IJCAI 2021 Tutorial

What is the relationship between interactions and visual concepts? —— Learning compositional and interpretable features.

Wen Shen, Quanshi Zhang

Background

- Neural activations of filters in traditional CNNs

Feature maps of a filter in traditional CNNs are usually chaotic.

Strongly interacted filters → meaningful concepts

Disentangling visual concepts from chaotic feature maps

Wen Shen

Quanshi Zhang

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

Disentangling visual concepts from chaotic feature maps


Wen Shen **Quanshi Zhang**

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]

Each node represents a pattern of an object part. A filter may encode multiple patterns (nodes).

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

Disentangling visual concepts from chaotic feature maps

Wen Shen **Quanshi Zhang**

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]

Each node represents a pattern of an object part. A filter may encode multiple patterns (nodes).

A filter only encode a specific pattern of an object part.

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Disentangling visual concepts from chaotic feature maps


Wen Shen **Quanshi Zhang**

Learning interpretable features

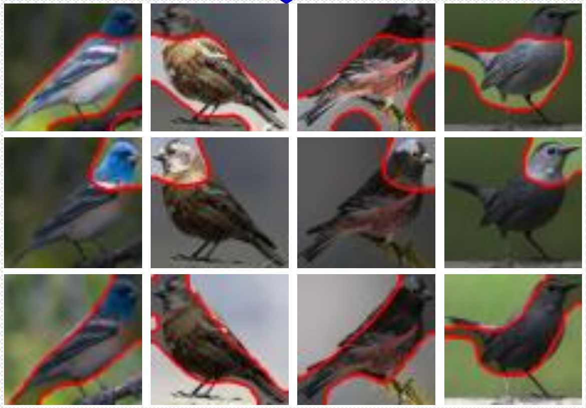
Build an explanatory graph to explain the semantic hierarchy[1]

Each node represents a pattern of an object part. A filter may encode multiple patterns (nodes).

The interpretable CNN, where each filter represents a specific object part[2]

A filter only encode a specific pattern of an object part.

The compositional CNN, where each filter represents a specific object part/image region[3]



A filter can encode a specific pattern of an object part or image region.

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

Disentangling visual concepts from chaotic feature maps

Wen Shen **Quanshi Zhang**

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]

- The interpretability of filters is gradually enhanced.

A filter may encode multiple patterns [1] → A filter only encodes a specific pattern [2][3].

- The representation power of filters is gradually enhanced.

A filter can only encodes an object part in ball-like areas [2] → A filter can encodes an object part with a specific shape or the image region without a specific structure [3].

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

The relationship between interactions and visual concepts

Wen Shen

Quanshi Zhang

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

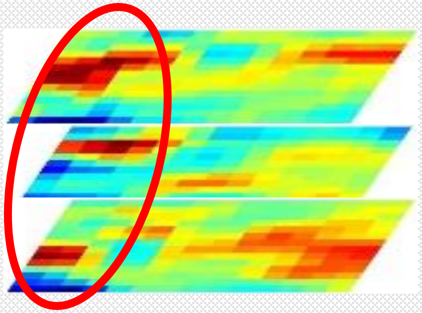
The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]

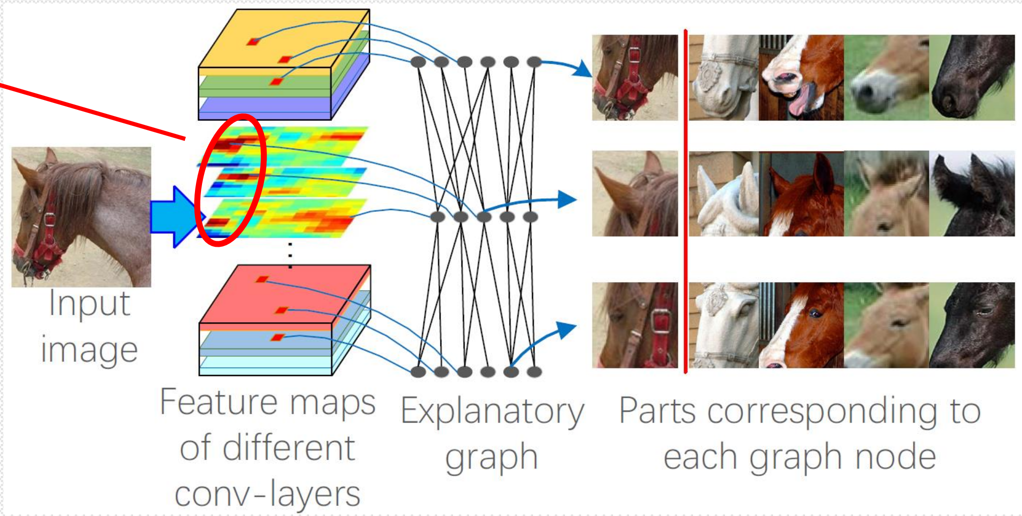
Filter 1

Filter 2

Filter 3



A node of the explanatory graph is encoded as the **highly interacted activations** of a few filters.



[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

The relationship between interactions and visual concepts


Wen Shen **Quanshi Zhang**

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

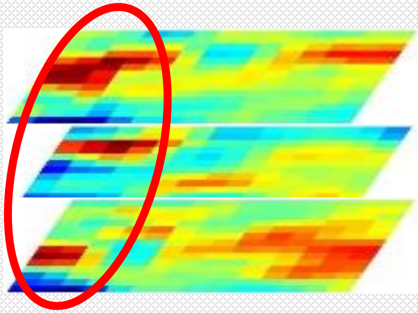
The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]

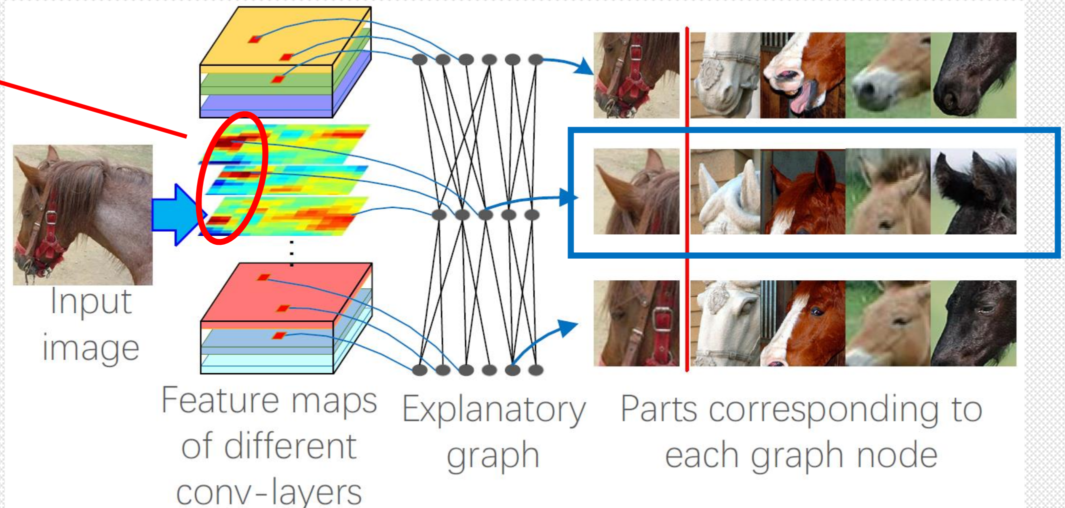
Filter 1

Filter 2

Filter 3



A node of the explanatory graph is encoded as the **highly interacted activations** of a few filters.



E.g., these filters with highly interacted activations in certain area represent the head of a horse.

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

The relationship between interactions and visual concepts

Wen Shen

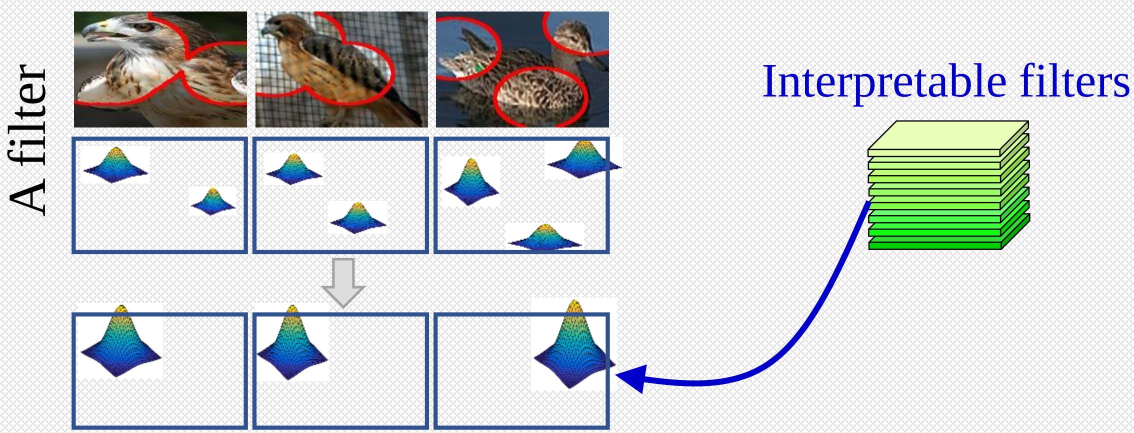
Quanshi Zhang

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The **interpretable CNN**, where each filter represents **a specific object part**[2]

The compositional CNN, where each filter represents a specific object part/image region[3]



Use **regional interaction activations** of a filter to represent object parts.

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

The relationship between interactions and visual concepts

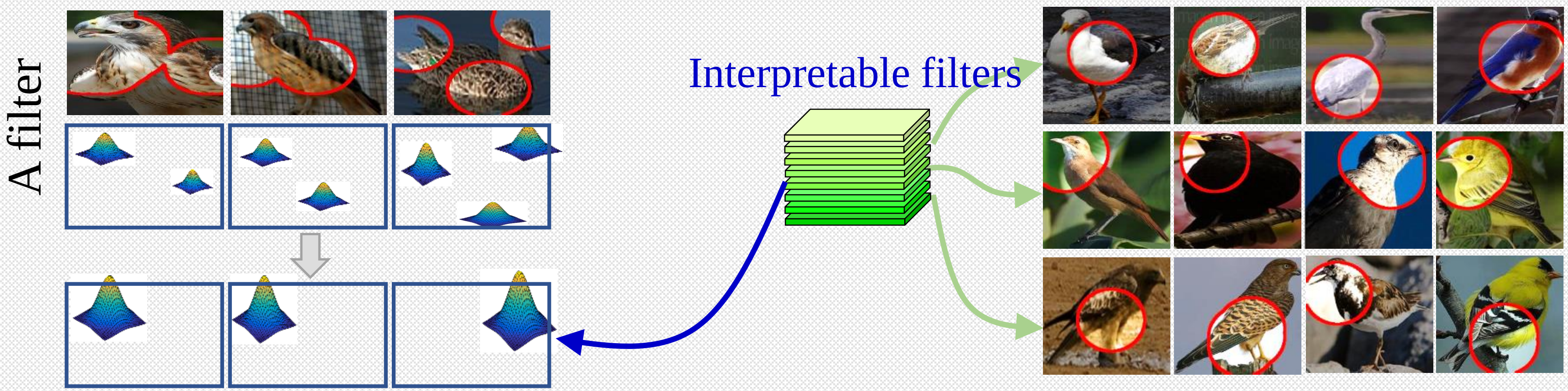
Wen Shen **Quanshi Zhang**

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]



Use **regional interaction activations** of a filter to represent object parts.

Each filter represents a specific part through different objects.

[1] Quanshi Zhang et al. “Interpreting CNN Knowledge via an Explanatory Graph” in AAAI 2018
[2] Quanshi Zhang et al. “Interpretable Convolutional Neural Networks” in CVPR 2018
[3] Wen Shen et al. “Interpretable Compositional Convolutional Neural Networks”in IJCAI 2021

Strongly interacted filters → meaningful concepts

The relationship between interactions and visual concepts

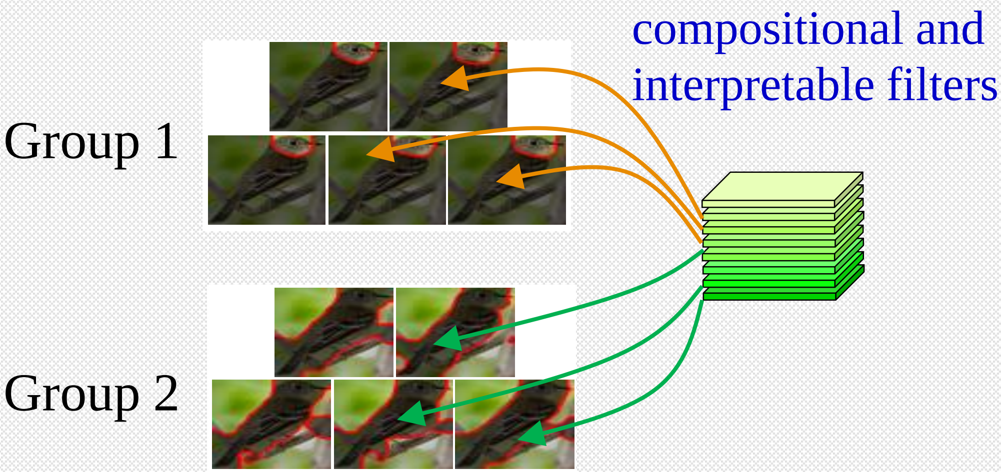

Wen Shen **Quanshi Zhang**

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]



A group of filters **cooperate with each other** to make inferences.

The **cooperative features have strong interactions.**

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

Strongly interacted filters → meaningful concepts

The relationship between interactions and visual concepts

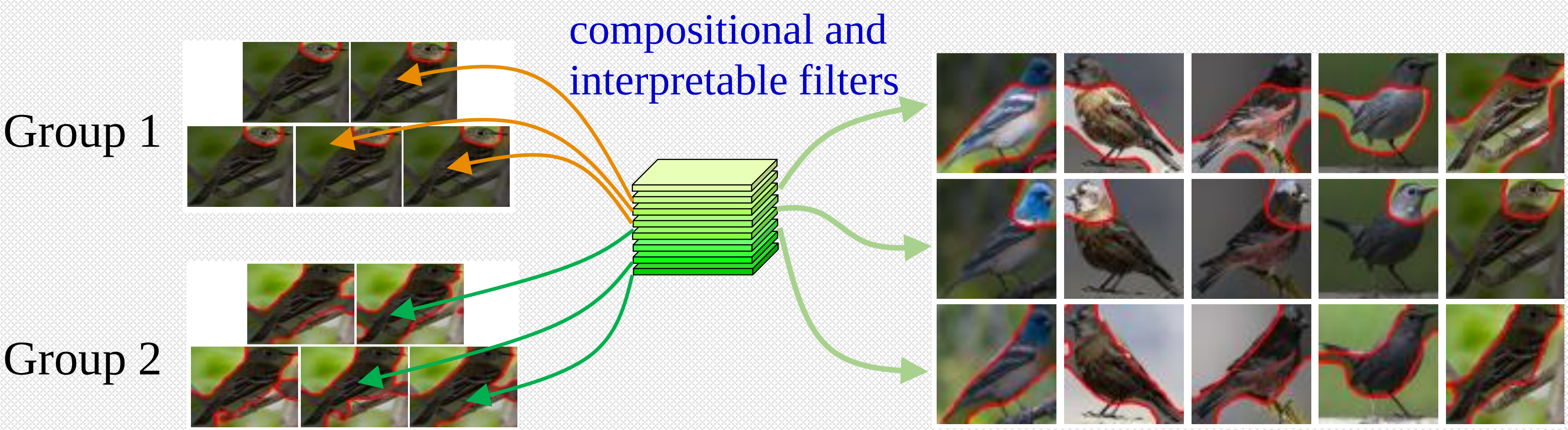
Wen Shen **Quanshi Zhang**

Learning interpretable features

Build an explanatory graph to explain the semantic hierarchy[1]

The interpretable CNN, where each filter represents a specific object part[2]

The compositional CNN, where each filter represents a specific object part/image region[3]



A group of filters **cooperate with each other** to make inferences.

