Interpretable Convolutional Neural Networks

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu
University of California, Los Angeles

CVPR 2018

Outline

- Background
- Motivation & Goal
- Analysis
- Proposed Method
- Experiments
- Conclusion

Background

Motivation & Goal

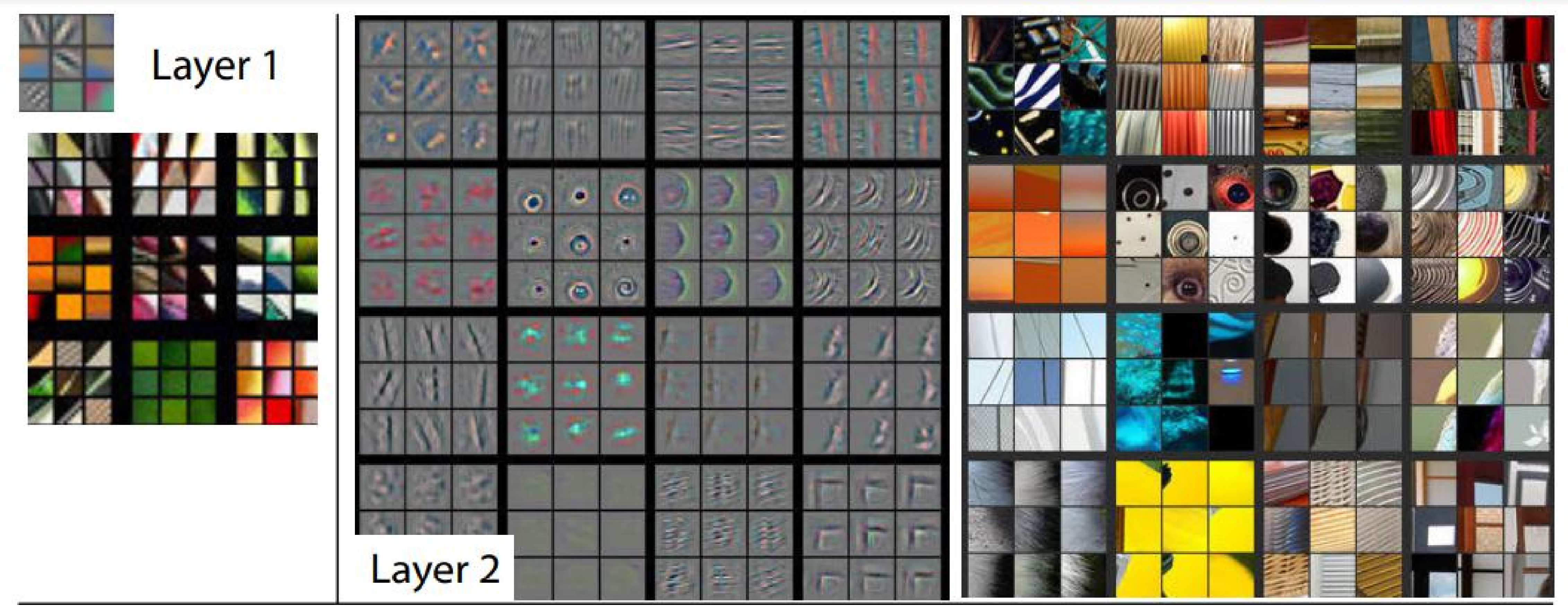
Motivation：

Without any additional human supervision, can we modify a CNN to make its conv-layers obtain interpretable knowledge representations?

