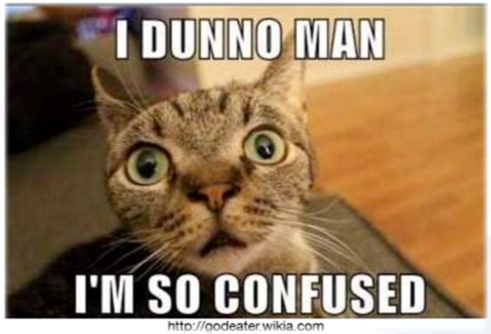
Explainable CNN

报告人：姚柯璐

2019-09-27

CNN Performance

- Deep learning 超强的表现终结了一批旧算法
- Deep learning 简化了算法设计的复杂度
- But
  - 端对端的训练一个black-box model会一直平稳的向下发展吗?
  - 随着网络结构和loss function的设计越来越复杂，神经网络真的会按照设计老老实实的去表达人们希望它表达的知识吗?

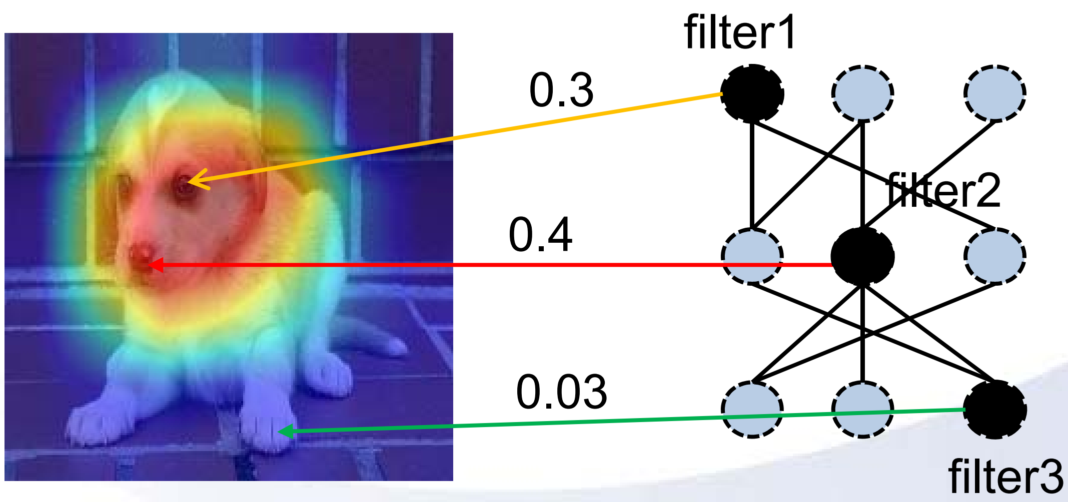


NN的可解释性研究层面

- 定量的解释一定比例的神经网络内部的推理逻辑
  - 拆分-哪些解释，哪些建模，哪些猜
- 在语义层面上建立认知与神经网络表达的信任关系
  - 人与人交流未必是完全的理解，而是依靠信任关系
  - NN可解释性保证一定置信度下的大致信任
- 依靠基础工具对越来越复杂的模型进行解读