The **cooperative features have strong interactions.**

[1] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018
[2] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018
[3] Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks"in IJCAI 2021

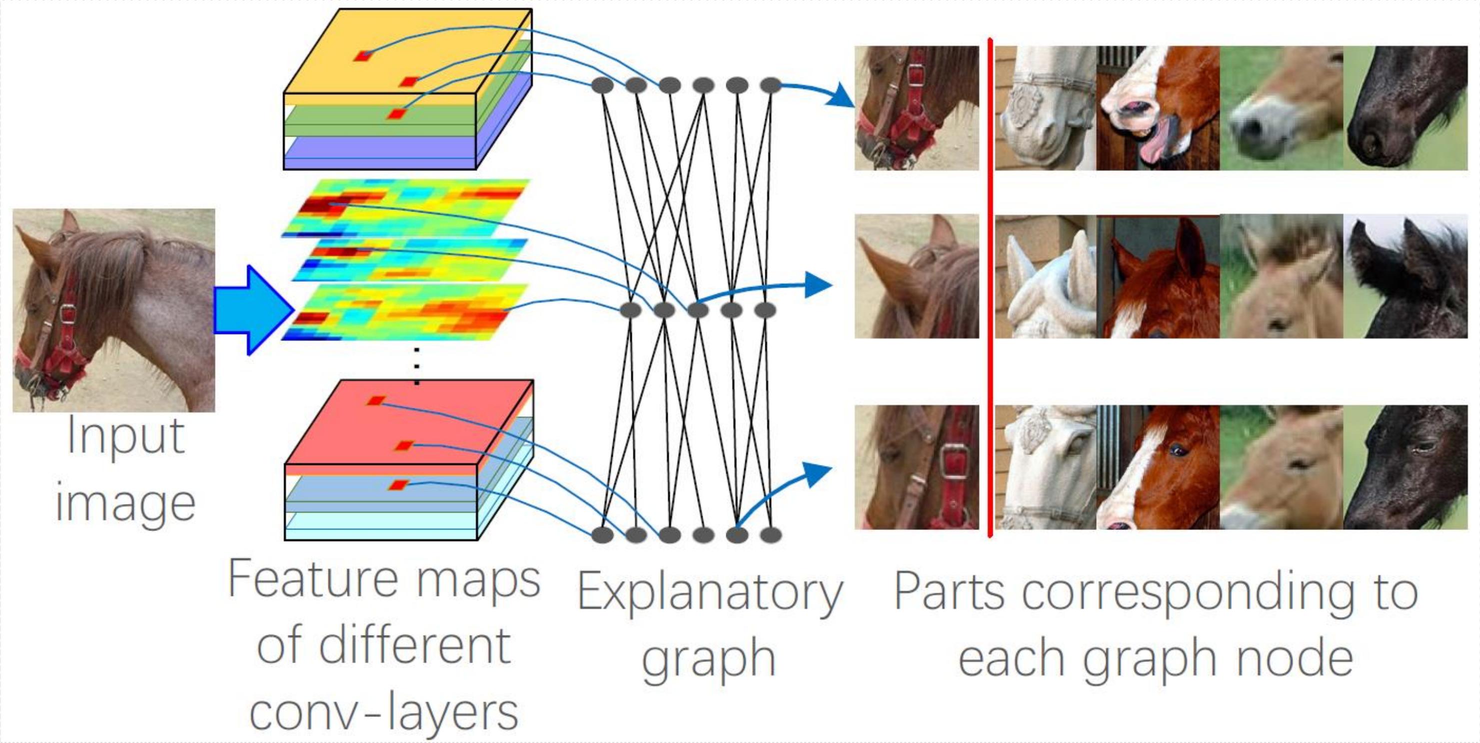
Strongly interacted filters → meaningful concepts

Wen Shen **Quanshi Zhang**

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

Wen Shen **Quanshi Zhang**

Objective

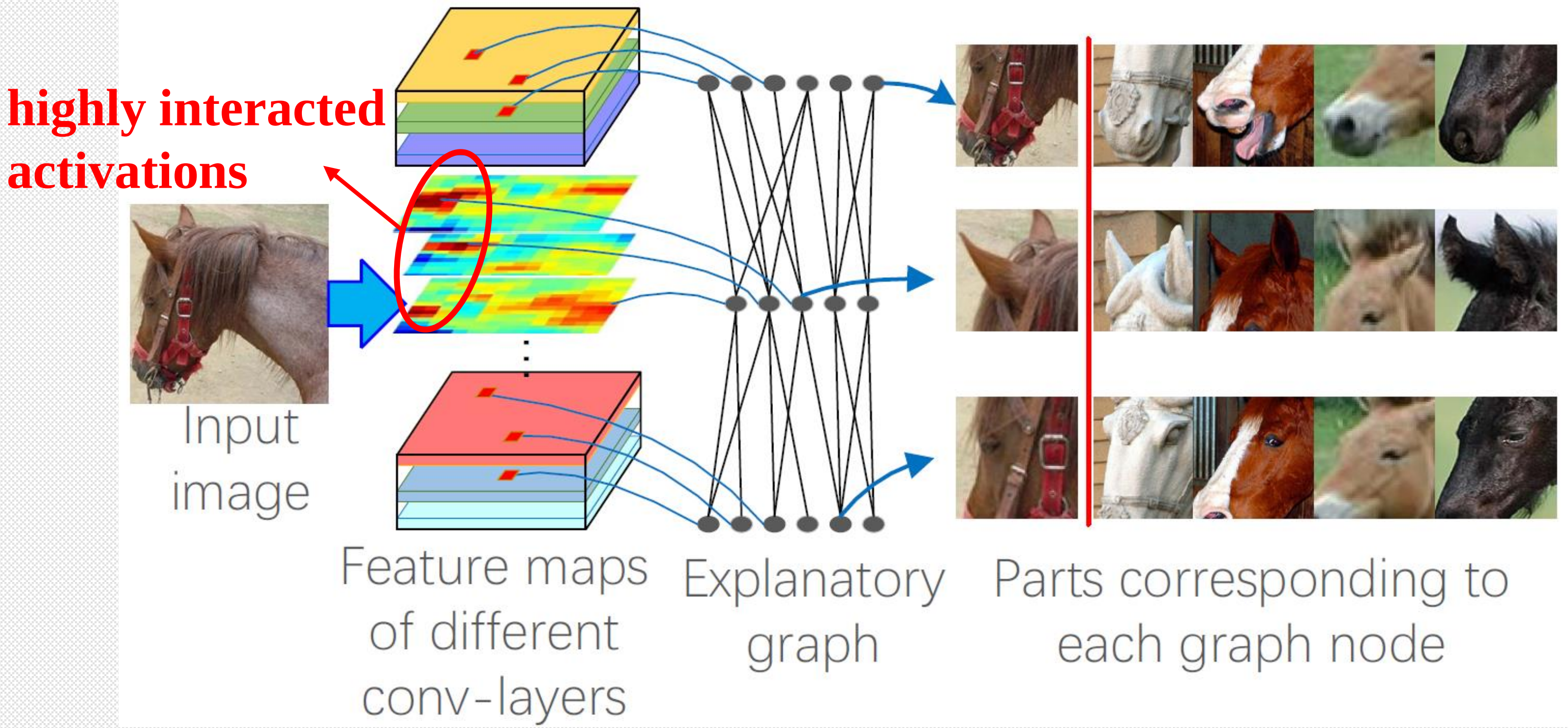


Build an explanatory graph to explain the semantic hierarchy hidden inside the network.

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


Wen Shen **Quanshi Zhang**

Objective

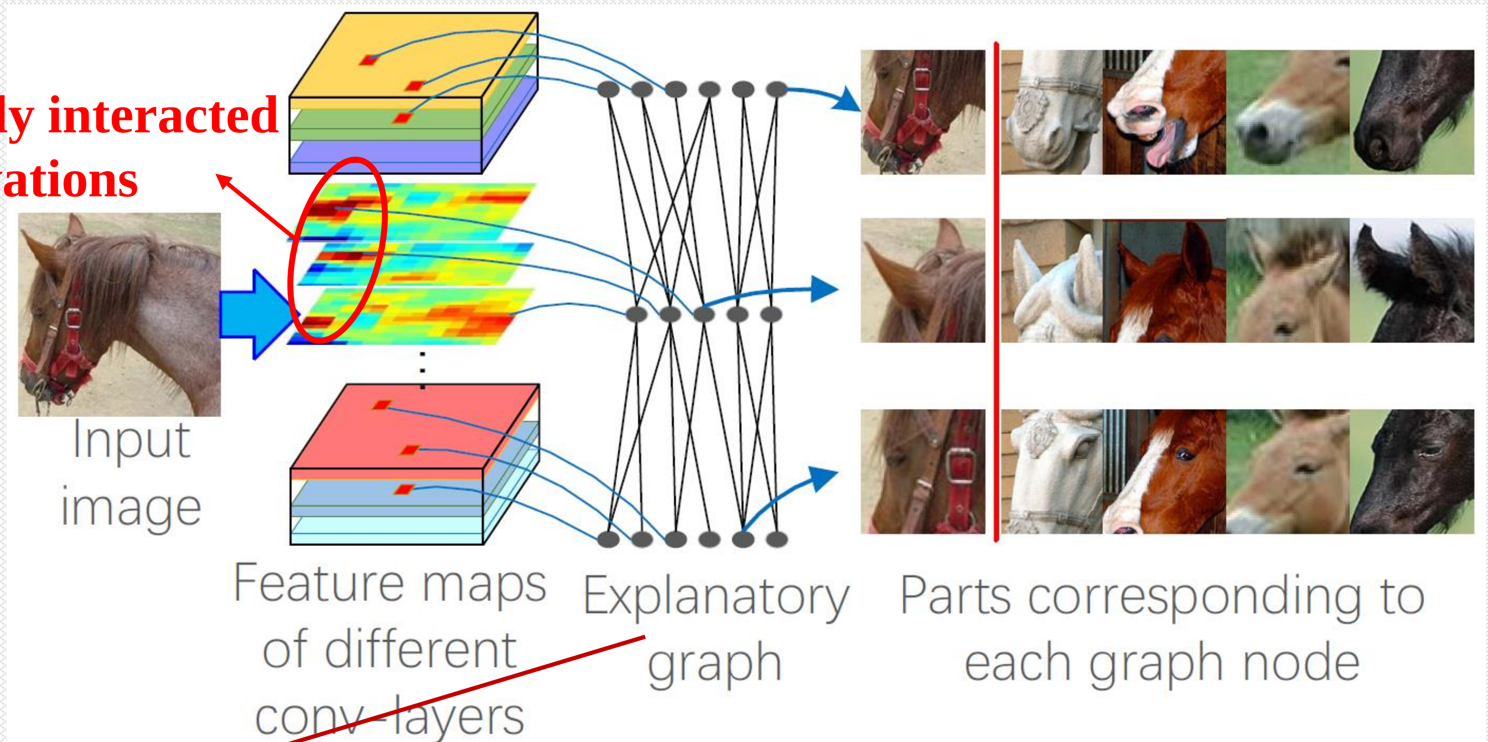


Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018


Wen Shen **Quanshi Zhang**

Objective

highly interacted activations



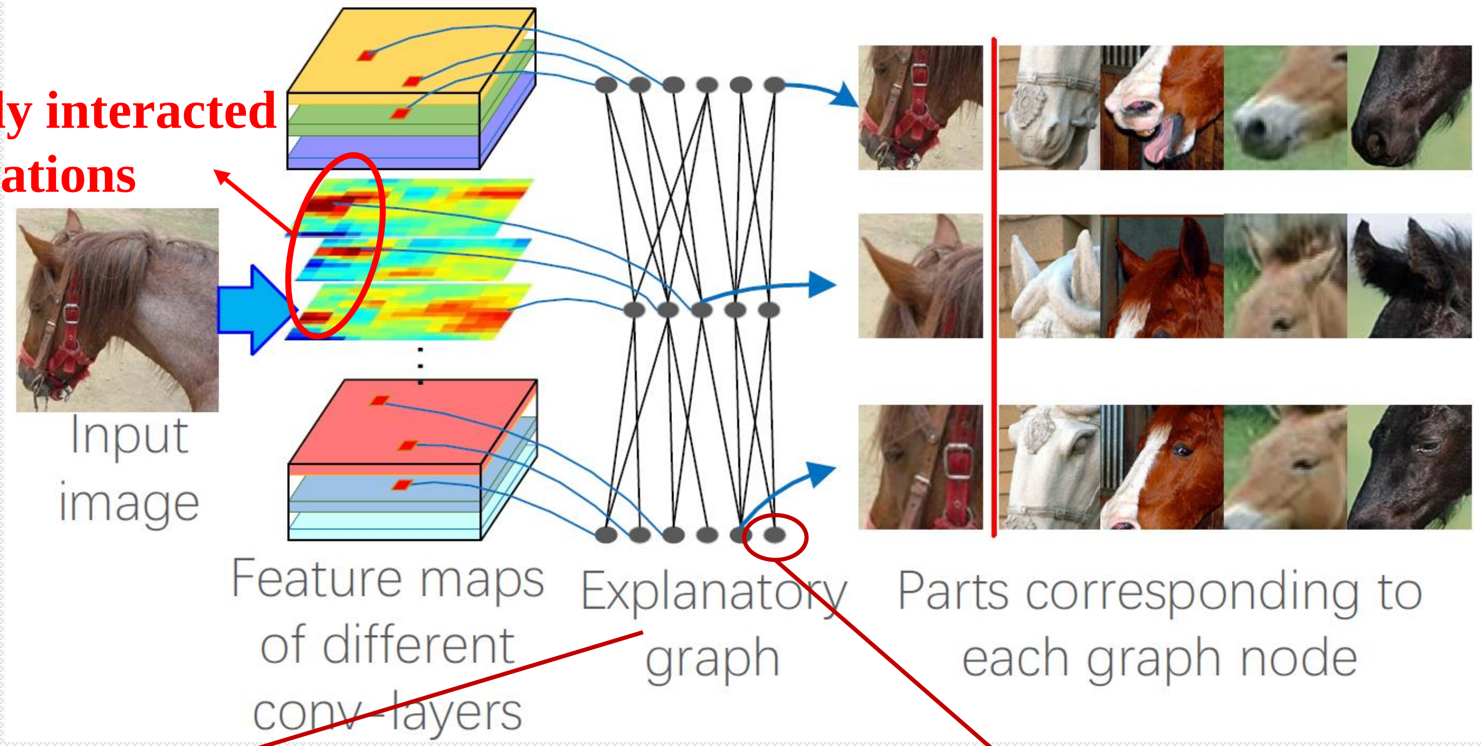
The graph has multiple layers →
multiple conv-layers of the CNN

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


Wen Shen **Quanshi Zhang**

Objective

highly interacted activations



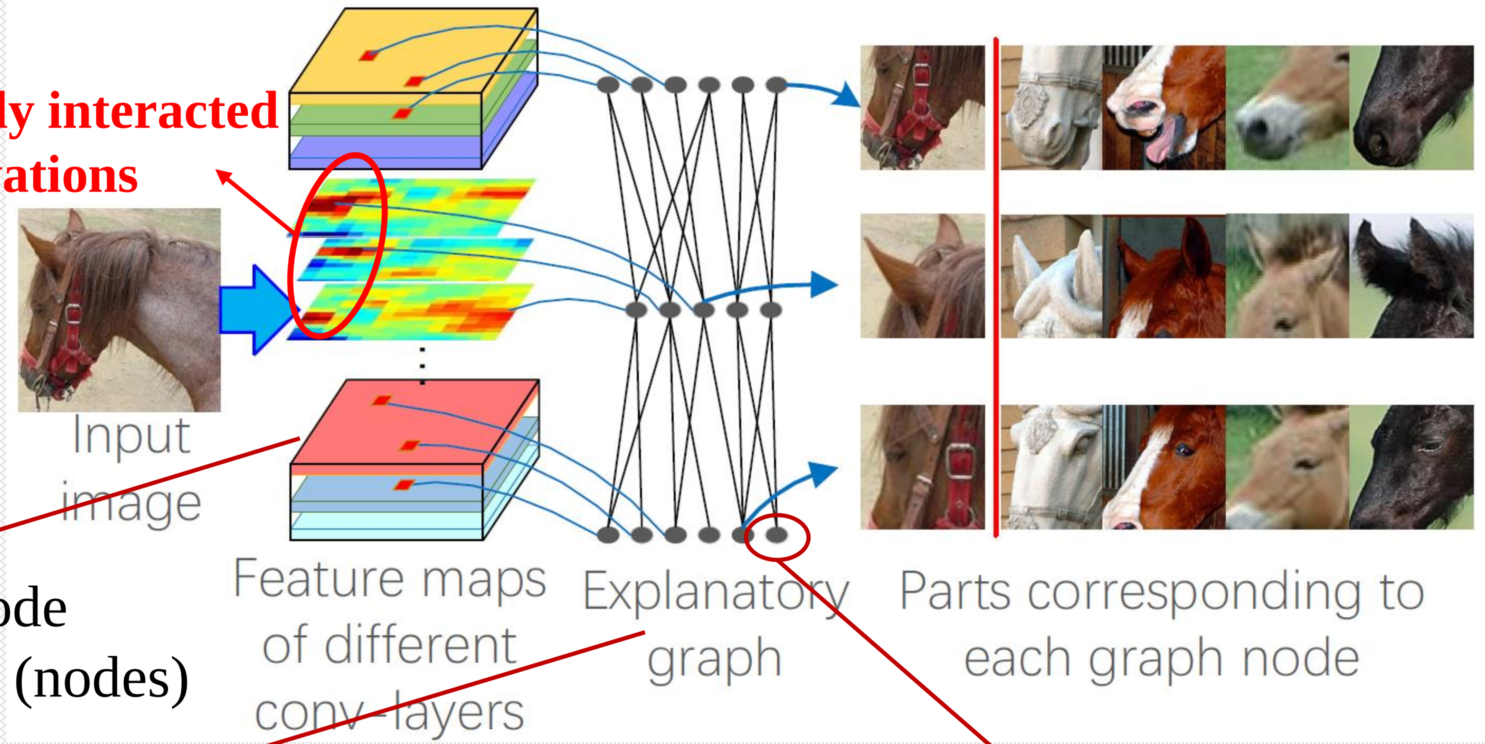
The graph has multiple layers → multiple conv-layers of the CNN