Goal：

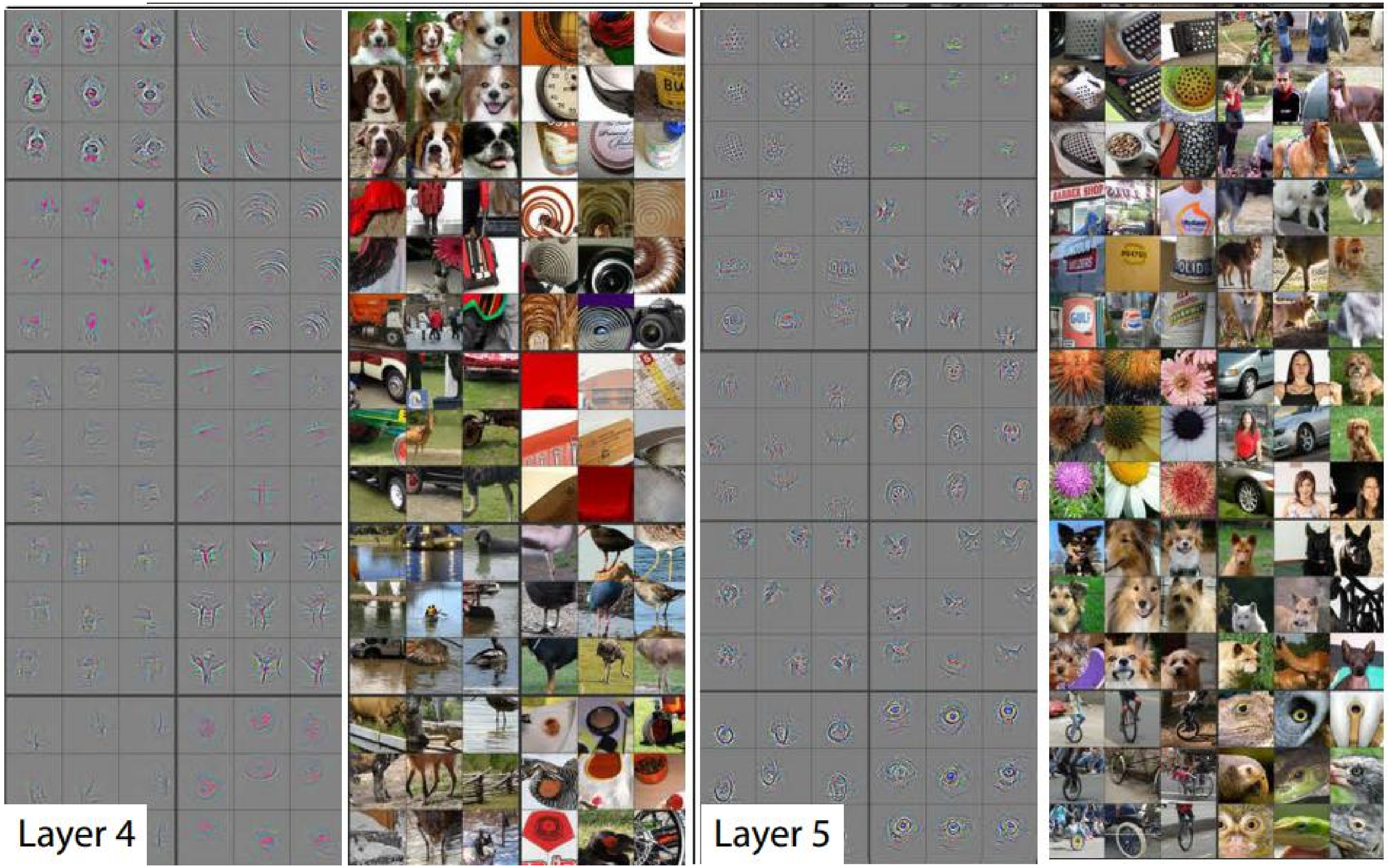
1. We slightly revise a CNN to improve its interpretability, which can be broadly applied to CNNs with different structures.
2. We learn interpretable filters for a CNN without any additional annotations of object parts or textures for supervision. Training samples are the same as the original CNN.
3. We do not hope the network interpretability greatly affects the discrimination power.

Analysis



Filters in low conv-layers usually describe simple textures.