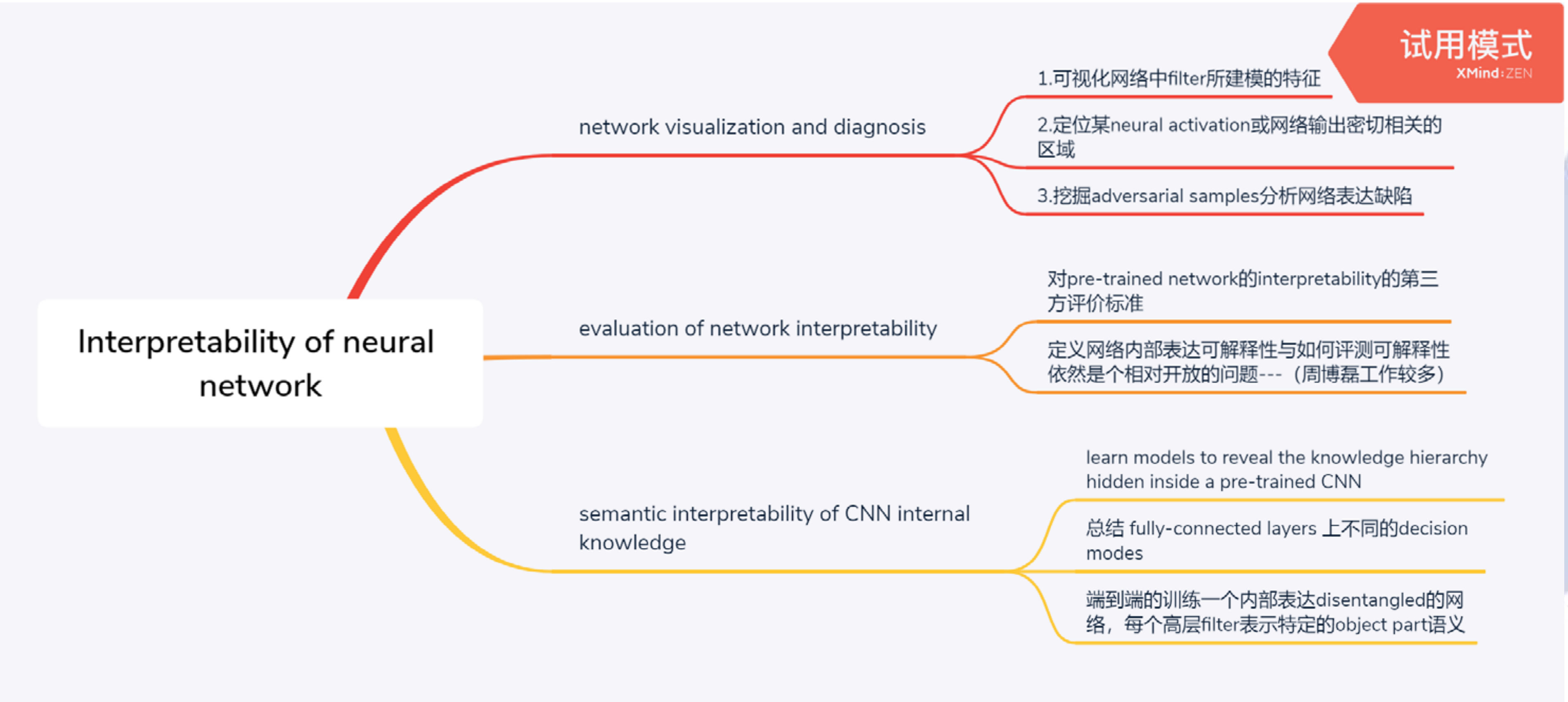
Explainable CNN

- Visualization of CNN Knowledge —— 可视化每个unit 知识表达
- 定义标准评测CNN知识的interpretability
- 提出让神经网络具有清晰的符号化的内部知识表达，在语义层对NN进行诊断、修改



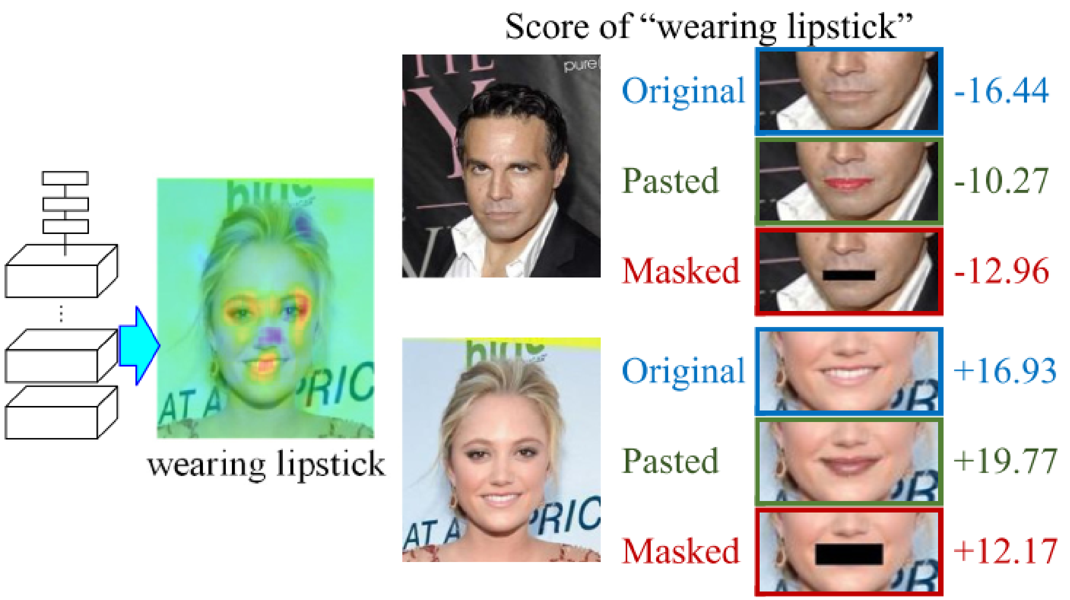
- End-to-End
- End-to-Middle?
- Middle-Middle?
- Debug CNN?
- Big Data?

