exhibited significantly lower localization instability than ordinary CNNs.

# Classification performance

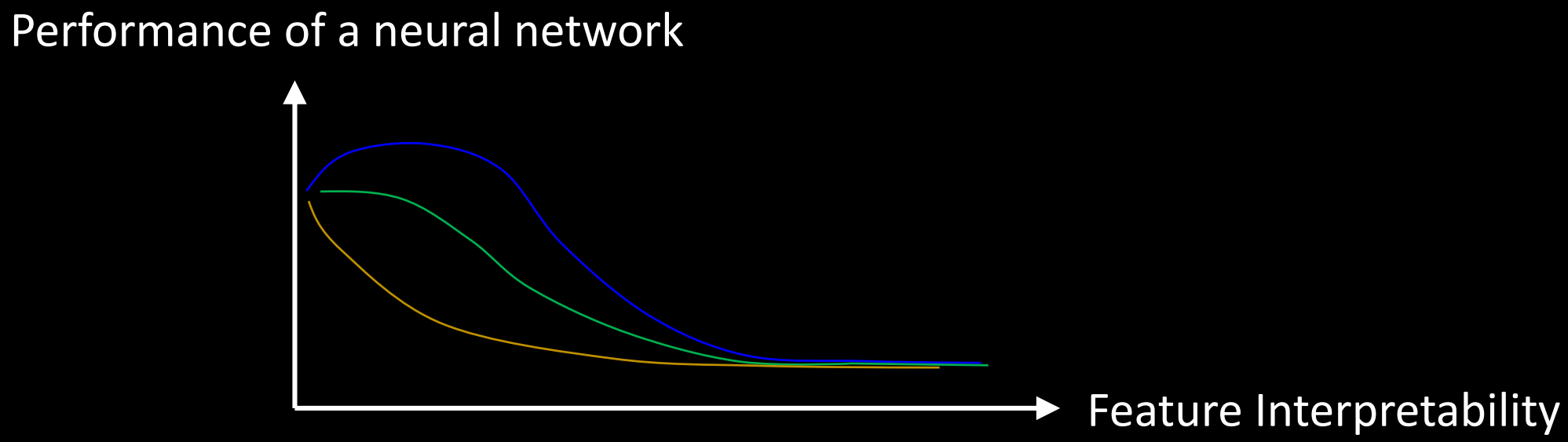
| | multi-category | | | single-category | | |
|---|---|---|---|---|---|---|
| | ILSVRC Part | VOC Part | | ILSVRC Part | VOC Part | CUB200 |
| | logistic[4] | logistic[4] | softmax | | | |
| AlexNet | – | – | – | **96.28** | **95.40** | **95.59** |
| interpretable | – | – | – | 95.38 | 93.93 | 95.35 |
| VGG-M | 96.73 | 93.88 | 81.93 | **97.34** | **96.82** | **97.34** |
| interpretable | **97.99** | **96.19** | **88.03** | 95.77 | 94.17 | 96.03 |
| VGG-S | 96.98 | 94.05 | 78.15 | **97.62** | **97.74** | **97.24** |
| interpretable | **98.72** | **96.78** | **86.13** | 95.64 | 95.47 | 95.82 |
| VGG-16 | – | 97.97 | 89.71 | **98.58** | **98.66** | **98.91** |
| interpretable | – | **98.50** | **91.60** | 96.67 | 95.39 | 96.51 |

Our interpretable CNNs outperformed traditional CNNs in multi-category classification.