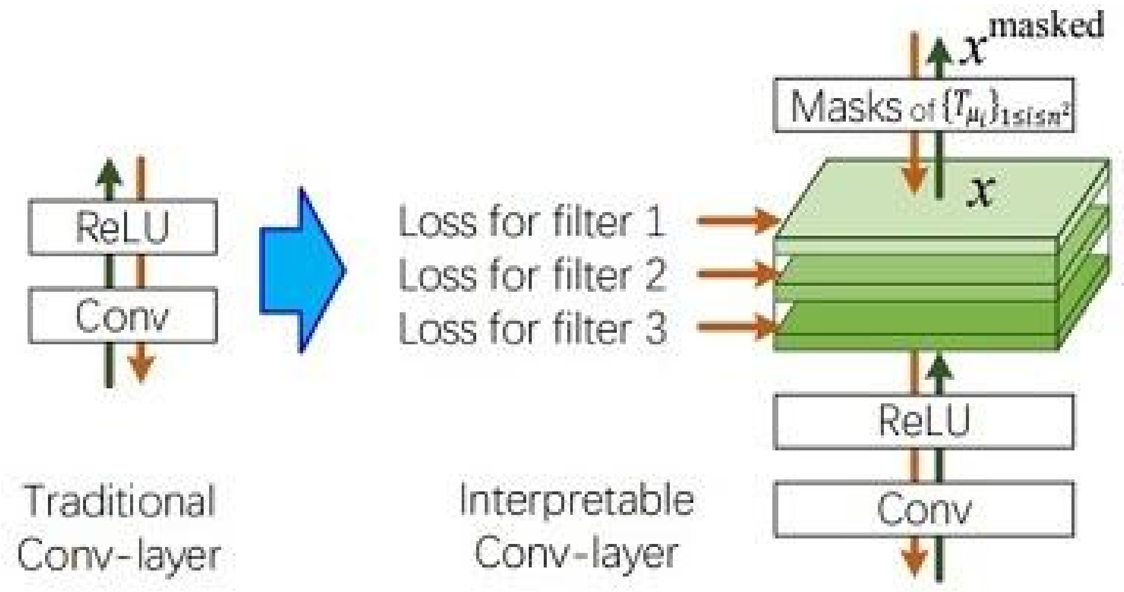
Analysis



Filters in high conv-layers are more likely to represent object parts.

Proposed Method

This paper proposes a simple yet effective loss to push a filter in a specific conv-layer of a CNN towards the representation of an object part.