network interpretability 研究方向和内容



Biased representations in a CNN

- a high accuracy on testing images cannot always ensure that a CNN learns correct representations
- The CNN may use unreliable co-appearing contexts to make predictions



- heat maps of inference patterns of the lipstick attribute
- The CNN mistakenly considers unrelated patterns as contexts to infer the lipstick

 Examining CNN Representations with respect to Dataset Bias (AAAI2018)

Interpreting CNN Knowledge via an Explanatory Graph (AAAI 2018)

- 所述领域：semantic interpretability of CNN
- 研究背景：black-box的表达难以避免representation bias 等问题，却保证了特征提取的灵活性与效率，而传统的图模型具有清晰的语义结构，却没有NN的效果
- 研究目的：“探索一种white-box的表达方式，同时又具有神经网络表达的flexibility和信息效率”

Interpreting CNN Knowledge via an Explanatory Graph (AAAI 2018)

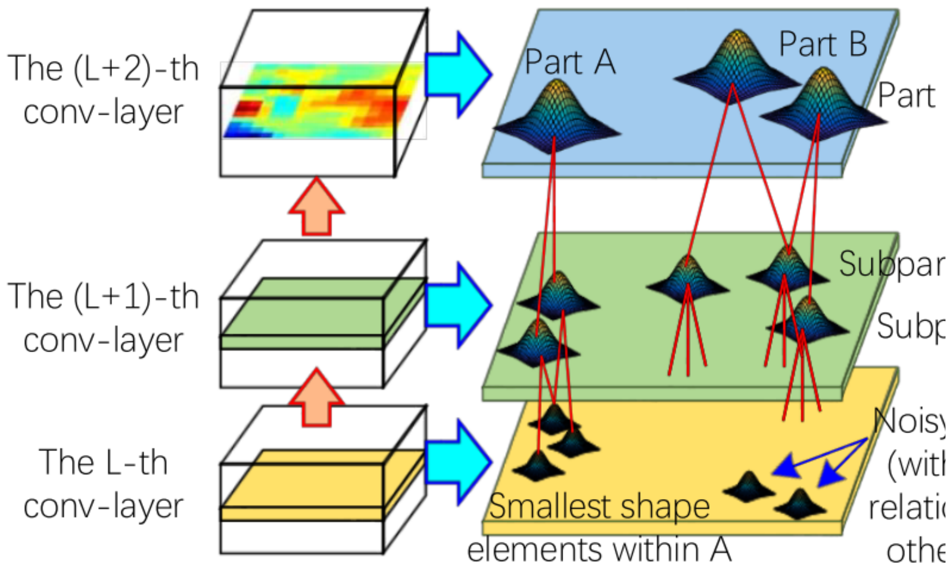
- 解决问题：CNN中层 Conv-layer 混乱的知识表达
- 提出方法：learn a explanatory graph.
  - Each filter represents each node represents a part pattern
  - each edge encodes co-activation relationships and spatial relationships between patterns

算法设计

$$P(\mathbf{X}_L|\mathbf{R}_{L+1},\boldsymbol{\theta}_L)=\prod_{x\in\mathbf{X}_L}P(\mathbf{p}_x|\mathbf{R}_{L+1},\boldsymbol{\theta}_L)^{F(x)}$$

$$=\prod_{x\in\mathbf{X}_L}\Big\{\sum_{V\in\Omega_{L,d}\cup\{V_{\text{none}}\}}P(V)P(\mathbf{p}_x|V,\mathbf{R}_{L+1},\boldsymbol{\theta}_L)\Big\}_{d=d_x}^{F(x)}$$

$$\text{argmax}_{\boldsymbol{\theta}_L}\prod_{I\in\mathbf{I}}P(\mathbf{X}_L^I|\mathbf{R}_{L+1}^I,\boldsymbol{\theta}_L)$$