Each node → a pattern of an object part

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

Wen Shen **Quanshi Zhang**

Objective



A filter may encode multiple patterns (nodes)

The graph has multiple layers → multiple conv-layers of the CNN

Each node → a pattern of an object part

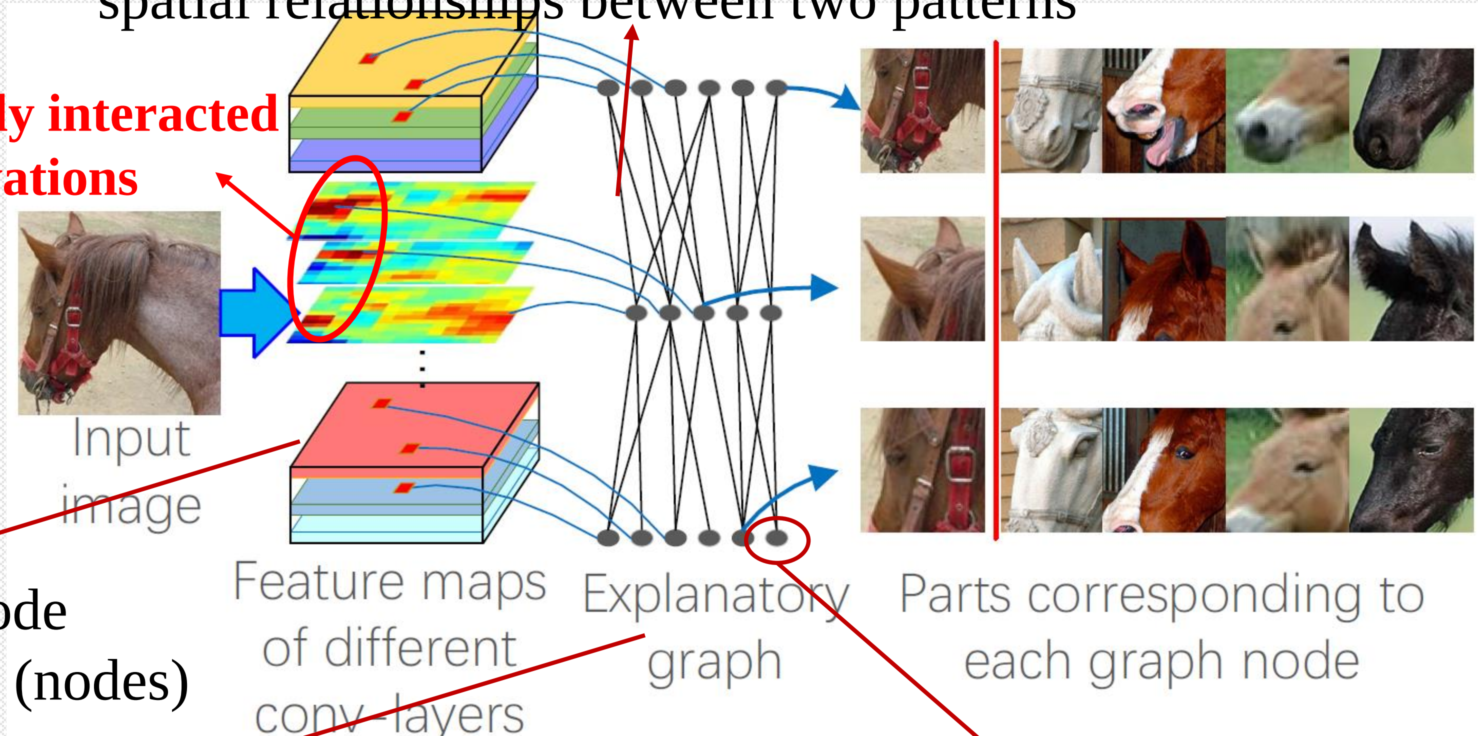
Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018


Wen Shen **Quanshi Zhang**

Objective

Each edge → co-activation relationships and spatial relationships between two patterns

highly interacted activations



A filter may encode multiple patterns (nodes)

The graph has multiple layers → multiple conv-layers of the CNN

Each node → a pattern of an object part

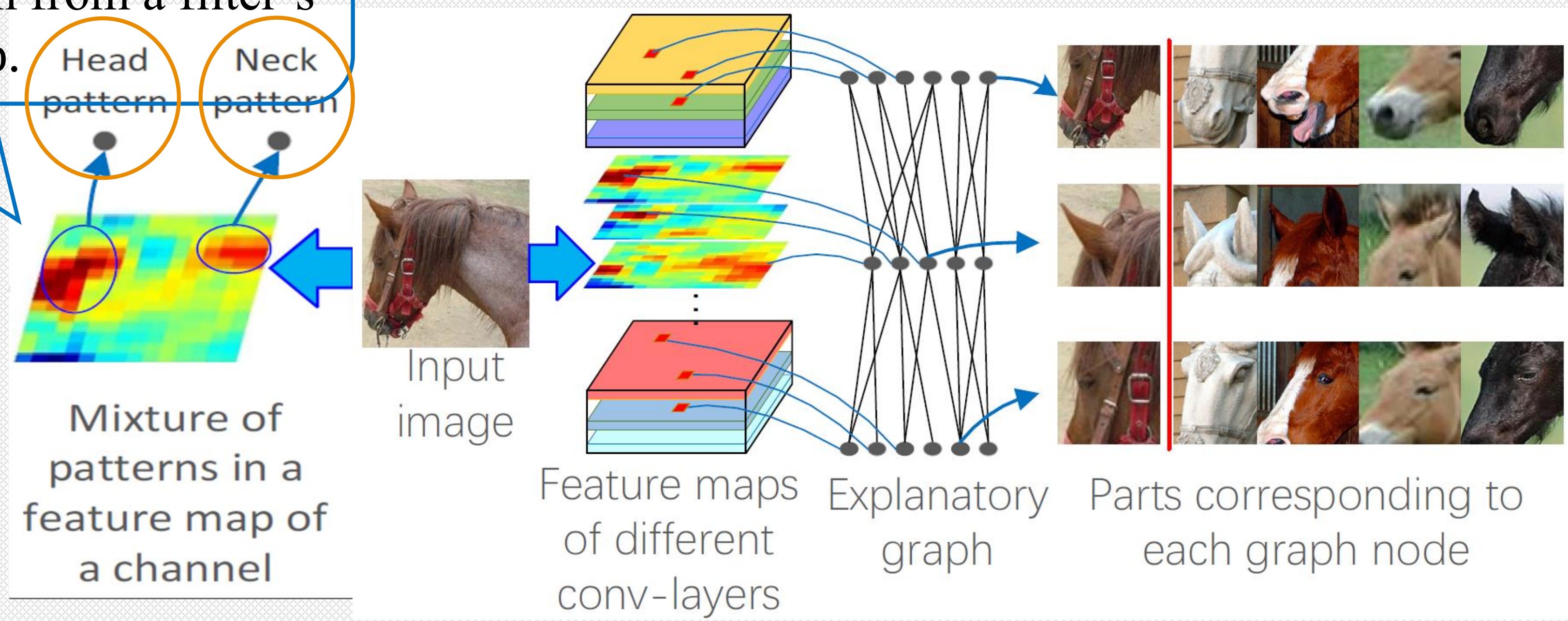
Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018



Wen Shen **Quanshi Zhang**

The explanatory graph

Disentangle the mixture of the head pattern and the neck pattern from a filter's feature map.



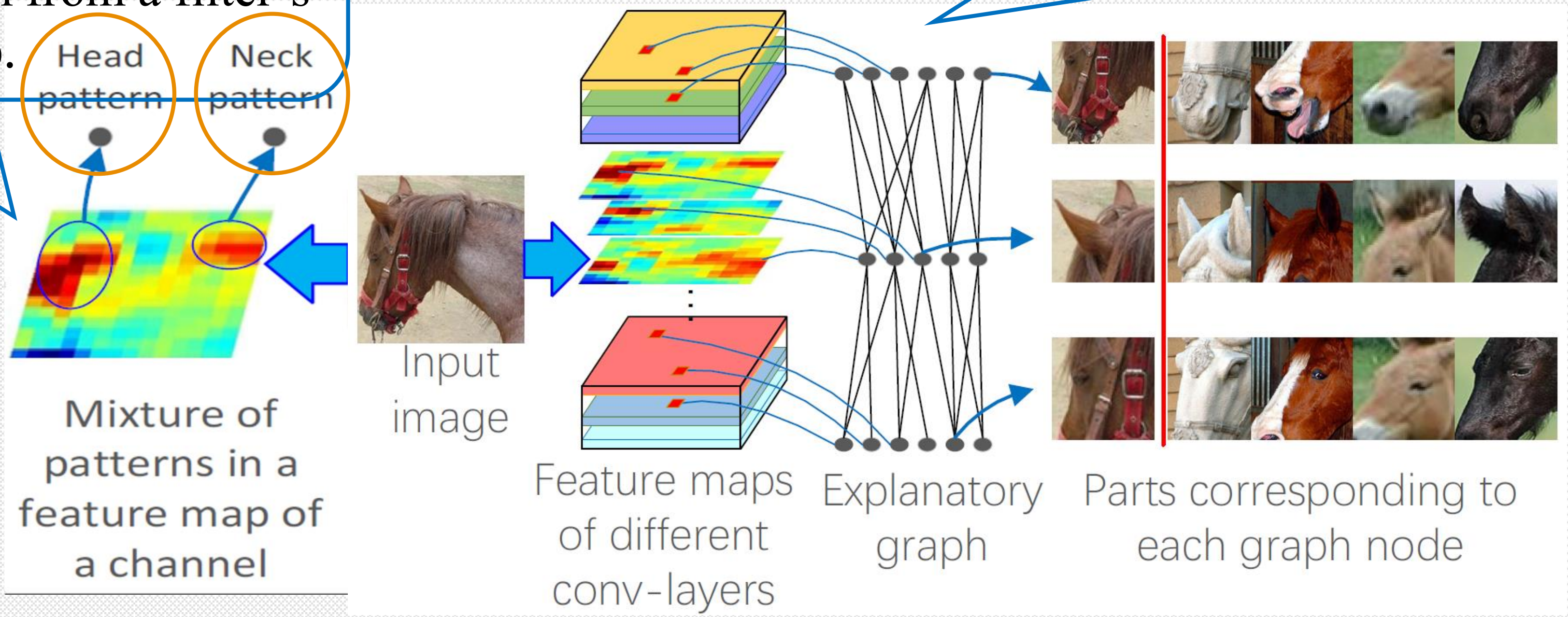
Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018


Wen Shen **Quanshi Zhang**

The explanatory graph

Disentangle the mixture of the head pattern and the neck pattern from a filter's feature map.

Summarize complex distributions of neural activations into a few patterns (nodes).



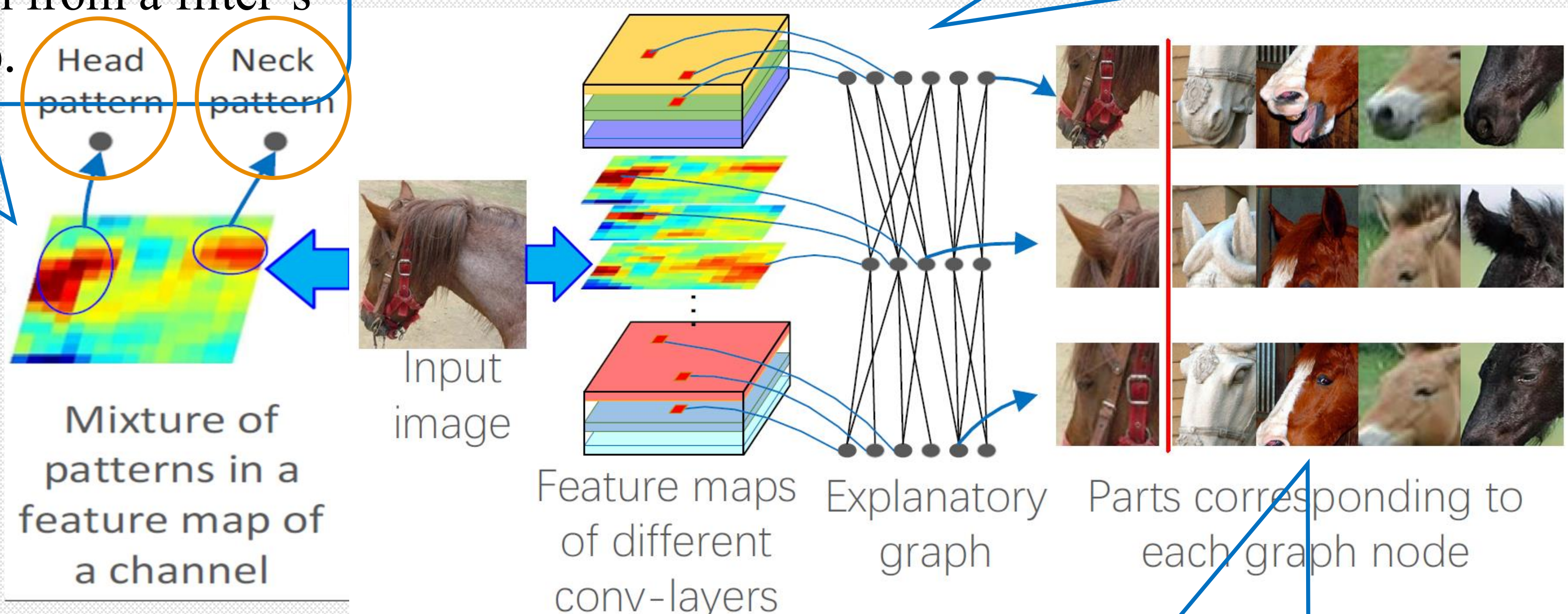
Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018


Wen Shen **Quanshi Zhang**

The explanatory graph

Disentangle the mixture of the head pattern and the neck pattern from a filter's feature map.

Summarize complex distributions of neural activations into a few patterns (nodes).



Each pattern consistently represents the same part among different images.