Proposed Method

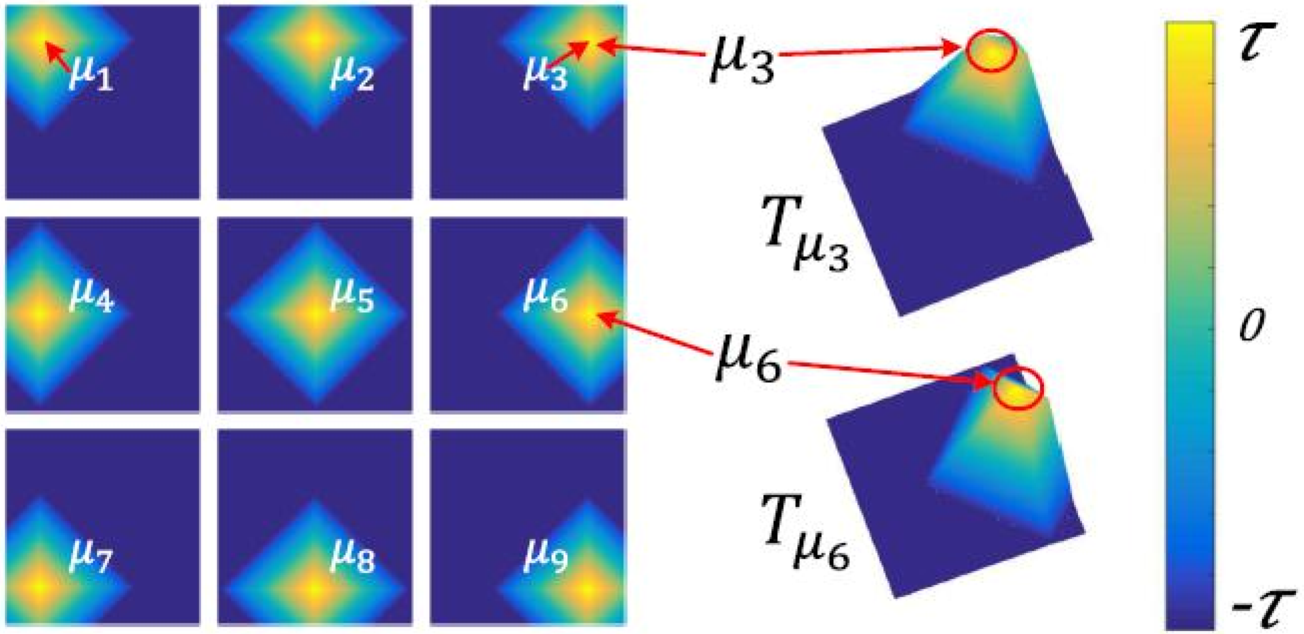


Figure 3. Templates of T_{μ_i}. Each template T_{μ_i} matches to a feature map x when the target part mainly triggers the i-th unit in x. In fact, the algorithm also supports a round template based on the L-2 norm distance. Here, we use the L-1 norm distance instead to speed up the computation.

Proposed Method

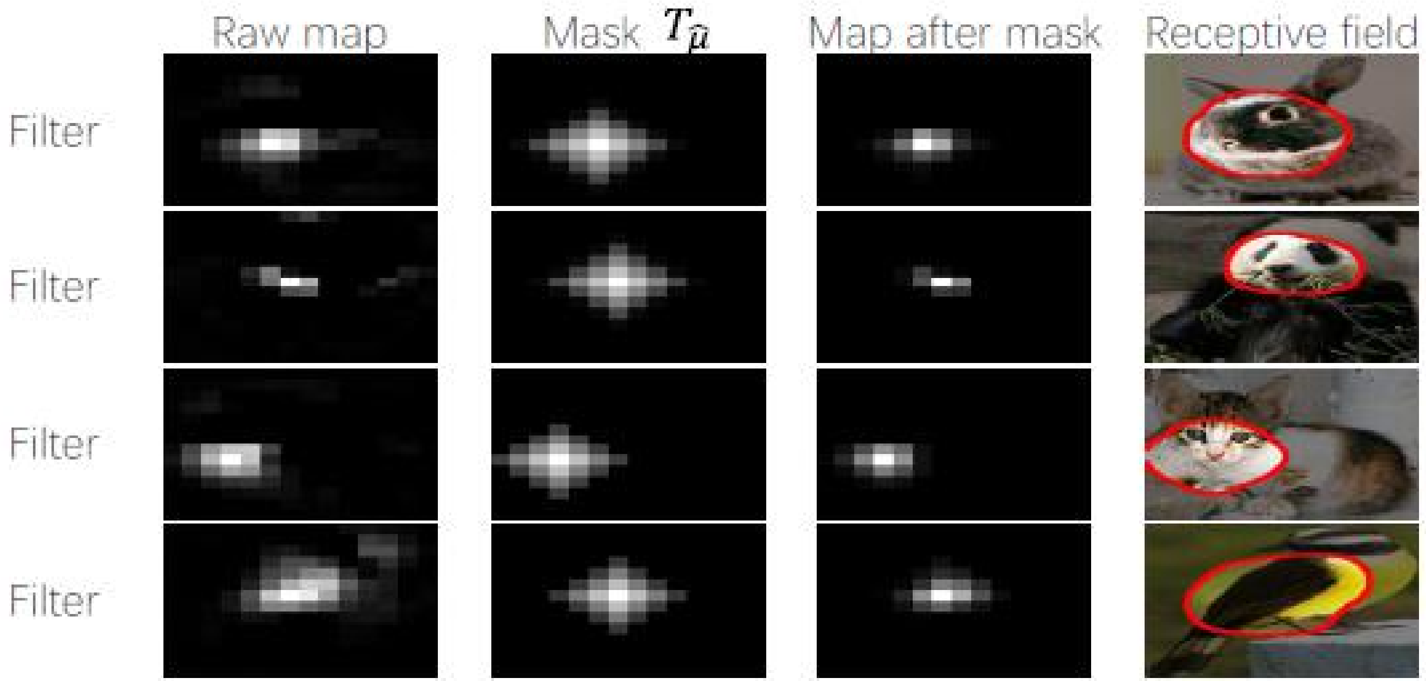


Figure 4. Given an input image I, from the left to the right, we consequently show the feature map of a filter after the ReLU layer x, the assigned mask $T_{\hat{\mu}}$, the masked feature map x^{masked}, and the image-resolution RF of activations in x^{masked} computed by [40].