Inputs: feature map $\mathbf{X}_L$ of the L-th conv-layer, inference results $\mathbf{R}_{L+1}$ in the upper conv-layer.
Outputs: μ_V, E_V for $\forall V\in\Omega_L$.
Initialization: $\forall V$, $E_V=\{V_{\text{dummy}}\}$, a random value for $\mu_V^{(0)}$
for $iter=1$ *to* T **do**
  $\forall V\in\Omega_L$, compute $P(\mathbf{p}_x,V|\mathbf{R}_{L+1},\boldsymbol{\theta}_L)$.
  for $V\in\Omega_L$ **do**
    1) Update μ_V via an EM algorithm,
    $\mu_V^{(iter)}=\mu_V^{(iter-1)}+\eta\sum_{I\in\mathbf{I},x\in\mathbf{X}_L}\mathbf{E}_{P(V|\mathbf{p}_x,\mathbf{R}_{L+1},\boldsymbol{\theta}_L)}\Big[F(x)\cdot\frac{\partial\log P(\mathbf{p}_x,V|\mathbf{R}_{L+1},\boldsymbol{\theta}_L)}{\partial\mu_V}\Big].$
    2) Select M patterns from $V'\in\Omega_{L+1}$ to construct E_V based on a greedy strategy, which maximize $\prod_{I\in\mathbf{I}}P(\mathbf{X}_L|\mathbf{R}_{L+1},\boldsymbol{\theta}_L)$.
  end
end

Algorithm 1: Learning sub-graph in the L-th layer

实验设计

- Four CNNs：VGG-16,ResNet50,152, VAE-GAN
- Three experiments to evaluate the explanatory graph
  - 1. visualize patterns in the graph
  - 2. evaluate the semantic interpretability of the part patterns
  - 3. multi-shot learning for part localization, in order to test the transferability of patterns in the graph
- Three benchmark datasets:
  - a total of 37 animal categories in three datasets: the ILSVRC 2013 DET Animal-Part dataset, the CUB200-2011 dataset, and the Pascal VOC Part dataset

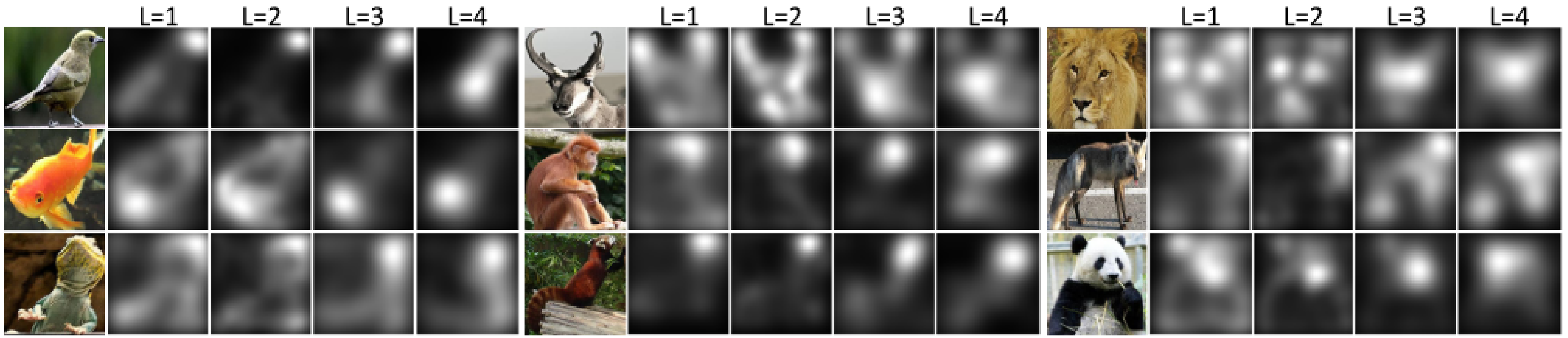
1. visualize patterns in the graph

- Given an explanatory graph for a VGG-16 network(trained/fine-tuned using object images of a category)
- visualizing part patterns in the graph in three ways
- 1.1 Top-ranked patches
  - Extract an images patch in the position of image plane with a scale of 70 pixels *70 pixels to represent pattern V

Figure 5: Image patches corresponding to different nodes in the explanatory graph.