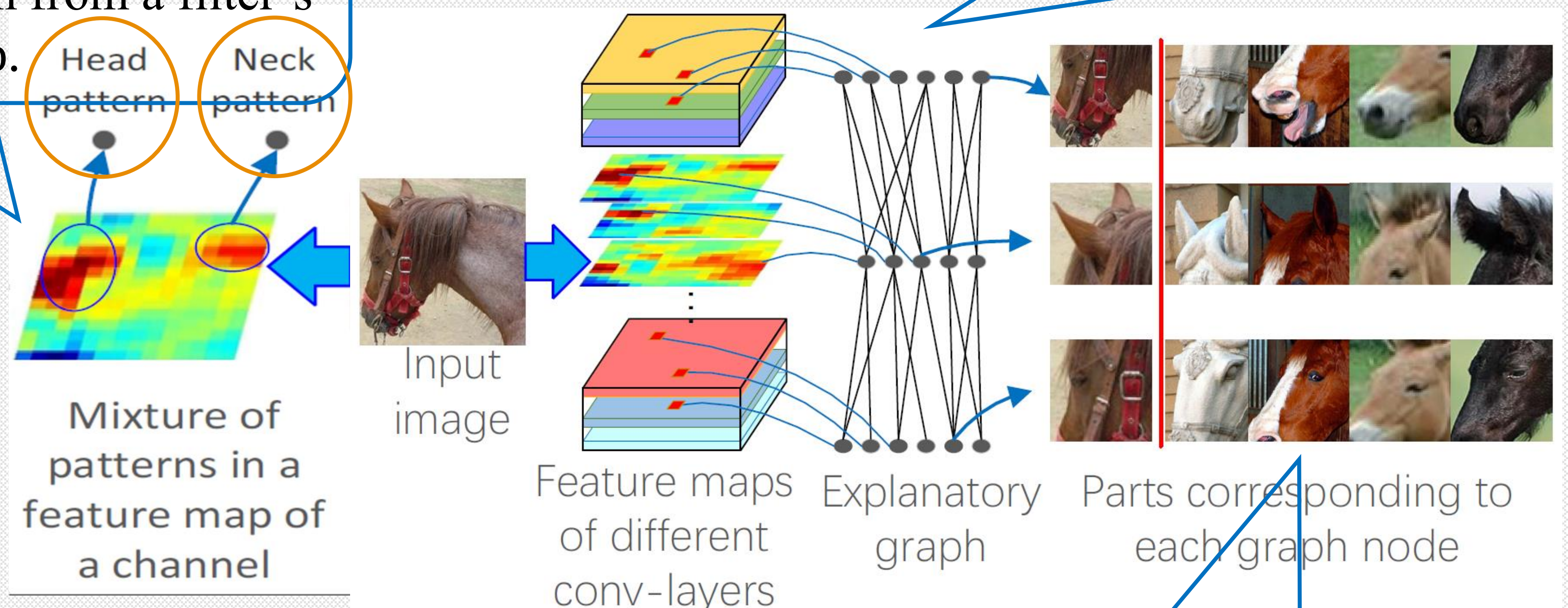
Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018


Wen Shen **Quanshi Zhang**

The explanatory graph

Disentangle the mixture of the head pattern and the neck pattern from a filter's feature map.

Summarize complex distributions of neural activations into a few patterns (nodes).



Filter out noisy activations on background from each feature map.

Each pattern consistently represents the same part among different images.

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


Wen Shen **Quanshi Zhang**

Input & Output

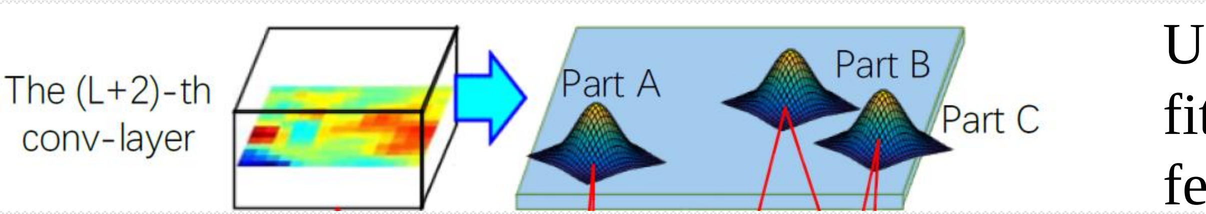
- Input:
  - A pre-trained CNN
    - trained for classification, segmentation, or ...
    - AlexNet, VGG-16, ResNet-50, ResNet-152, and etc.
  - Its training images with object bounding boxes

- Output: an explanatory graph

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


Wen Shen **Quanshi Zhang**

Mining an explanatory graph



Use a mixture of patterns to fit activation distributions of a feature map (just like GMM)

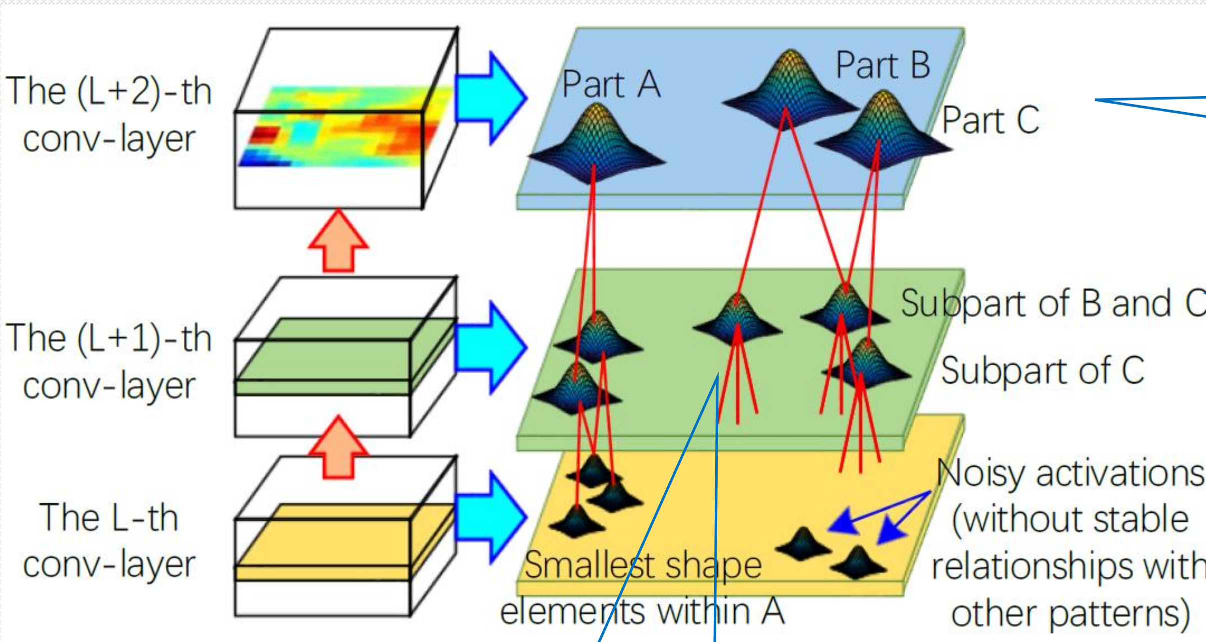
a feature map of a filter

→ a distribution of "activation entities"

[2] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018


Wen Shen **Quanshi Zhang**

Mining an explanatory graph



Patterns for large parts

Patterns for subparts

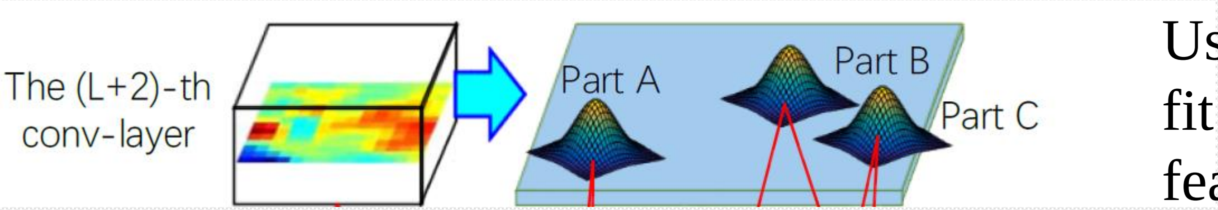
Patterns for even smaller parts

Edges: spatial relationships between co-activated patterns

[2] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018

Wen Shen **Quanshi Zhang**

Mining an explanatory graph



Use a mixture of patterns to fit activation distributions of a feature map (just like GMM)

Need to learn

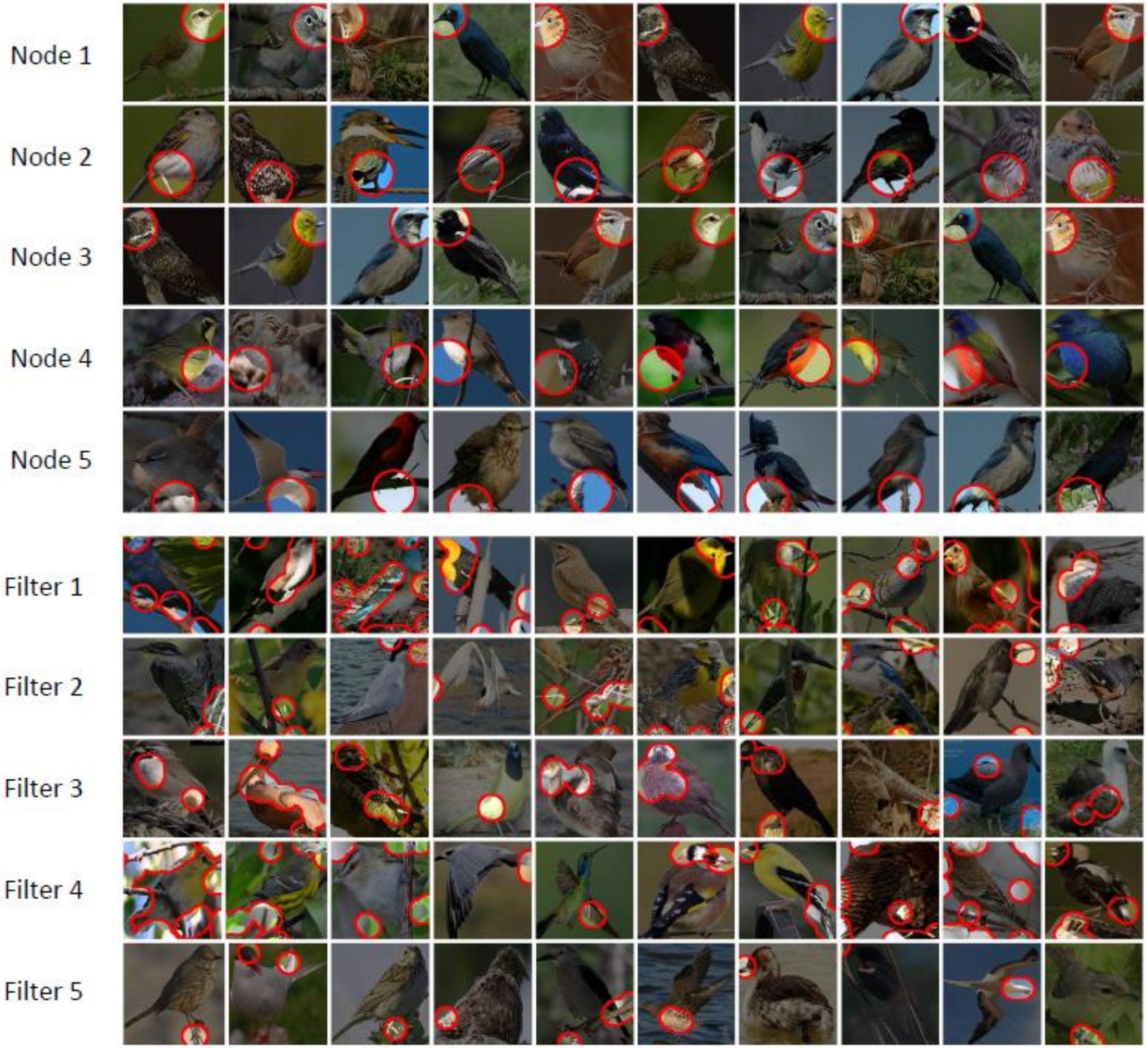
1. Connections between nodes
2. Spatial relationships between connected nodes

Use such spatial relationships to disentangle feature maps of conv-layers.

[2] Quanshi Zhang et al. "Interpreting CNN Knowledge via an Explanatory Graph" in AAAI 2018

Strongly interacted filters → meaningful concepts


Wen Shen **Quanshi Zhang**



Performance of nodes in the explanatory graph

Disentangle each pattern component from each filter's feature map.

Performance of raw filters in the CNN

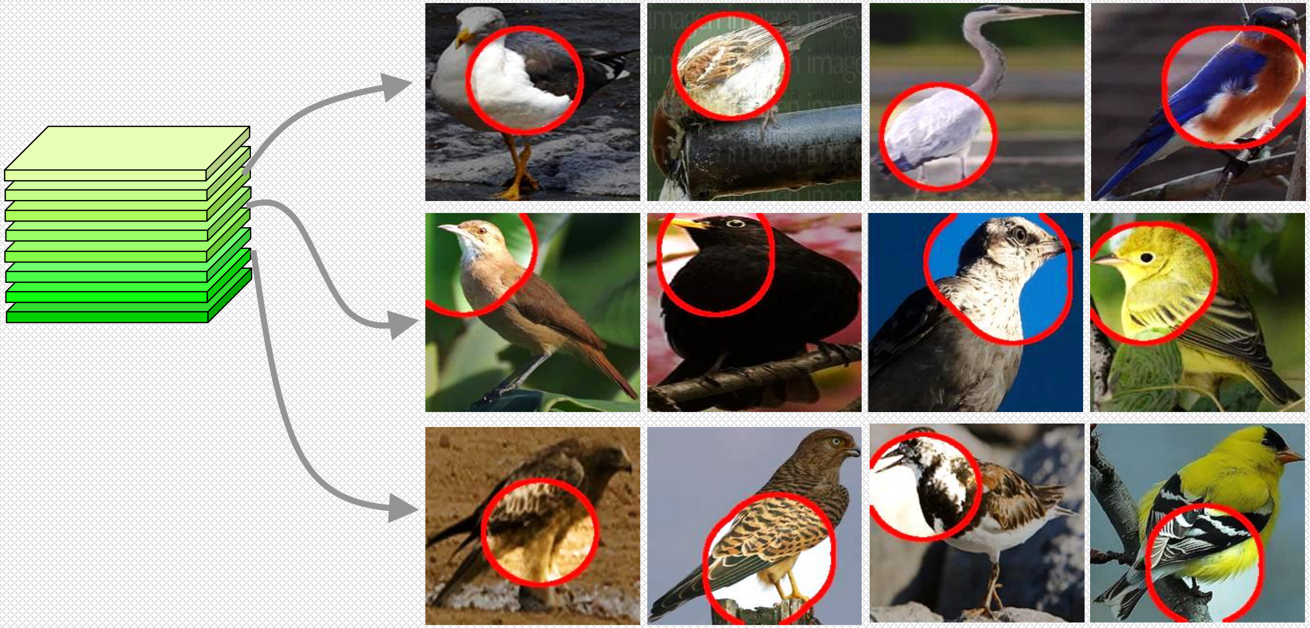

Wen Shen **Quanshi Zhang**

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


Wen Shen **Quanshi Zhang**

Objective

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



Feature maps of Filter 1

Feature maps of Filter 2

Feature maps of Filter 3

Neural activations of 3 interpretable filters

[1] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018


Wen Shen **Quanshi Zhang**

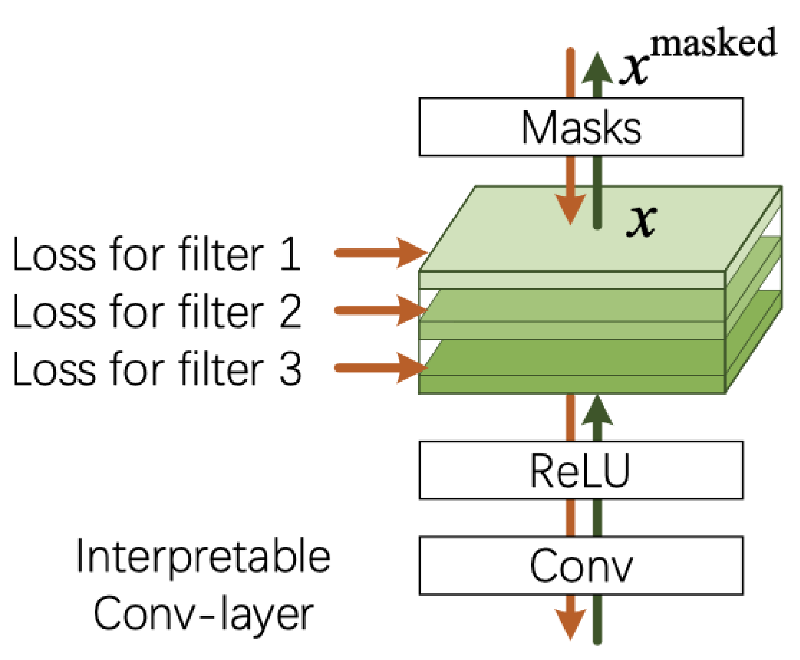
Input & Output

- Input
  - Training samples (X_i, Y_i) for a certain task
  - **No** annotations of parts or textures are used.
- Output
  - An interpretable CNN with disentangled filters


Wen Shen **Quanshi Zhang**

Network structure

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 templates $\boldsymbol{T}$.