# Outline

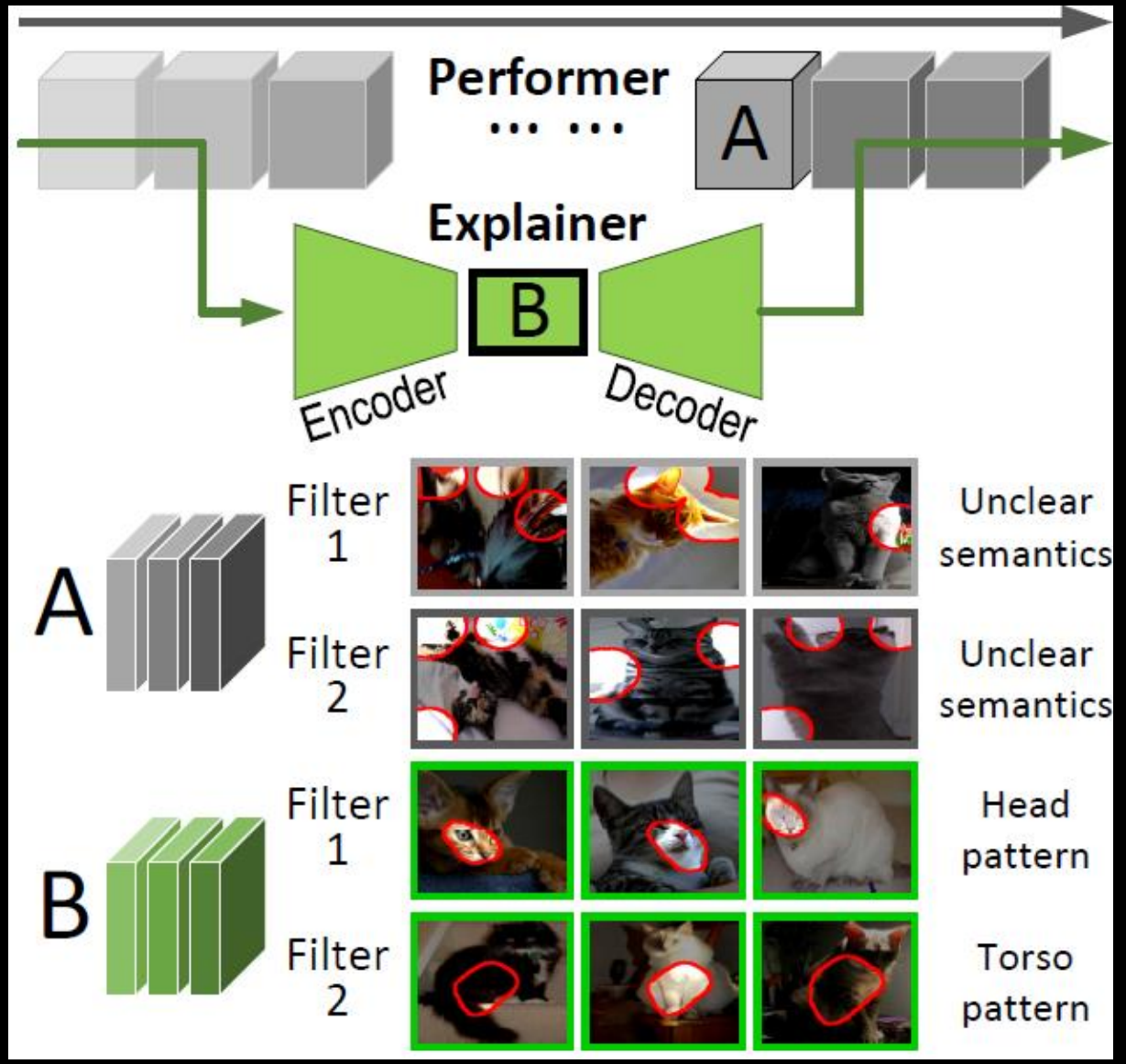
- How to represent CNNs using semantic graphical models
- How to learn disentangled, interpretable features in middle layers
- **How to boost interpretability without hurting the discrimination power**
- How to learn networks with functionally interpretable structures