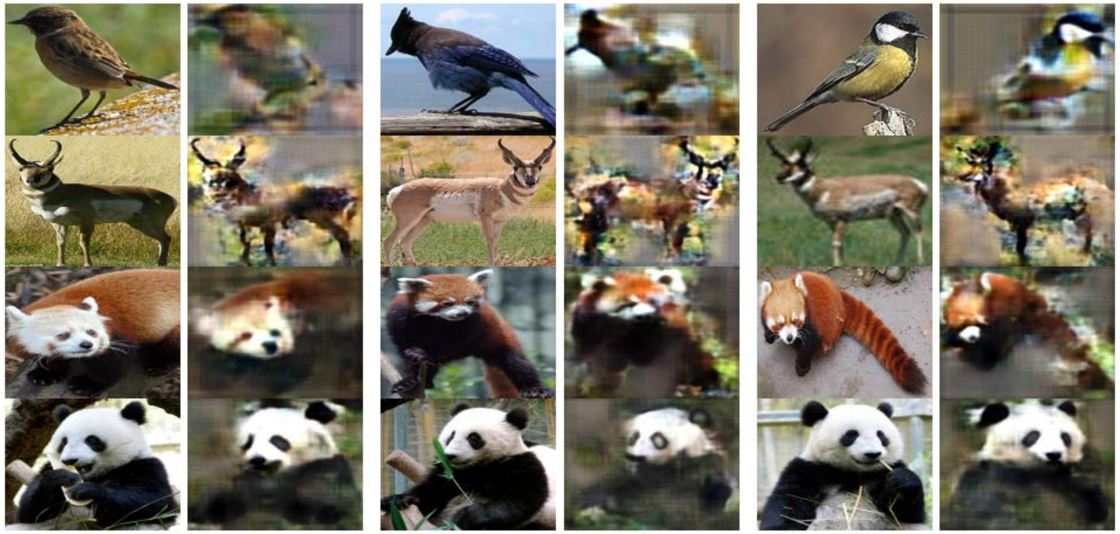
1. visualize patterns in the graph

- 1.2 Heat maps of patterns
- Given a cropped object image I, we used the explanatory graph to infer its patterns on image I, and drew heat maps to show the spatial distribution of the inferred patterns



1. visualize patterns in the graph

- 1.3 Pattern-based image synthesis
  - Given an object image I, we used the explanatory graph for pattern inference
  - Let the top-10% patterns with highest scores of $S^{I}_{V \to x}$ as valid ones.
  - We filtered out all neural responses of units, which were not assigned to valid patterns, from feature maps (setting these responses to zero)
  - then used (Dosovitskiy and Brox 2016) to synthesize the appearance corresponding to the modified feature maps



2. semantic interpretability of patterns

- This paper tests whether each pattern in an explanatory graph consistently represented the same object region among different images
  - four explanatory graphs for a VGG-16 network, two residual networks, and a VAE-GAN that were trained/fine-tuned using the CUB2002011 dataset
  - two methods to evaluate the semantic interpretability of patterns
    - Part interpretability of patterns
    - Location instability of inference positions

2. semantic interpretability of patterns

- 2.1 Part interpretability of patterns
  - Extracted patterns from high conv-layers, and as discussed in (Bau et al. 2017), high conv-layers contain large-scale part patterns
  - Used people to manually evaluate the pattern’s interpretability
  - how many inference results among the top K described the same object part, in order to compute the purity of part semantics of pattern V
  - graph nodes encod ns than raw filters

	VGG 16
Raw filter map	19.
Raw filter peak	43.
Ours	**95**

Purity o

2. semantic interpretability of patterns

- 2.2 Location instability of inference positions
  - We assumed that if a pattern was always triggered by the same object part through different images, then the distance between the pattern’s inference position and a ground-truth landmark of the object part should not change a lot among various images.