$$Loss_f = -MI(\mathbf{X}; \mathbf{T}) \qquad \text{for filter } f$$

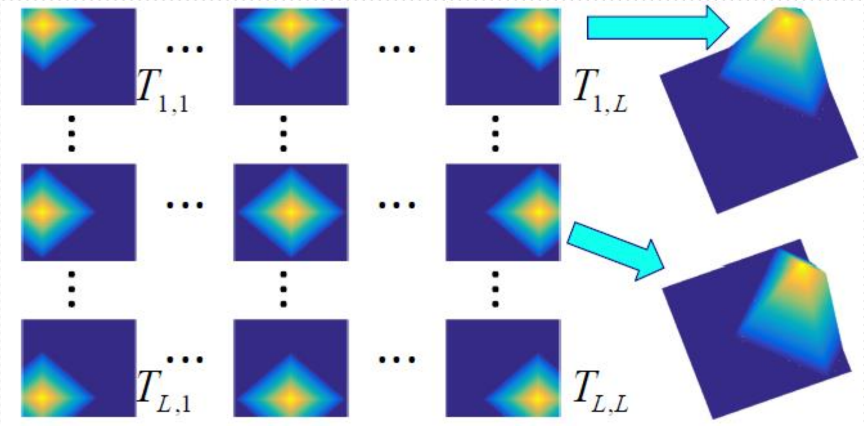
Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018


Wen Shen **Quanshi Zhang**

Network structure

Understanding the filter loss: the filter loss boosts the mutual information between feature maps ***X*** and a set of pre-defined templates ***T***.

$$Loss_f = -MI(\mathbf{X};\mathbf{T})$$



templates ***T***

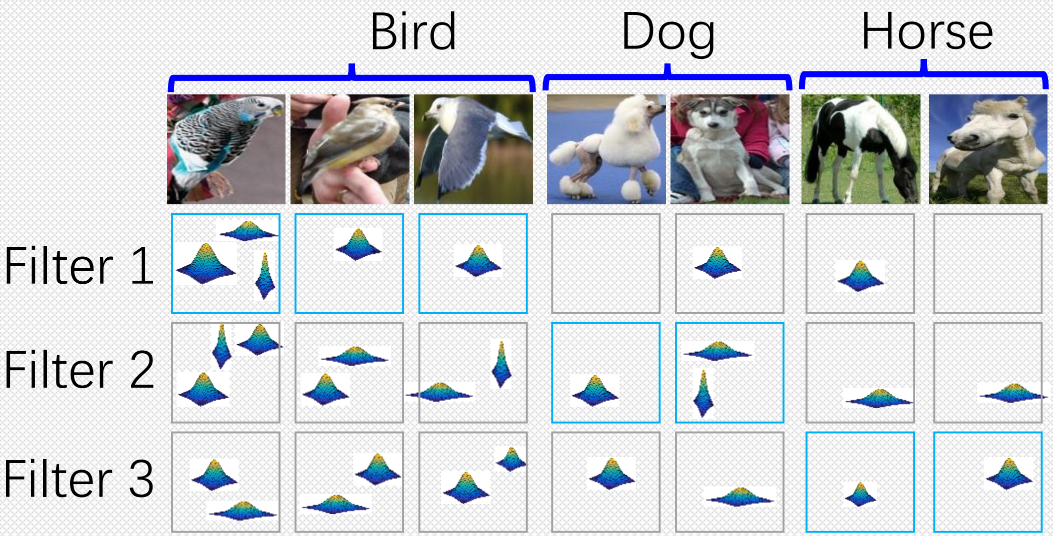
$$Loss_f = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X})}_{\text{Entropy of inter-category activations}} + \sum_x p(\mathbf{T}^+, x)\underbrace{H\left(\mathbf{T}^+ = \{T_\mu\}\middle|X{=}x\right)}_{\text{Entropy of the spatial distribution of activations}}$$


Wen Shen **Quanshi Zhang**

Learning

From chaotic feature maps to the disentangled maps of object parts

$$Loss_f = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X})}_{\text{Entropy of inter-category activations}} + \sum_x p(\mathbf{T}^+, x) \underbrace{H\left(\mathbf{T}^+ = \{T_\mu\}|X=x\right)}_{\text{Entropy of the spatial distribution of activations}}$$



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

Strongly interacted filters → meaningful concepts

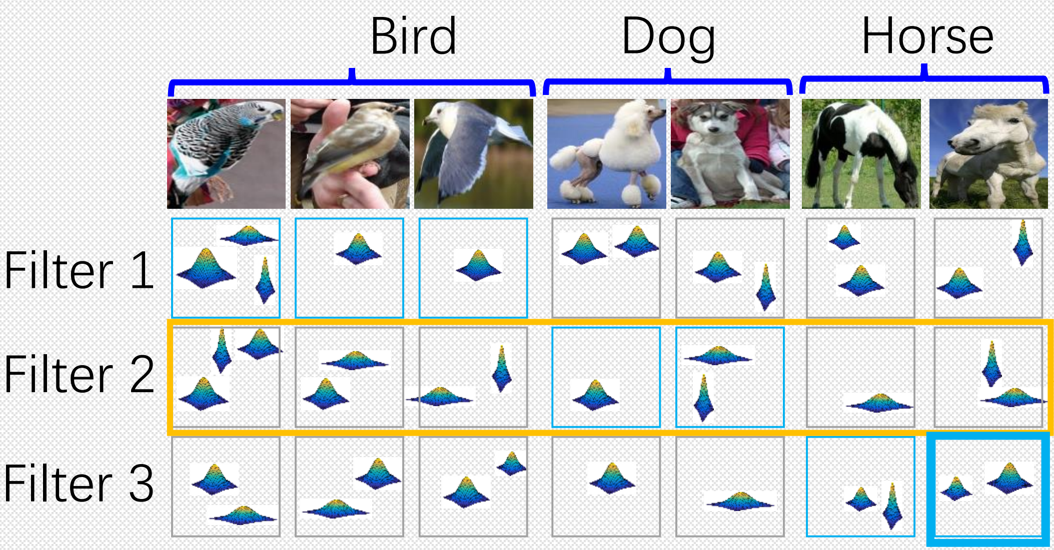


Wen Shen **Quanshi Zhang**

Learning

From chaotic feature maps to the disentangled maps of object parts

$$Loss_f = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\} | \mathbf{X})}_{\text{Entropy of inter-category activations}} + \sum_x p(\mathbf{T}^+, x) \underbrace{H\big(\mathbf{T}^+ = \{T_\mu\} \big| X = x\big)}_{\text{Entropy of the spatial distribution of activations}}$$



The loss encourages a low inter-class entropy, i.e., **increasing regional activations with strong interactions.**

The loss encourages a low spatial entropy.

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

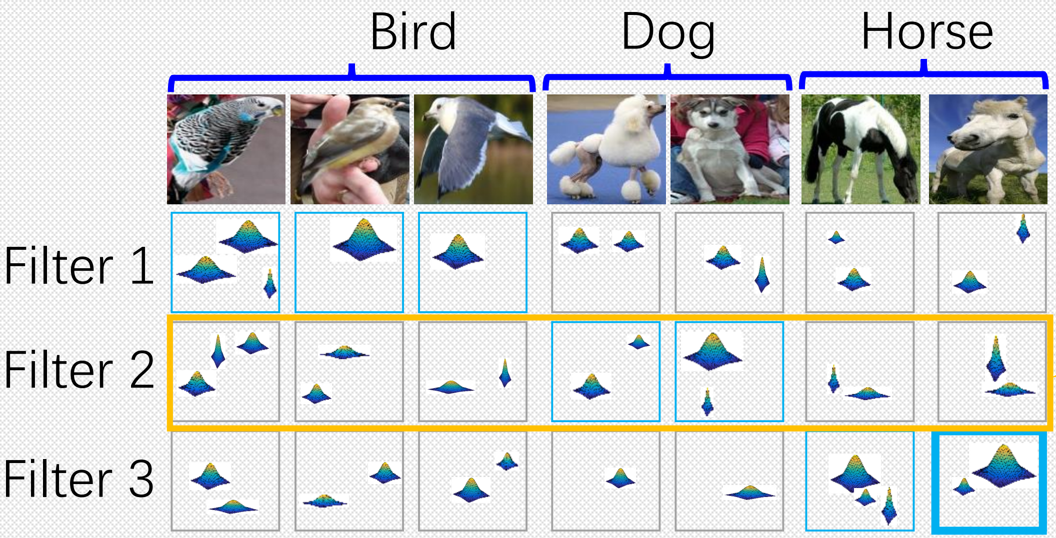
Strongly interacted filters → meaningful concepts


Wen Shen **Quanshi Zhang**

Learning

From chaotic feature maps to the disentangled maps of object parts

$$Loss_f = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X})}_{\text{Entropy of inter-category activations}} + \sum_x p(\mathbf{T}^+, x) \underbrace{H\left(\mathbf{T}^+ = \{T_\mu\} \middle| X{=}x\right)}_{\text{Entropy of the spatial distribution of activations}}$$



The loss encourages a low inter-class entropy, i.e., **increasing regional activations with strong interactions.**

The loss encourages a low spatial entropy.

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

Strongly interacted filters → meaningful concepts

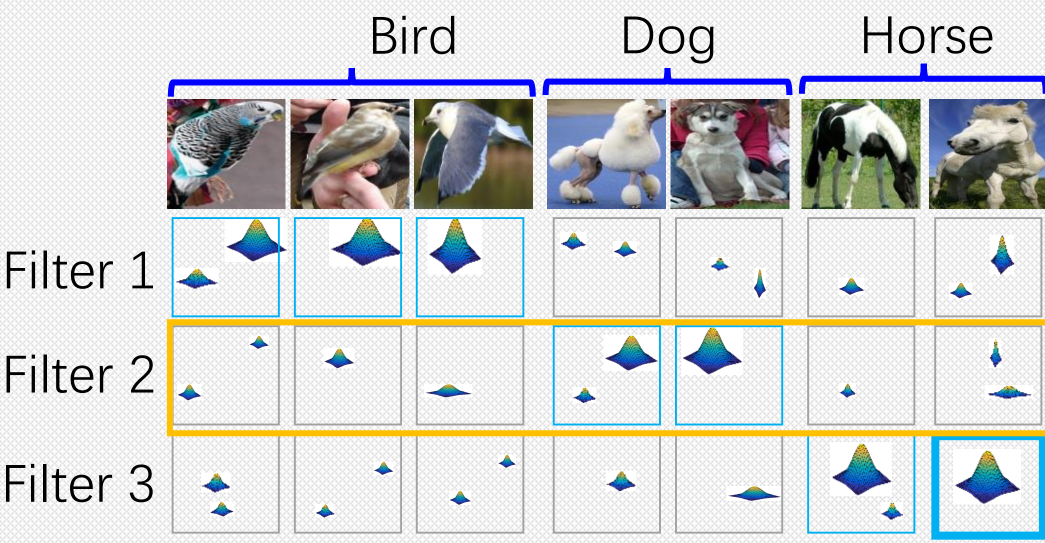
Wen Shen

Quanshi Zhang

Learning

From chaotic feature maps to the disentangled maps of object parts

$$Loss_f = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X})}_{\text{Entropy of inter-category activations}} + \sum_x p(\mathbf{T}^+, x)\underbrace{H\left(\mathbf{T}^+ = \{T_\mu\}\middle|X{=}x\right)}_{\text{Entropy of the spatial distribution of activations}}$$



The loss encourages a low inter-class entropy, i.e., **increasing regional activations with strong interactions.**

The loss encourages a low spatial entropy.

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

Strongly interacted filters → meaningful concepts

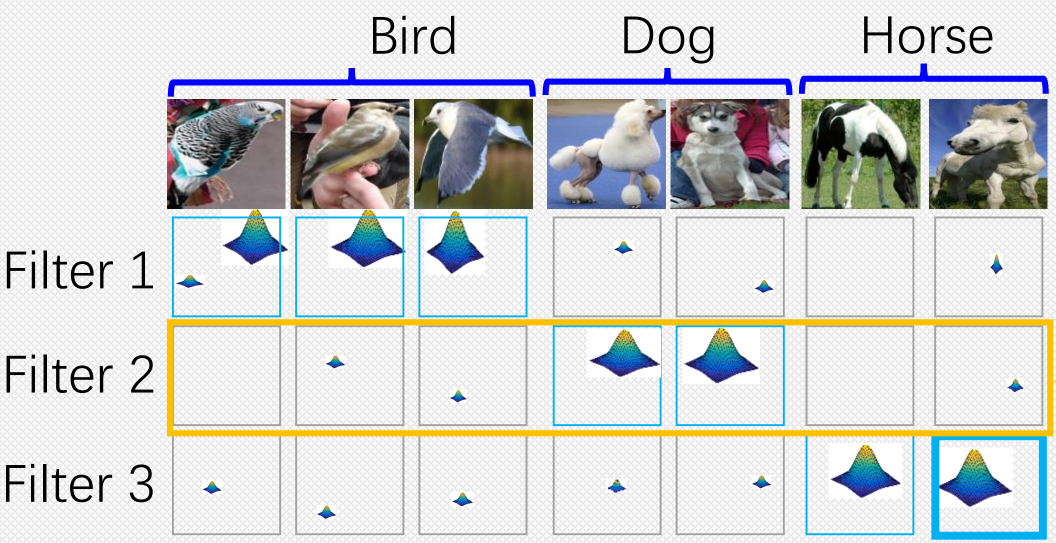


Wen Shen **Quanshi Zhang**

Learning

From chaotic feature maps to the disentangled maps of object parts

$$Loss_f = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X})}_{\text{Entropy of inter-category activations}} + \sum_x p(\mathbf{T}^+, x)\underbrace{H\left(\mathbf{T}^+ = \{T_\mu\}\middle|X{=}x\right)}_{\text{Entropy of the spatial distribution of activations}}$$



The loss encourages a low inter-class entropy, i.e., **increasing regional activations with strong interactions.**

The loss encourages a low spatial entropy.

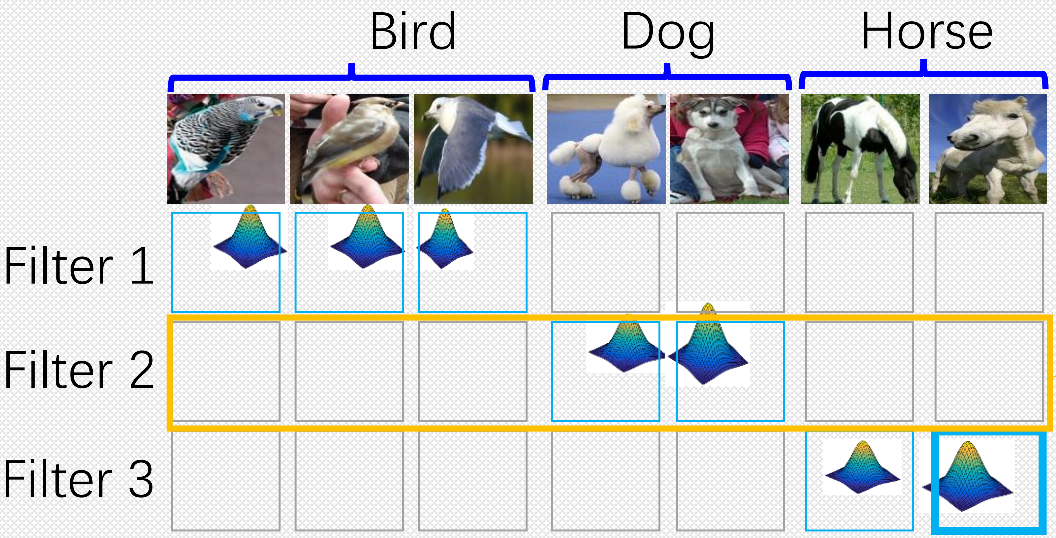
Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018


Wen Shen **Quanshi Zhang**

Learning

From chaotic feature maps to the disentangled maps of object parts

$$Loss_f = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X})}_{\text{Entropy of inter-category activations}} + \sum_x p(\mathbf{T}^+, x) \underbrace{H\left(\mathbf{T}^+ = \{T_\mu\}\middle|X{=}x\right)}_{\text{Entropy of the spatial distribution of activations}}$$



The loss encourages a low inter-class entropy, i.e., **increasing regional activations with strong interactions.**

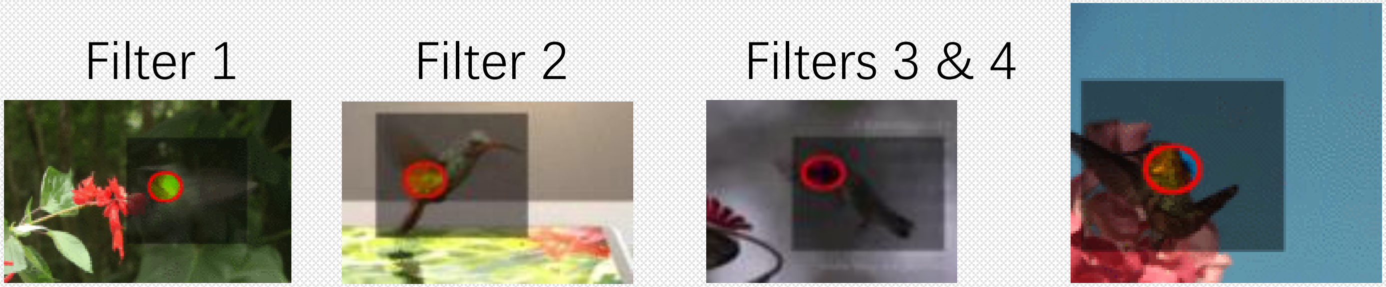
The loss encourages a low spatial entropy.