# Motivation: Unsupervised Learning of Neural Networks to Explain Neural Networks



# Performer & Explainer

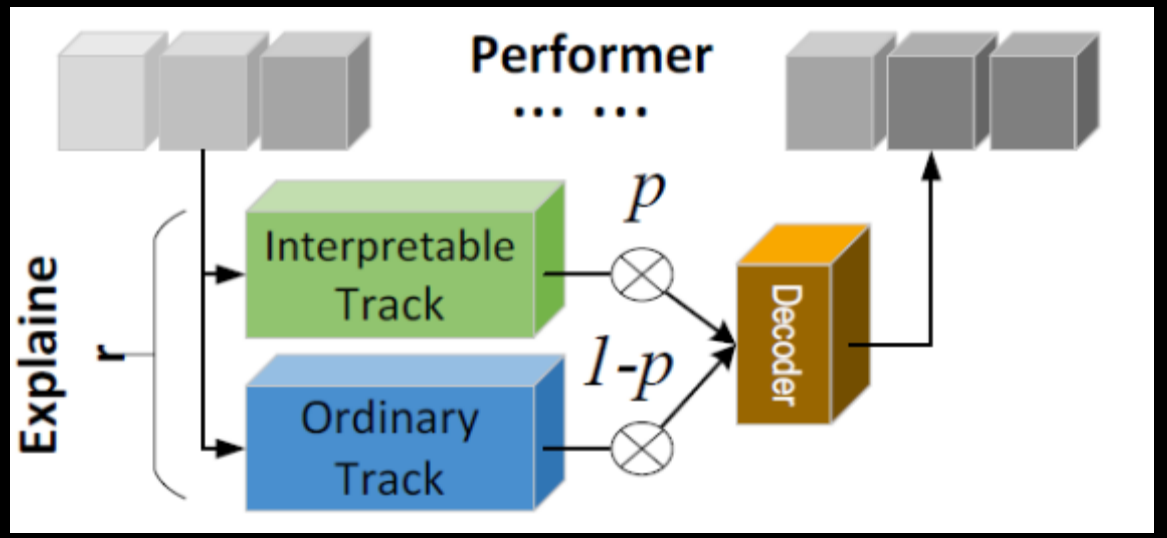
- Black-box performer
- Interpretable explainer
- Using interpretable feature maps in the explainer network to reconstruct feature maps in the performer network



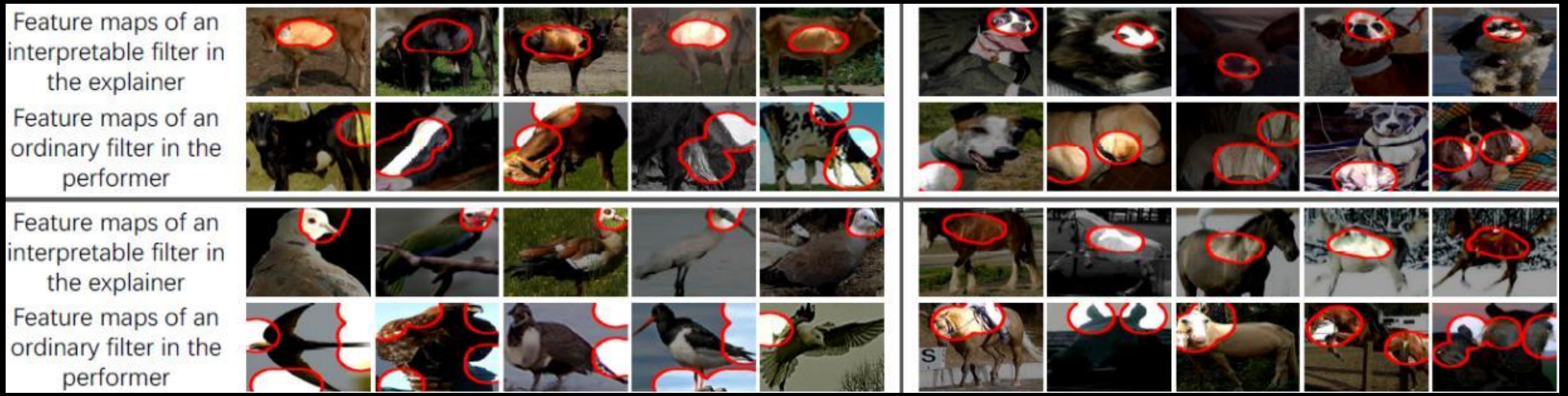

Quanshi Zhang et al. "Unsupervised Learning of Neural Networks to Explain Neural Networks" extended abstract in AAAI Workshop on Network Interpretability for Deep Learning, 2019