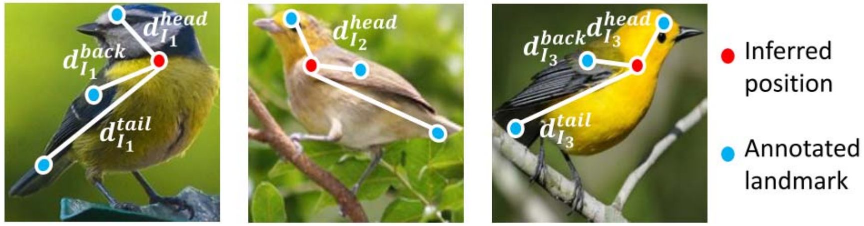


Figure 9: Notation for the computation of location instability.

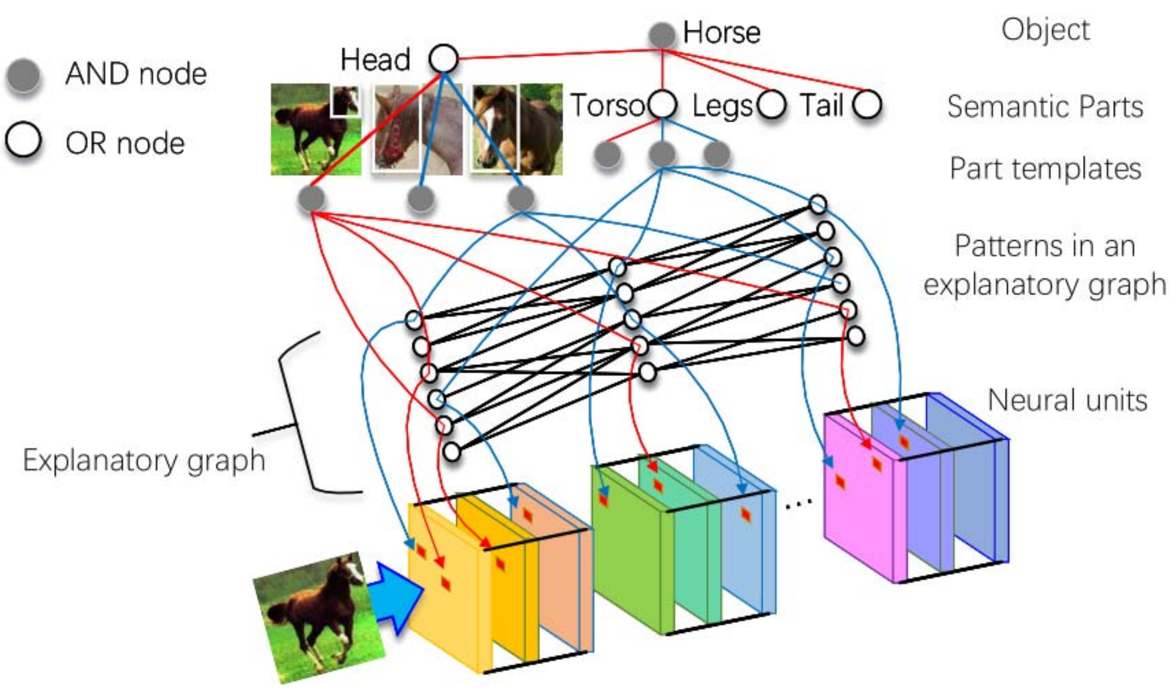
$$(\sqrt{var(d_I^{\text{head}})} + \sqrt{var(d_I^{\text{back}})} + \sqrt{var(d_I^{\text{tail}})})/3$$

	ResNet-50	ResNet-152	VGG-16	VAE-GAN
Raw filter (Zhou et al. 2015)	0.1328	0.1346	0.1398	0.1944
Ours	**0.0848**	**0.0858**	**0.0638**	**0.1066**
(Singh, Gupta, and Efros 2012)	0.1341			
(Simon, Rodner, and Denzler 2014)	0.2291			

Table 1: Location instability of patterns.

3. multi-shot part localization

- And-Or graph for semantic parts



总结

- proposed a simple yet effective method to learn an explanatory graph that reveals knowledge hierarchy inside conv-layers of a pre-trained CNN
- Experiments showed that patterns had significantly higher stability than baselines
- Partlocalization experiments well demonstrated the good transferability

请各位老师批评指正

谢谢！