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


Wen Shen **Quanshi Zhang**

Activation regions of interpretable filters.

Filter 1 Filter 2 Filters 3 & 4



Filter
Filter
Filter
Filter

Filter
Filter
Filter
Filter

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

Strongly interacted filters → meaningful concepts

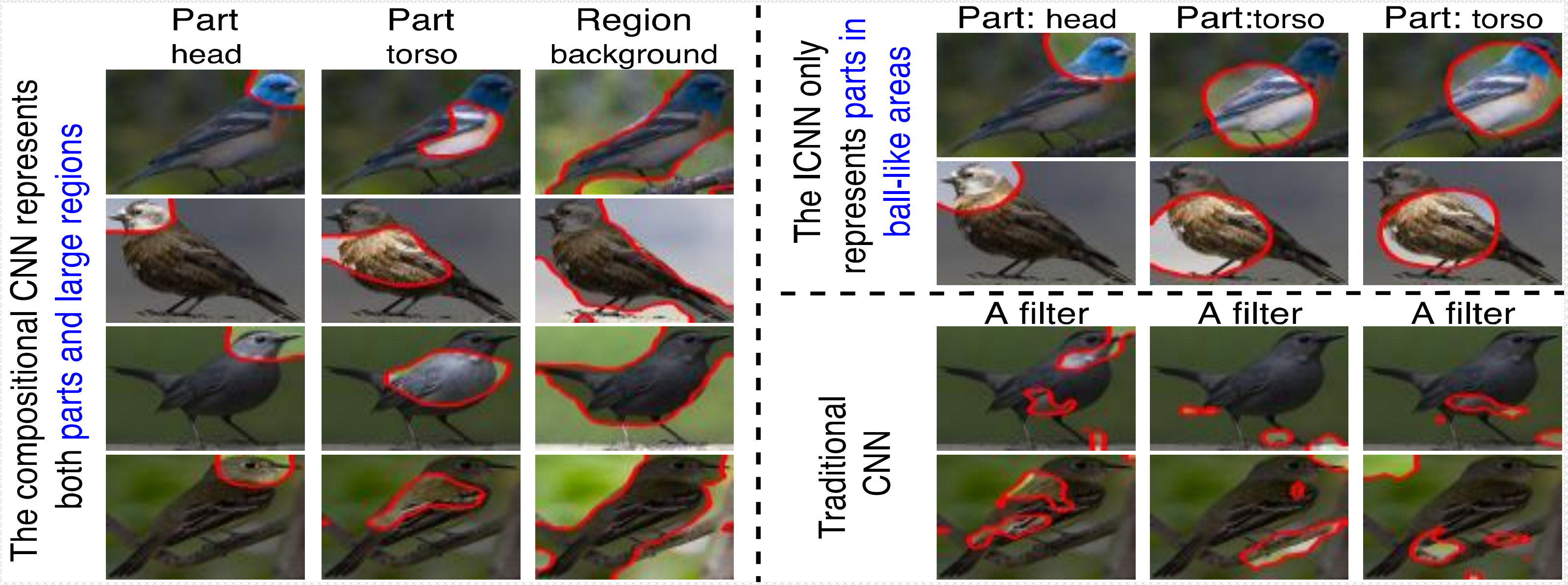
Wen Shen

Quanshi Zhang

Wen Shen et al. "Interpretable Compositional Convolutional Neural Networks" in IJCAI 2021

Objective


Wen Shen **Quanshi Zhang**



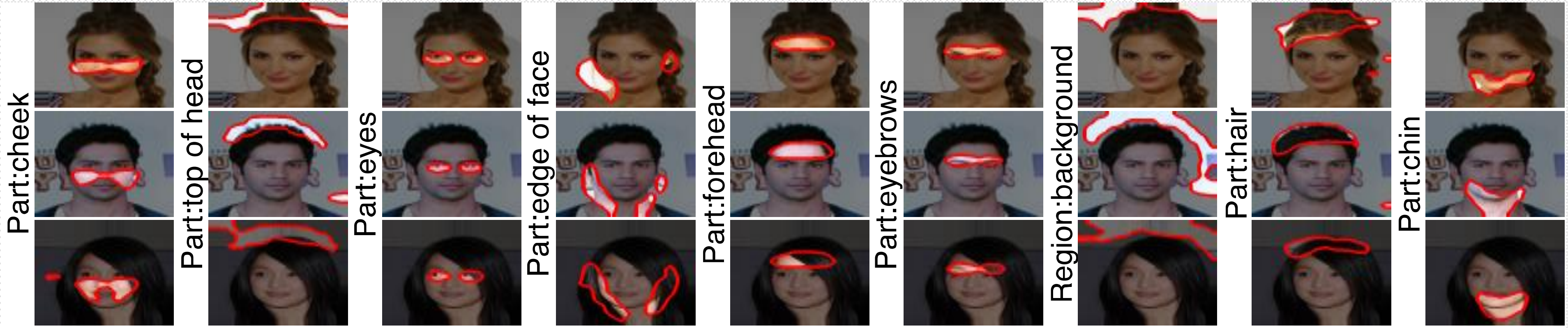
- Traditional CNN: has no self-reflection of its representations.
- ICNN[1]: only represents object parts in ball-like areas.
- **Our compositional CNN: represents both object parts with specific shapes and image regions without specific structures.**

[1] Quanshi Zhang et al. "Interpretable Convolutional Neural Networks" in CVPR 2018

Wen Shen **Quanshi Zhang**

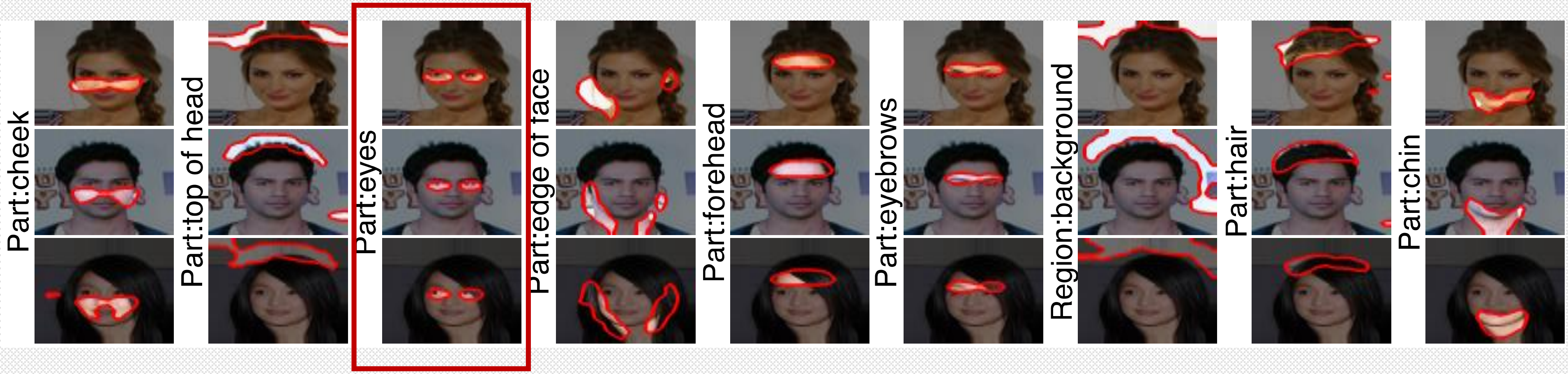
Objective



- Compositional interpretable filters should satisfy the following **two properties.**
  - ➢ **Consistency.**
  - ➢ **Diversity.**


Wen Shen **Quanshi Zhang**

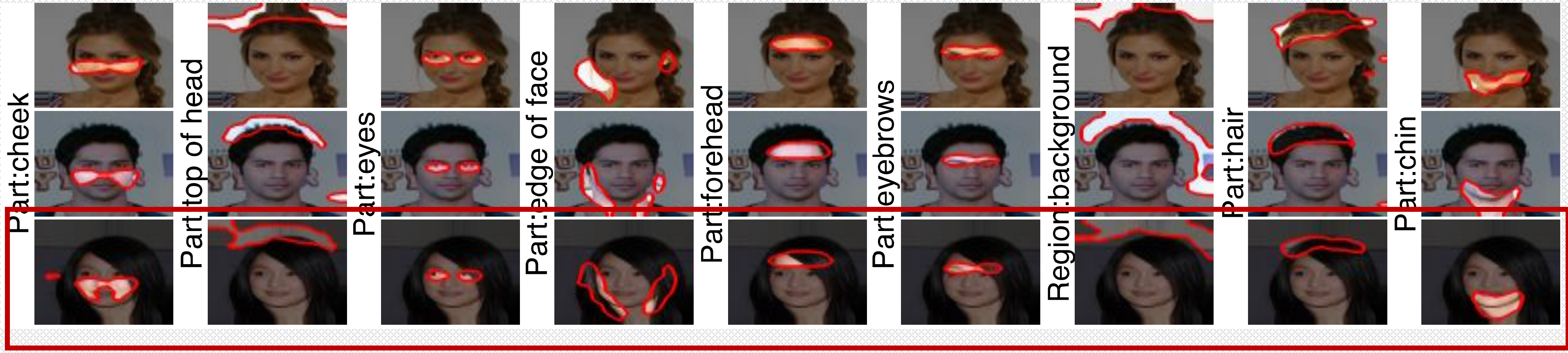
Objective



- Compositional interpretable filters should satisfy the following **two properties.**
  - ➢ **Consistency.** Each filter is supposed to be consistently activated by the same object part or the same image region through different images.
  - ➢ **Diversity.**

Wen Shen **Quanshi Zhang**

Objective



- Compositional interpretable filters should satisfy the following **two properties.**
  - **Consistency.** Each filter is supposed to be consistently activated by the same object part or the same image region through different images.
  - **Diversity.** Different filters are supposed to be activated by different object parts or image regions.

Input & Output


Wen Shen **Quanshi Zhang**

- Input
  - Training samples (X_i,Y_i) for a certain task.
  - **No** annotations of object parts or image regions are used.
- Output
  - An interpretable compositional CNN with disentangled filters.



Wen Shen **Quanshi Zhang**

Method

- To satisfy the properties of consistency and diversity
  - use a set of filters to jointly represent a specific part/region, instead of using a single filter → **consistency**

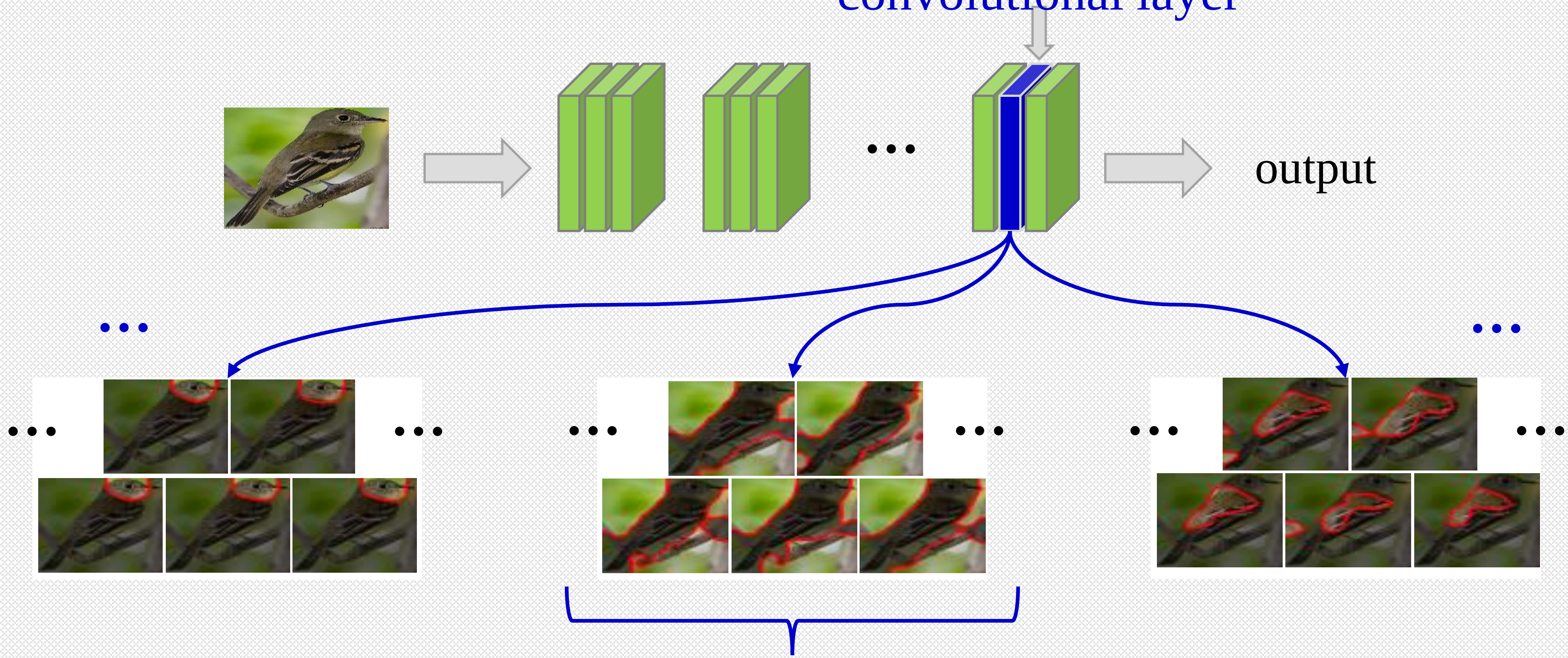
Strongly interacted filters → meaningful concepts

Wen Shen

Quanshi Zhang

Method

An interpretable compositional convolutional layer



A group of filters **cooperate with each other** to make inferences → **Consistency**

The cooperative features have strong interactions.

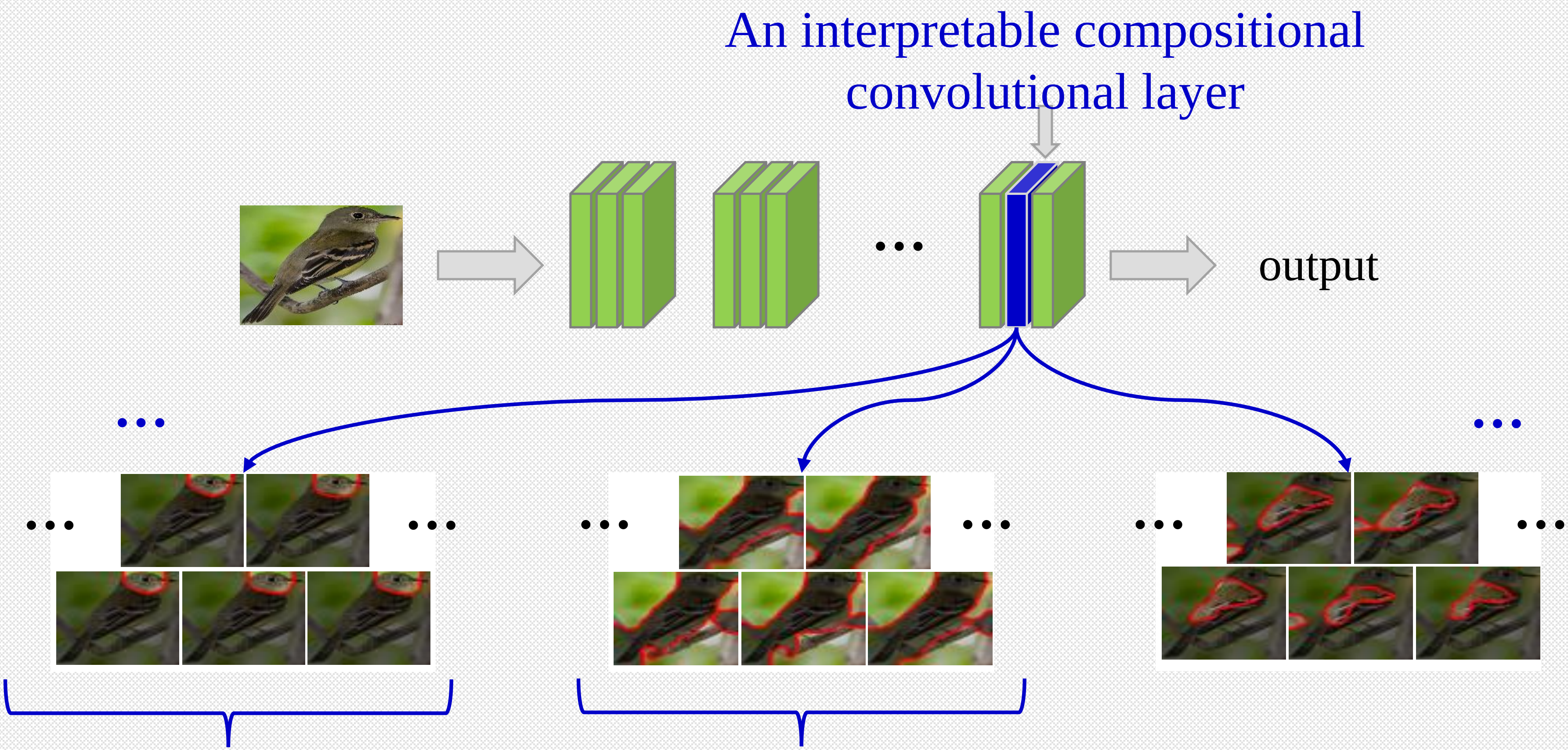
Wen Shen **Quanshi Zhang**

Method

- To satisfy the properties of consistency and diversity
  - use a set of filters to jointly represent a specific part/region, instead of using a single filter → **consistency**
  - use different sets of filters to represent different parts/regions → **diversity**

Wen Shen **Quanshi Zhang**

Method



Different groups of filters represent different parts/regions. → **Diversity**

Features of filters in different groups have weak interactions.