Proposed Method

Filter Loss Function:

$$\mathbf{Loss}_f = -MI(\mathbf{X};\mathbf{T}) = -\sum_T p(T)\sum_x p(x|T)\log\frac{p(x|T)}{p(x)}$$

$MI(\cdot)$ denotes the mutual information

$$\forall T \in \mathbf{T}, \qquad p(x|T) = \frac{1}{Z_T}\exp\left[tr(x \cdot T)\right]$$

$$Z_T = \sum_{x\in\mathbf{X}} \exp(tr(x \cdot T))$$

$$tr(x \cdot T) = \sum_{ij} x_{ij}t_{ij}.\ p(x) = \sum_T p(T)p(x|T)$$

$$p(x) = \sum_T p(T)p(x|T)$$

注：最大化mutual information，因此loss function前面有负号
Mutual Information https://www.wikiwand.com/en/Mutual_information

Experiments

Average part interpretability

	bird	cat	cow	dog	horse	sheep	Avg.
AlexNet	0.332	0.363	0.340	0.374	0.308	0.373	0.348
AlexNet, interpretable	**0.770**	**0.565**	**0.618**	**0.571**	**0.729**	**0.669**	**0.654**
VGG-16	0.519	0.458	0.479	0.534	0.440	0.542	0.495
VGG-16, interpretable	**0.818**	**0.653**	**0.683**	**0.900**	**0.795**	**0.772**	**0.770**
VGG-M	0.357	0.365	0.347	0.368	0.331	0.373	0.357
VGG-M, interpretable	**0.821**	**0.632**	**0.634**	**0.669**	**0.736**	**0.756**	**0.708**
VGG-S	0.251	0.269	0.235	0.275	0.223	**0.287**	0.257
VGG-S, interpretable	**0.526**	**0.366**	**0.291**	**0.432**	**0.478**	0.251	**0.390**

Table 1. Part interpretability of filters in CNNs for single-category classification based on the Pascal VOC Part dataset [3].

Network	Logistic log loss[4]	Softmax log loss
VGG-16	0.710	0.723
VGG-16, interpretable	**0.938**	**0.897**
VGG-M	0.478	0.502
VGG-M, interpretable	**0.770**	**0.734**
VGG-S	0.479	0.435
VGG-S, interpretable	**0.572**	**0.601**

Table 2. Part interpretability of filters in CNNs that are trained for multi-category classification based on the VOC Part dataset [3]. Filters in our interpretable CNNs exhibited significantly better part interpretability than ordinary CNNs in all comparisons.

Experiments

Location instability

	bird	cat	cow	dog	horse	sheep	Avg.
AlexNet	0.153	0.131	0.141	0.128	0.145	0.140	0.140
AlexNet, interpretable	**0.090**	**0.089**	**0.090**	**0.088**	**0.087**	**0.088**	**0.088**
VGG-16	0.145	0.133	0.146	0.127	0.143	0.143	0.139
VGG-16, interpretable	**0.101**	**0.098**	**0.105**	**0.074**	**0.097**	**0.100**	**0.096**
VGG-M	0.152	0.132	0.143	0.130	0.145	0.141	0.141
VGG-M, interpretable	**0.086**	**0.094**	**0.090**	**0.087**	**0.084**	**0.084**	**0.088**
VGG-S	0.152	0.131	0.141	0.128	0.144	0.141	0.139
VGG-S, interpretable	**0.089**	**0.092**	**0.092**	**0.087**	**0.086**	**0.088**	**0.089**

Table 4. Location instability of filters ($\mathbf{E}_{f,k}[D_{f,k}]$) in CNNs that are trained for single-category classification using the Pascal VOC Part dataset [3]. Filters in our interpretable CNNs exhibited significantly lower localization instability than ordinary CNNs.