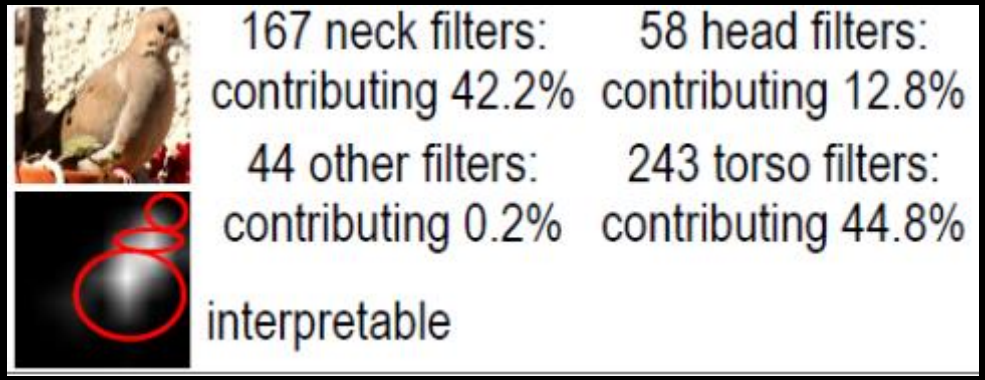
# Explainer network

- Interpretable track
  - Disentangled feature maps → object parts
- Ordinary track
  - Ordinary feature maps
- $\mathrm{x} = px_{interpretable} + (1-p)x_{ordinary}$
  - P: the ratio of interpretable information
  - Increase P




Quanshi Zhang et al. "Unsupervised Learning of Neural Networks to Explain Neural Networks" extended abstract in AAAI Workshop on Network Interpretability for Deep Learning, 2019


Feature maps of an interpretable filter in the explainer
Feature maps of an ordinary filter in the performer
Feature maps of an interpretable filter in the explainer
Feature maps of an ordinary filter in the performer


167 neck filters:
contributing 42.2%
58 head filters:
contributing 12.8%
44 other filters:
contributing 0.2%
243 torso filters:
contributing 44.8%
interpretable

# Disentangle part information from features

Table 2: Average $p$ values of explainers in different experiments.

| | Pascal-Part [4] | | CUB200 |
|---|---|---|---|
| | Single | Multi | -2011 [33] |
| AlexNet | – | 0.7137 | 0.5810 |
| VGG-M | 0.9012 | 0.8066 | 0.8611 |
| VGG-S | 0.9270 | 0.8996 | 0.9533 |
| VGG-16 | 0.8593 | 0.8718 | 0.9579 |

# Outline

- How to represent CNNs using semantic graphical models
- How to learn disentangled, interpretable features in middle layers
- How to boost interpretability without hurting the discrimination power
- **How to learn networks with functionally interpretable structures**

# Outline

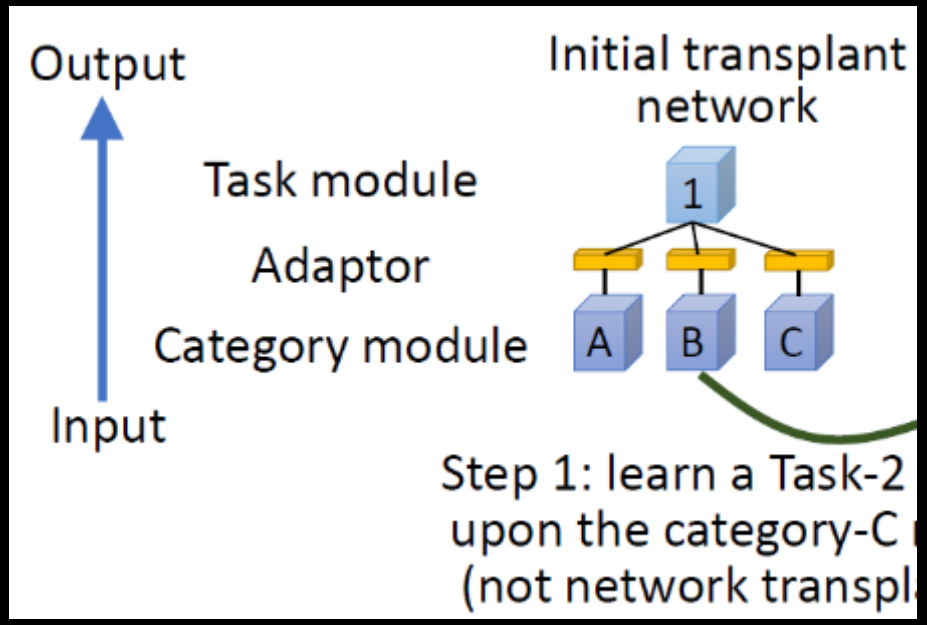
- How to represent CNNs using semantic graphical models
- How to learn disentangled, interpretable features in middle layers
- How to boost interpretability without hurting the discrimination power
- **How to learn networks with functionally interpretable structures**
  - **Learning a universal neural network for massive tasks and massive categories**