Wen Shen **Quanshi Zhang**

Method

- To satisfy the properties of consistency and diversity
  - use a set of filters to jointly represent a specific part/region, instead of using a single filter → **consistency**
  - use different sets of filters to represent different parts/regions → **diversity**

- We add a loss to the target convolutional layer to construct an compositional interpretable layer

$$\mathbf{L}(\theta, \mathbf{A}) = \underbrace{\lambda \mathrm{Loss}(\theta, \mathbf{A})}_{\text{filter loss}} + \frac{1}{n}\sum\nolimits_{I \in \mathbf{I}} \underbrace{\mathrm{L}^{\mathrm{cls}}(\hat{y}_I, y_I^*; \theta)}_{\text{task loss}},$$

Wen Shen **Quanshi Zhang**

Method

- To satisfy the properties of consistency and diversity
  - use a set of filters to jointly represent a specific part/region, instead of using a single filter → **consistency**
  - use different sets of filters to represent different parts/regions → **diversity**

$$\mathrm{Loss}(\theta, \mathbf{A}) = -\sum_{k=1}^{K} \frac{S_k^{\text{within}}}{S_k^{\text{all}}} = -\sum_{k=1}^{K} \frac{\sum_{i,j \in A_k} \boxed{s_{ij}}}{\sum_{i \in A_k, j \in \Omega} s_{ij}}$$

Measure the similarity between filters in the group A_k

These four filters have similar activation regions (i.e. these filters have **strong interactions**)


Wen Shen **Quanshi Zhang**

Method

- To satisfy the properties of consistency and diversity
  - use a set of filters to jointly represent a specific part/region, instead of using a single filter → **consistency**
  - use different sets of filters to represent different parts/regions → **diversity**

$$\mathrm{Loss}(\theta, \mathbf{A}) = -\sum_{k=1}^{K} \frac{S_k^{\text{within}}}{S_k^{\text{all}}} = -\sum_{k=1}^{K} \frac{\sum_{i,j\in A_k} \boxed{s_{ij}}}{\sum_{i\in A_k, j\in\Omega} s_{ij}}$$

Measure the similarity between filters in the group A_k

These four filters have similar activation regions (i.e. these filters have **strong interactions**)

Increase the similarity between filters in the **same** group to ensure the consistency.


Wen Shen **Quanshi Zhang**

Method

- To satisfy the properties of consistency and diversity
  - use a set of filters to jointly represent a specific part/region, instead of using a single filter → **consistency**
  - use different sets of filters to represent different parts/regions → **diversity**

$$\mathrm{Loss}(\theta, \mathbf{A}) = -\sum_{k=1}^{K} \frac{S_k^{\mathrm{within}}}{S_k^{\mathrm{all}}} = -\sum_{k=1}^{K} \frac{\sum_{i,j\in A_k} s_{ij}}{\sum_{i\in A_k, j\in\Omega} \boxed{s_{ij}}}$$

Measure the similarity between filters in A_k and *all* filters (Ω) in the target layer.

Filters in A_k

Other filters

These four filters have different activation regions (i.e. these filters have **weak interactions**)


Wen Shen **Quanshi Zhang**

Method

- To satisfy the properties of consistency and diversity
  - use a set of filters to jointly represent a specific part/region, instead of using a single filter → **consistency**
  - use different sets of filters to represent different parts/regions → **diversity**

$$\mathrm{Loss}(\theta, \mathbf{A}) = -\sum_{k=1}^{K} \frac{S_k^{\text{within}}}{S_k^{\text{all}}} = -\sum_{k=1}^{K} \frac{\sum_{i,j \in A_k} s_{ij}}{\sum_{i \in A_k, j \in \Omega} \boxed{s_{ij}}}$$

Measure the similarity between filters in A_k and *all* filters (Ω) in the target layer.

Filters in A_k

Other filters

Decrease the similarity between filters in **different** groups to ensure the diversity.

These four filters have different activation regions (i.e. these filters have **weak interactions**)

Strongly interacted filters → meaningful concepts


Wen Shen


Quanshi Zhang

Method

- We add a loss to the target convolutional layer to construct an compositional interpretable layer, where filters satisfy the properties of consistency and diversity.

$$\mathrm{Loss}(\theta, \mathbf{A}) = -\sum_{k=1}^{K} \frac{S_k^{\mathrm{within}}}{S_k^{\mathrm{all}}} = -\sum_{k=1}^{K} \frac{\sum_{i,j\in A_k} s_{ij}}{\sum_{i\in A_k, j\in\Omega} s_{ij}}$$

- The similarity between filters *i*, *j* is implemented as a kernel function.

$$s_{ij} = \mathcal{K}(X_i, X_j) = \boxed{\rho_{ij}} + 1 = \frac{\mathrm{cov}(X_i, X_j)}{\sigma_i \sigma_j} + 1 \geq 0,$$

The Pearson's correlation coefficient between variables x_i^l and x_j^l through different images.


Wen Shen **Quanshi Zhang**

Method

- We add a loss to the target convolutional layer to construct an compositional interpretable layer, where filters satisfy the properties of consistency and diversity.

$$\text{Loss}(\theta, \mathbf{A}) = -\sum_{k=1}^{K} \frac{S_k^{\text{within}}}{S_k^{\text{all}}} = -\sum_{k=1}^{K} \frac{\sum_{i,j\in A_k} \boxed{s_{ij}}}{\sum_{i\in A_k, j\in\Omega} s_{ij}}$$

Measure the similarity between feature maps of filters *i*, *j*.

- The similarity between filters *i*, *j* is implemented as a kernel function.

$$s_{ij} = \mathcal{K}(X_i, X_j) = \boxed{\rho_{ij}} + 1 = \frac{\text{cov}(X_i, X_j)}{\sigma_i \sigma_j} + 1 \geq 0,$$

The Pearson's correlation coefficient between variables x_i^l and x_j^l through different images.


Wen Shen **Quanshi Zhang**

Learning of the filter loss

The minimization of $\text{Loss}(\theta, \mathbf{A})$ is essentially equivalent to the problem of the spectral clustering[2].

$$\frac{1}{2}(\text{Loss}(\theta, \mathbf{A}) + K) = \frac{1}{2}\sum_{k=1}^{K} \frac{\sum_{i \in A_k, j \notin A_k} s_{ij}}{\sum_{i \in A_k, j \in \Omega} s_{ij}}$$


Wen Shen


Quanshi Zhang

Broad applicability

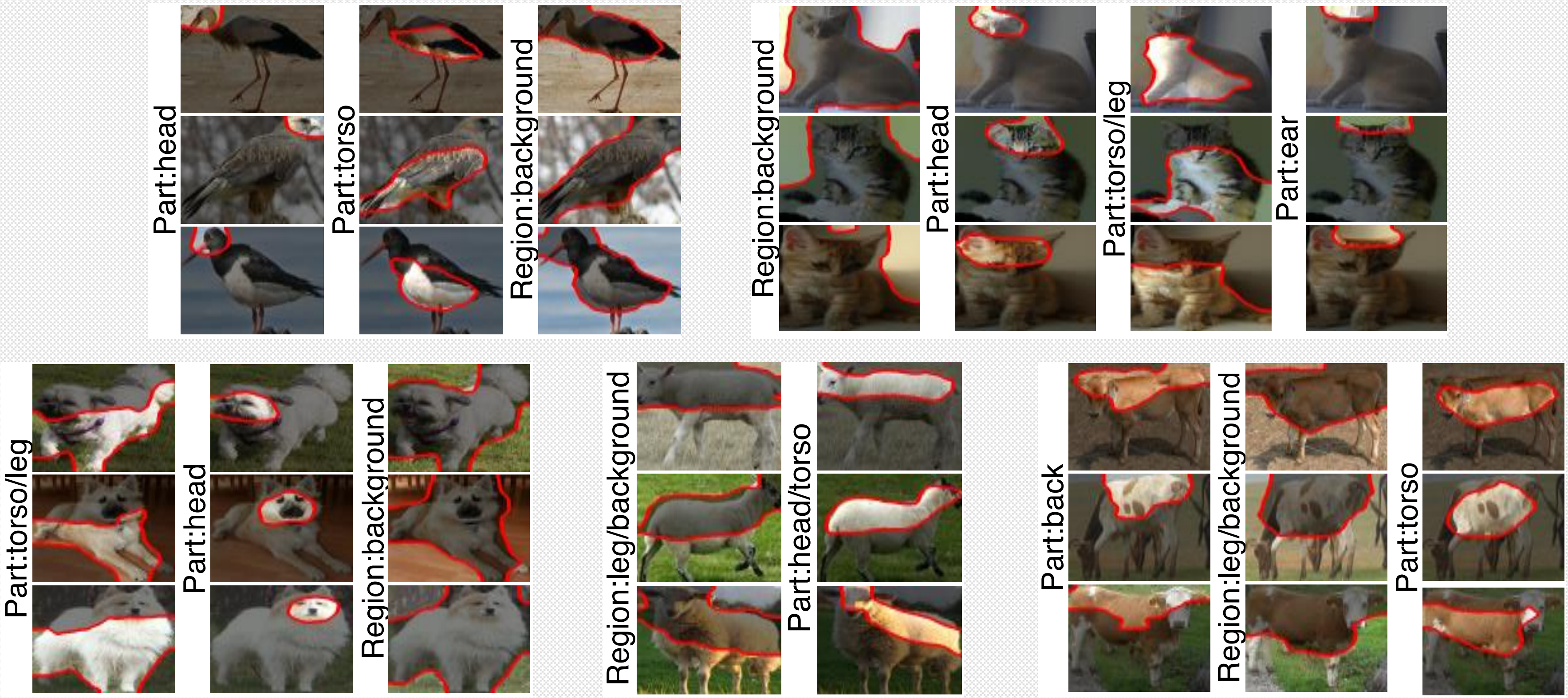
- Can be applied to different task
  - e.g. object classification, segmentation, etc.

- Tested on different CNNs
  - VGG-13
  - VGG-16
  - ResNet-18
  - ResNet-50
  - DenseNet-121
  - DenseNet-161


Wen Shen **Quanshi Zhang**

Activation regions of interpretable filters

Binary classification of a single category.

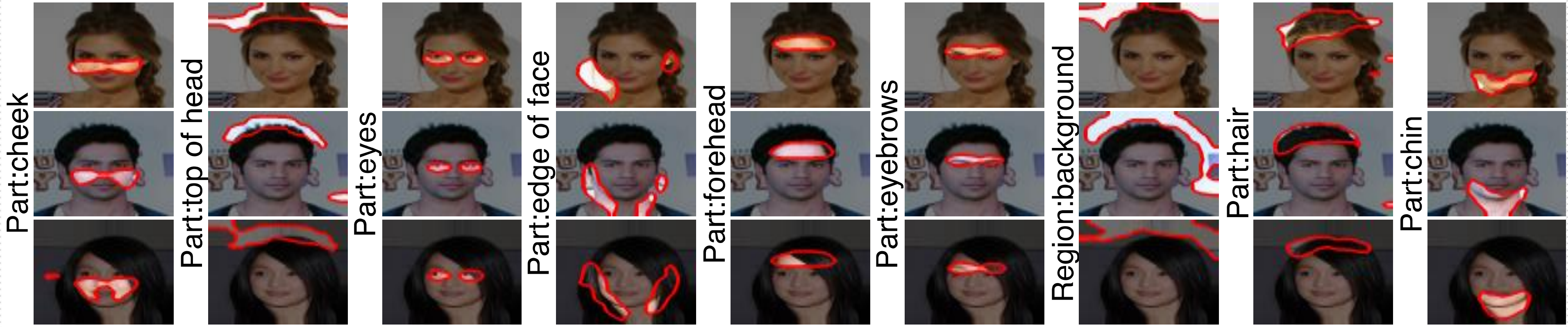


Each filter in a compositional CNN consistently represented the same object part or the same image region, while different filters represented different parts and regions.

Activation regions of interpretable filters


Wen Shen **Quanshi Zhang**

Multi-label classification.

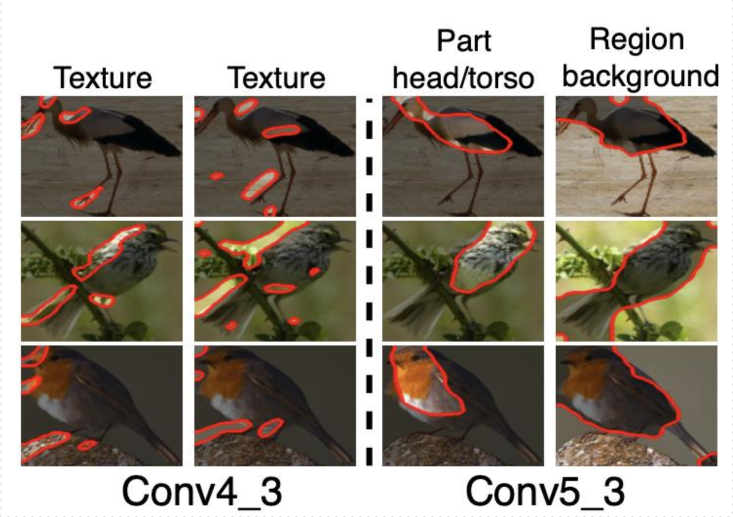


Interpretable filters in a ICNN encoded **very few** types of patterns, which are concentrated in the center of the face.

Interpretable filters in a compositional CNN encoded **diverse** patterns, covering almost all elements of the face image, such as forehead, eyes, nose, etc.

Wen Shen **Quanshi Zhang**

More visualization



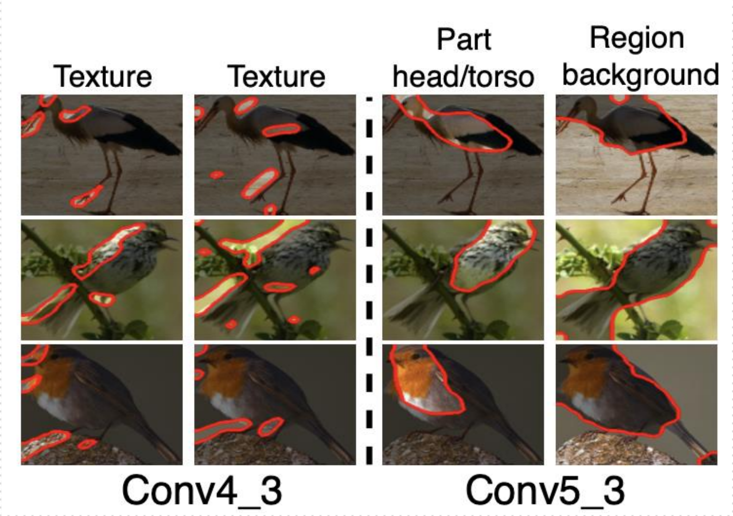
Comparison of interpretable filters of a high convolutional layer and a middle convolutional layer.