Network	Avg. location instability
AlexNet	0.150
AlexNet, interpretable	**0.070**
VGG-16	0.137
VGG-16, interpretable	**0.076**
VGG-M	0.148
VGG-M, interpretable	**0.065**
VGG-S	0.148
VGG-S, interpretable	**0.073**

Table 5. Location instability of filters ($\mathbf{E}_{f,k}[D_{f,k}]$) in CNNs for single-category classification using the CUB200-2011 dataset.

Experiments

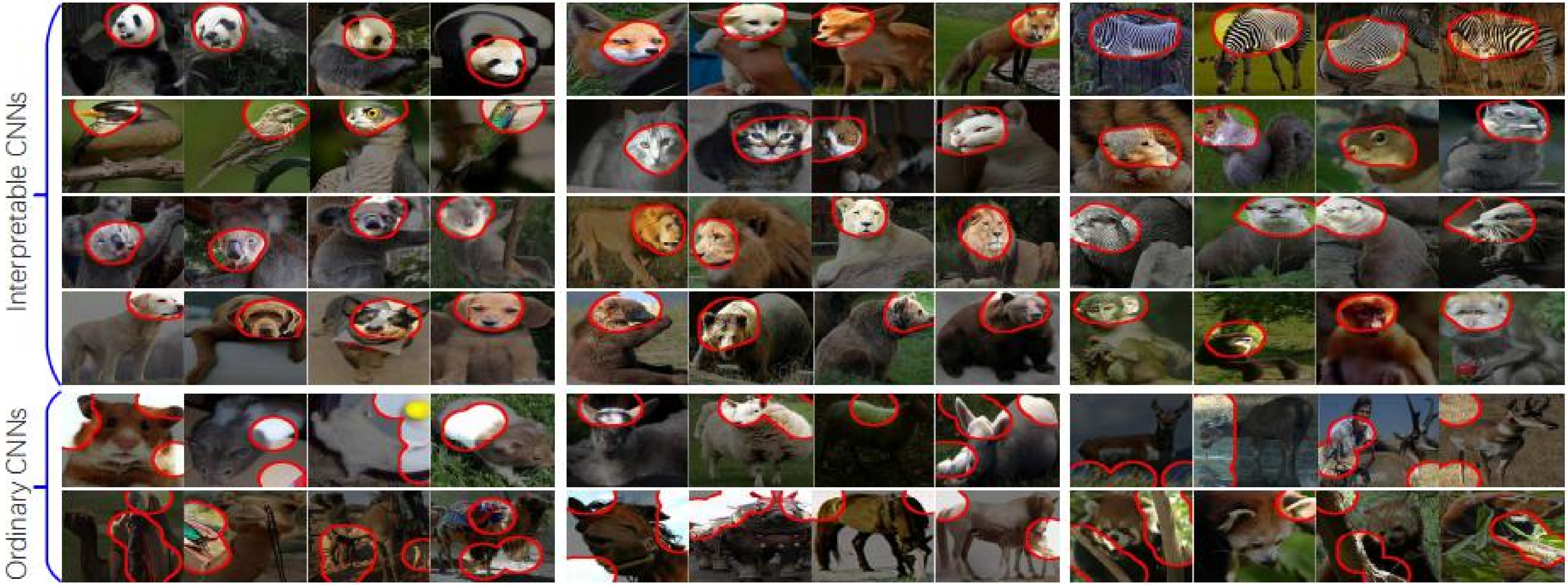
Classification Accuracy

	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

Table 7. Classification accuracy based on different datasets. In single-category classification, ordinary CNNs performed better, while in multi-category classification, interpretable CNNs exhibited superior performance.

Experiments

Performances



Conclusion

1. This paper proposes a general method to enhance feature interpretability of CNNs.

2. We design a loss to push a filter in high conv-layers towards the representation of an object part during the learning process without any part annotations.