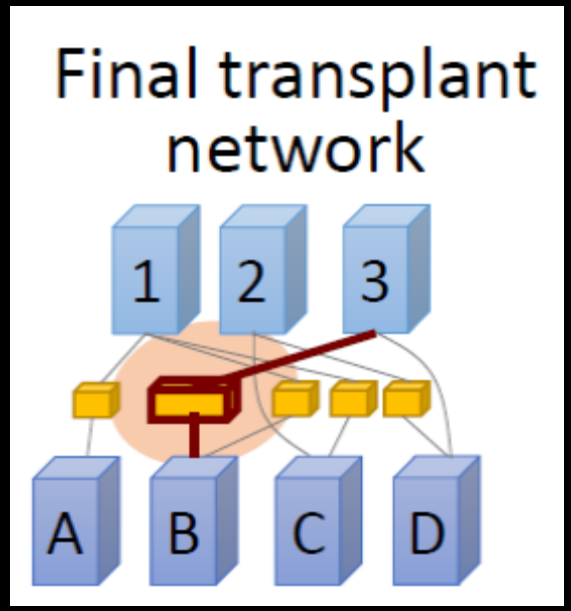
从学习“通用网络”的角度来看，deep learning需要具备哪些特质呢？

# Network Transplanting

神经网络表达与图模型表达的殊途同归

- Interpretable network structures
  - Category modules
  - Task modules
  - Adapter modules
- Gradually merge specific nets → a large generic, distributed, net

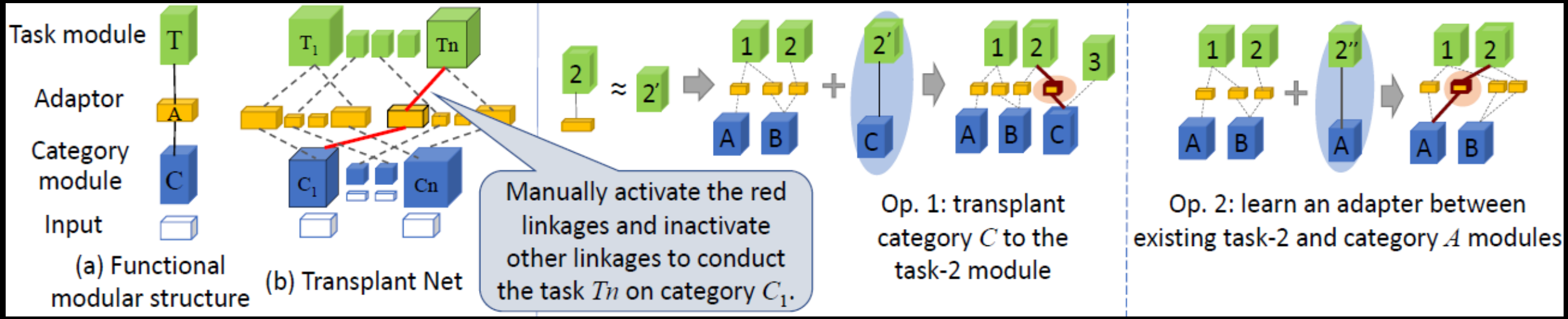



Quanshi Zhang et al. “Network Transplanting” extended abstract in AAAI Workshop on Network Interpretability for Deep Learning, 2019