High convolutional layer: Interpretable filters usually represent object parts or image regions;

Low convolutional layer: Interpretable filters usually represent local textures or shapes.


Wen Shen **Quanshi Zhang**

More visualization

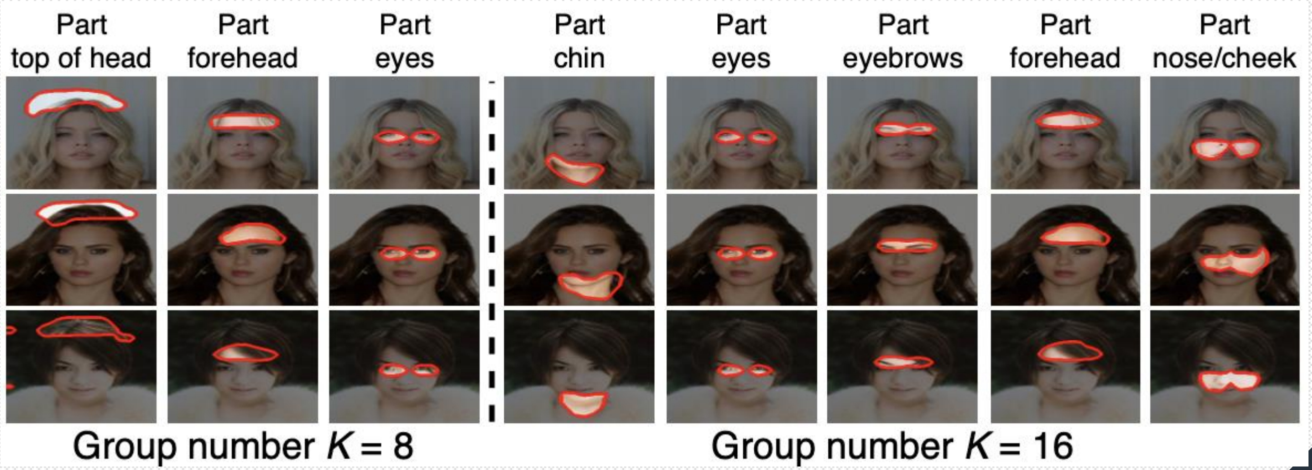


Comparison of interpretable filters of a high convolutional layer and a middle convolutional layer.

High convolutional layer: Interpretable filters usually represent object parts or image regions; **Low convolutional layer:** Interpretable filters usually represent local textures or shapes.

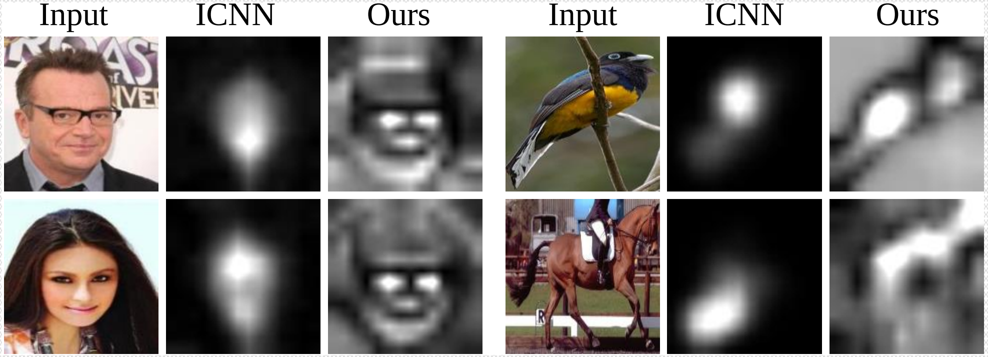
Comparison of interpretable filters learned with different values of

As group number increases, more detailed visual patterns are learned.




Wen Shen **Quanshi Zhang**

More visualization

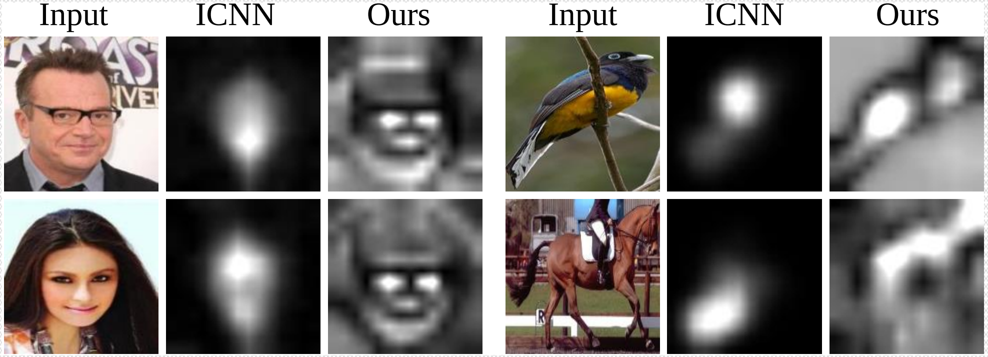


Visualizing distributions of visual patterns that are encoded in interpretable filters.

Interpretable filters of a compositional CNN explain much more regions in an image than those of an ICNN.

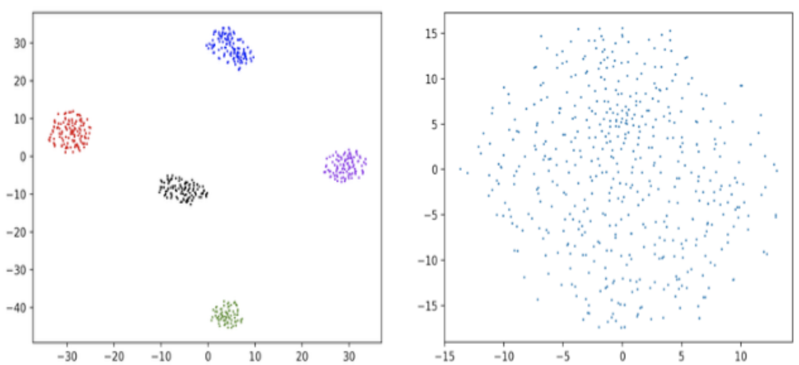
Wen Shen **Quanshi Zhang**

More visualization



Visualizing distributions of visual patterns that are encoded in interpretable filters.

Interpretable filters of a compositional CNN explain much more regions in an image than those of an ICNN.



(c1) compositional CNN (c2) traditional CNN

Visualizing filters in a compositional CNN and a traditional CNN using t-SNE

Feature maps of a compositional CNN seem more clustered than those of a traditional CNN.

Quantitative Evaluation of Filter Interpretability


Wen Shen **Quanshi Zhang**

- **Inconsistency of Visual Patterns** measures the consistency of visual patterns represented by a filter through different images.
  - Ideally, an interpretable filter was supposed to have high consistency.

$$H = -\sum_{j=1}^{T} P_j \log P_j$$

The entropy of such probabilities over different semantic concepts.

$$P_j = \frac{\sum_{I \in \mathbf{I}^{\text{test}}} \sum_{u=1}^{M} \min\{\tilde{Q}_u(I), G_u^j(I)\}}{\sum_{I \in \mathbf{I}^{\text{test}}} \sum_{u=1}^{M} \tilde{Q}_u(I)}$$

The probability of a filter being associated with a ground-truth semantic concept in a specific image.

Quantitative Evaluation of Filter Interpretability


Wen Shen


Quanshi Zhang

- **Diversity of Visual Patterns** evaluates whether a CNN learned various visual patterns.

$$Diversity = \frac{1}{M}\mathbb{E}_I\left[\sum_{u=1}^{M} \mathbb{1}\left(\left(\frac{1}{d}\sum_{i=1}^{d} \tilde{Q}_u^i(I)\right) \geq \gamma\right)\right]$$

A pixel is explained by a CNN, if this pixel was explained by some filters.

The diversity of visual patterns was approximately quantified as the number of pixels which had been explained by a CNN.

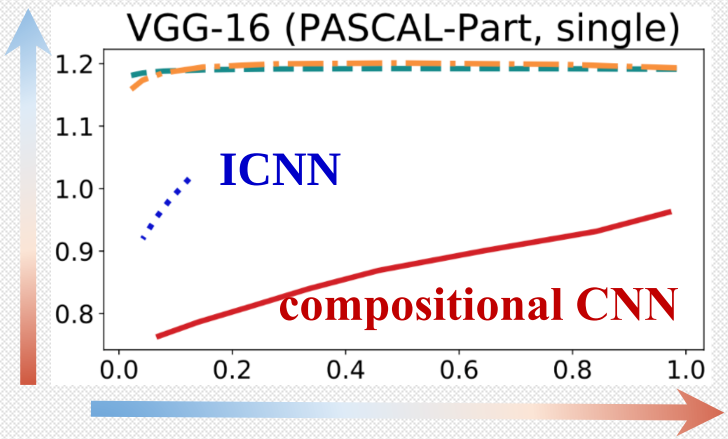
Strongly interacted filters → meaningful concepts

Our method learns filters with much higher consistency and diversity

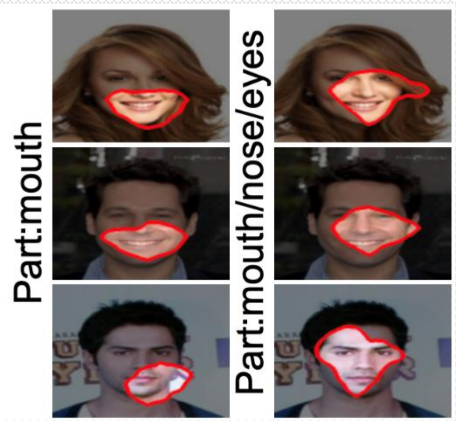
Wen Shen **Quanshi Zhang**

Low consistency

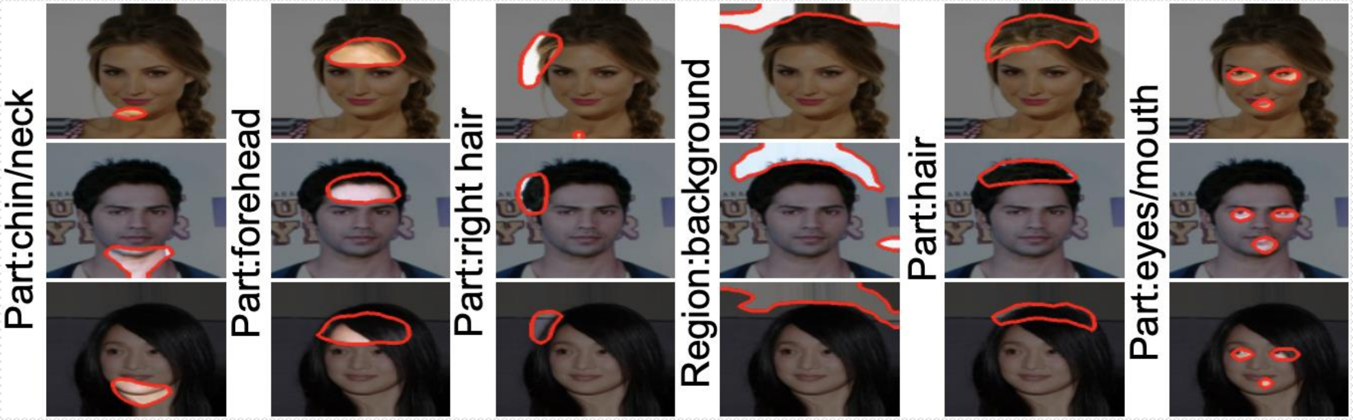
High consistency



Low diversity High diversity



Activation regions of interpretable filters in ICNN

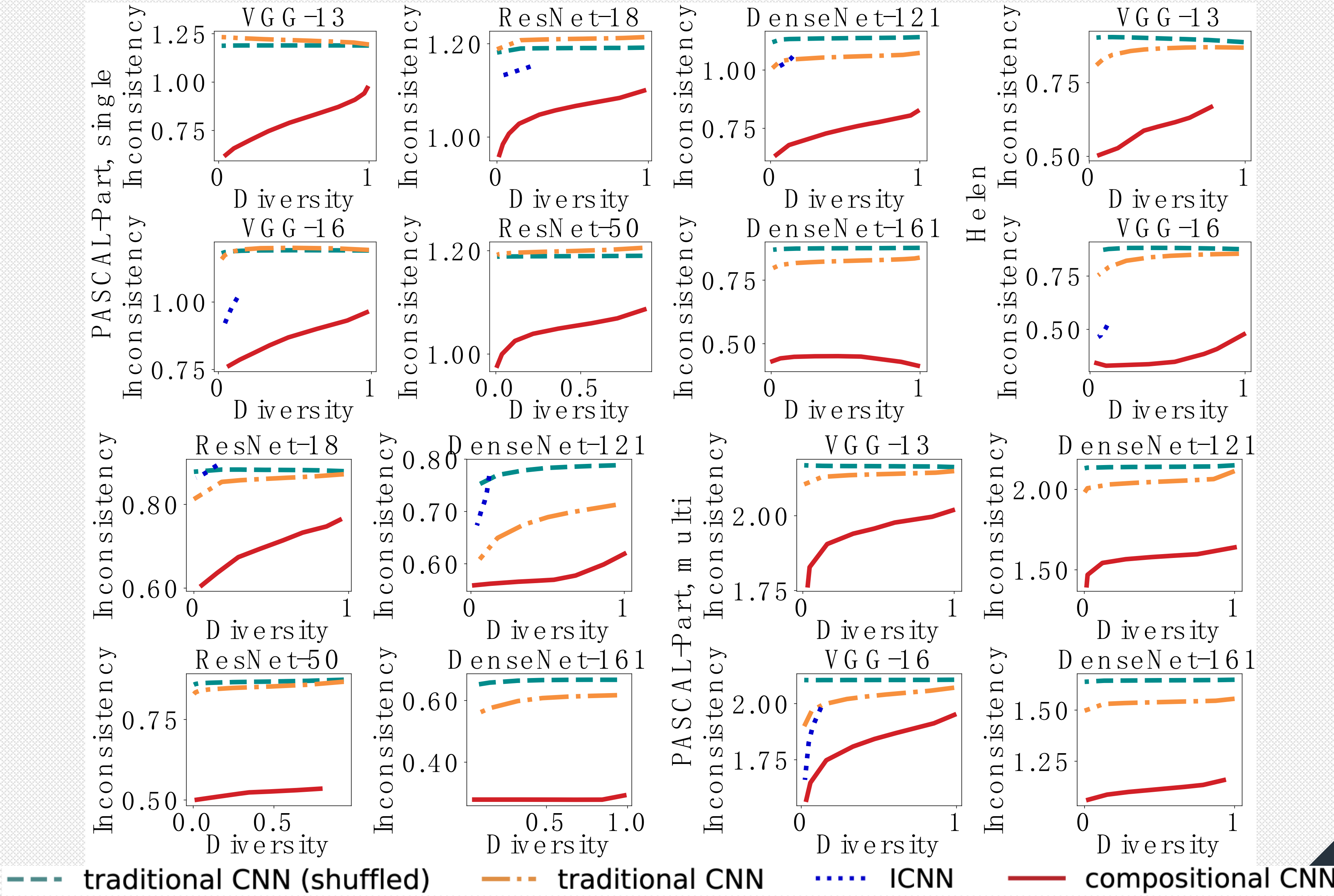


Activation regions of interpretable filters in compositional CNN

Strongly interacted filters → meaningful concepts

Our method learns filters with much higher consistency and diversity


Wen Shen **Quanshi Zhang**



Wen Shen **Quanshi Zhang**

Classification performance

	single-category			multi-category
	PASCAL-Part	CUB200	CelebA	PASCAL-Part
VGG-13	**97.07**	**99.76**	–	**87.51**
compositional CNN	96.29	99.41	–	86.37
VGG-16	**98.66**	**99.86**	90.47	89.71
ICNN	95.39	96.51	89.11	**91.60**
compositional CNN	97.12	99.27	**90.70**	87.51
ResNet-18	**97.77**	**99.81**	89.60	–
ICNN	93.30	97.12	–	–
compositional CNN	96.90	98.49	**89.76**	–
ResNet-50	**97.88**	**99.88**	**90.21**	–
compositional CNN	97.30	99.27	89.63	–
DenseNet-121	**98.29**	**99.92**	–	91.28
ICNN	96.55	99.22	–	–
compositional CNN	97.52	98.83	–	**91.75**
DenseNet-161	**98.70**	**99.96**	–	**93.48**
compositional CNN	98.14	99.61	–	92.66

Compositional CNNs exhibite comparable classification performance with traditional CNNs. Besides, compositional CNNs achieve higher accuracy than ICNNs in most comparisons.