# Core task of network transplanting

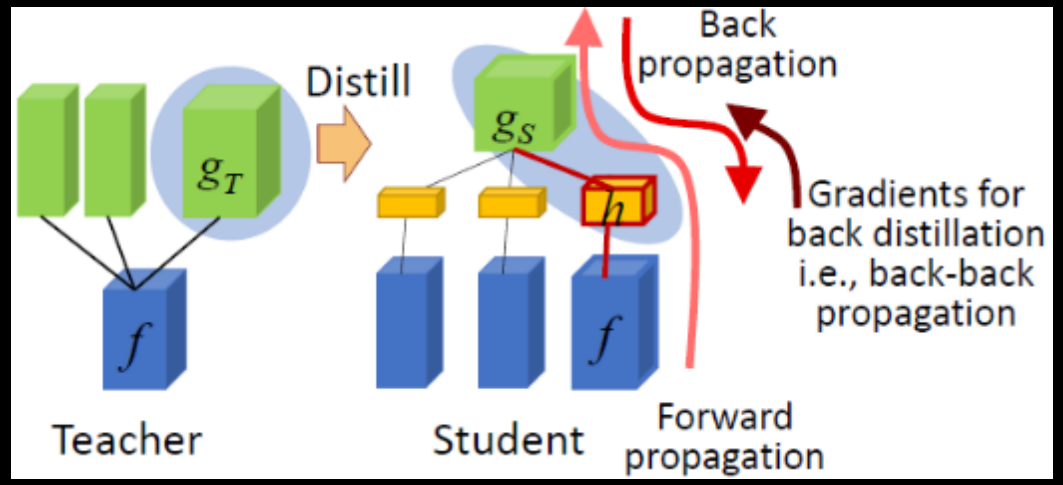
- Given Net A : Classification of the bird
- Given Net B : Segmentation of the cat
- Transplant the classification module from A to B
  - Enable the classification of the cat
- Transplant the segmentation module from B to A
  - Enable the segmentation of the bird

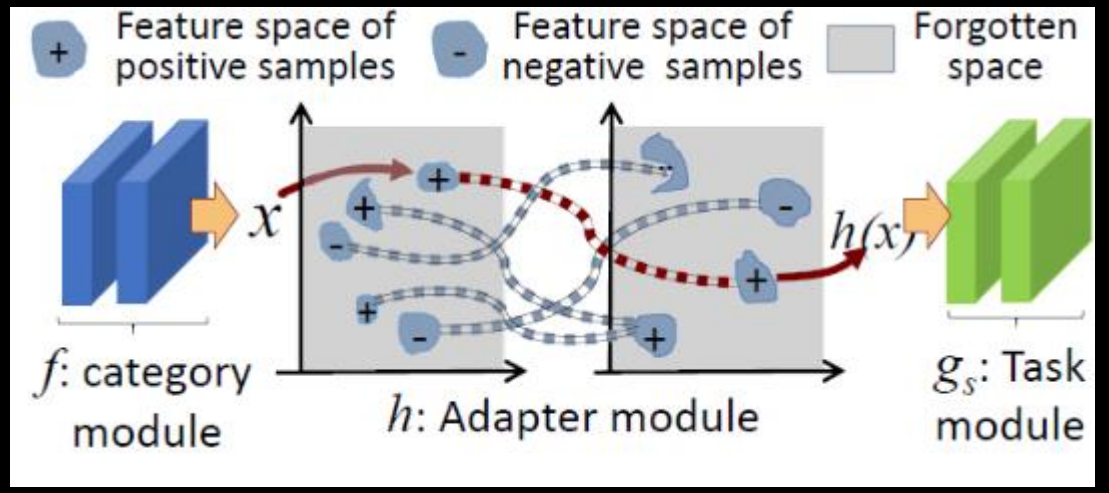



Quanshi Zhang et al. “Network Transplanting” extended abstract in AAAI Workshop on Network Interpretability for Deep Learning, 2019

# Challenge: how to project features between two spaces

- Only learning the adapter module, without destroying the generality of the category module and the task module
  - BP does not work
  - Propose a new optimization method: back distillation




Quanshi Zhang et al. “Network Transplanting” extended abstract in AAAI Workshop on Network Interpretability for Deep Learning, 2019

Table 1: Error rate of classification when we insert one conv-layer and one ReLU layer to a pre-trained CNN as the adapter.

| # of samples | | cat | cow | dog | horse | sheep | **Avg.** |
|---|---|---|---|---|---|---|---|
| 100 | OL | 12.89 | 3.09 | 12.89 | 10.82 | 9.28 | 9.79 |
| | BD | **1.55** | **0.52** | **3.61** | **1.55** | **1.03** | **1.65** |
| 50 | OL | 13.92 | 15.98 | 12.37 | 16.49 | 15.46 | 14.84 |
| | BD | **1.55** | **0.52** | **3.61** | **1.55** | **1.03** | **1.65** |
| 20 | OL | 16.49 | 26.80 | 28.35 | 32.47 | 25.77 | 25.98 |
| | BD | **1.55** | **0.52** | **3.09** | **1.55** | **1.03** | **1.55** |
| 10 | OL | 39.18 | 39.18 | 35.05 | 41.75 | 38.66 | 38.76 |
| | BD | **1.55** | **0.52** | **3.61** | **1.55** | **1.03** | **1.65** |
| **0** | OL | – | – | – | – | – | – |
| | BD | **1.55** | **0.52** | **4.12** | **1.55** | **1.03** | **1.75** |

Table 2: Error rate of classification when we insert three conv-layers and three ReLU layers to a pre-trained CNN as the adapter.

| # of samples | | cat | cow | dog | horse | sheep | **Avg.** |
|---|---|---|---|---|---|---|---|
| 100 | OL | 9.28 | 6.70 | 12.37 | 11.34 | 3.61 | 8.66 |
| | BD | **1.03** | **2.58** | **4.12** | **1.55** | **2.58** | **2.37** |
| 50 | OL | 14.43 | 13.92 | 15.46 | 8.76 | 7.22 | 11.96 |
| | BD | **3.09** | **3.09** | **4.12** | **2.06** | **4.64** | **3.40** |
| 20 | OL | 22.16 | 25.77 | 32.99 | 22.68 | 22.16 | 25.15 |
| | BD | **7.22** | **6.70** | **7.22** | **2.58** | **5.15** | **5.77** |
| 10 | OL | 36.08 | 32.99 | 31.96 | 34.54 | 34.02 | 33.92 |
| | BD | **8.25** | **15.46** | **10.31** | **13.92** | **10.31** | **11.65** |
| **0** | OL | – | – | – | – | – | – |
| | BD | **50.00** | **50.00** | **50.00** | **49.48** | **50.00** | **49.90** |

Quanshi Zhang et al. “Network Transplanting” extended abstract in AAAI Workshop on Network Interpretability for Deep Learning, 2019

# Network transplanting vs. traditional learning

- Modular interpretability → more controllability
- Much weaker supervision than learning from data
- Incremental adding new modules (new tasks, new categories) to the net
  - Generic net for multiple categories and multiple tasks
  - No need to simultaneously prepare training samples for all tasks and all categories
    - Valuable when there are lots of categories and tasks
- Catastrophic forgetting

Quanshi Zhang et al. “Network Transplanting” extended abstract in AAAI Workshop on Network Interpretability for Deep Learning, 2019

- Summarization
  - Explaining pre-trained deep model: transforming CNN representations into semantic graphs
  - Learning interpretable features for DNNs
  - Conflicts between the interpretability and the discrimination power
  - Learning functionally interpretable structures

# Panel discussion

- 可解释性学习未来的发展方向是什么？
- 可解释对于深度学习的意义和研究目标是什么？
- 现在部分学者认为可解释性在学习中不是必要条件，大家怎么看？
- 可解释性学习如何来进行科学的度量？解释结果的客观性与可靠性怎么评价？
- 可解释性的应用场景有哪些？
- 可解释性学习的最大挑战是什么？