Visual Interpretability for Deep Learning: a Survey

Mohammad Samavatian

CSE-5559

What is Paper About?

What?

- Understanding neural-network representations.
- Learning neural networks with interpretable/disentangled middle layer representations.

Why?

Interpretability is always the Achilles' heel of deep neural networks

- DNN has high discrimination power

But

- Low interpretability

Paper Agenda

- Visualization of CNN representations in intermediate network layers
- Diagnosis of CNN representations
- Disentanglement of “the mixture of patterns” encoded in each filter of CNNs
- Building explainable models
- Semantic-level middle-to-end learning via human computer interaction

Paper Agenda

- **Visualization of CNN representations in intermediate network layers**
- Diagnosis of CNN representations
- Disentanglement of “the mixture of patterns” encoded in each filter of CNNs
- Building explainable models
- Semantic-level middle-to-end learning via human computer interaction

Visualization of CNN Representations

Exploring visual patterns hidden inside a neural unit

- Gradient-based methods:
  - use the gradients to estimate the image appearance that maximizes the unit score
- Up-convolutional net:
  - inverts CNN feature maps to images
- Compute the image-resolution receptive field of neural activations in a feature map

Paper Agenda

- Visualization of CNN representations in intermediate network layers
- **Diagnosis of CNN representations**
- Disentanglement of “the mixture of patterns” encoded in each filter of CNNs
- Building explainable models
- Semantic-level middle-to-end learning via human computer interaction

Diagnosis of CNN Representations

Obtain insight understanding of features encoded in a CNN

- Analyze CNN features from a global view
- extracts image regions that directly contribute the network output for a label/attribute
- Estimation of vulnerable points in the feature space
- Refine network representations based on the analysis of network feature spaces
- Discover potential, biased representations of a CNN

Paper Agenda

- Visualization of CNN representations in intermediate network layers
- Diagnosis of CNN representations
- **Disentanglement of “the mixture of patterns” encoded in each filter of CNNs**
- Building explainable models
- Semantic-level middle-to-end learning via human computer interaction

Disentangling CNN representations into explanatory graphs

Disentangling CNN features into human interpretable graphical representations

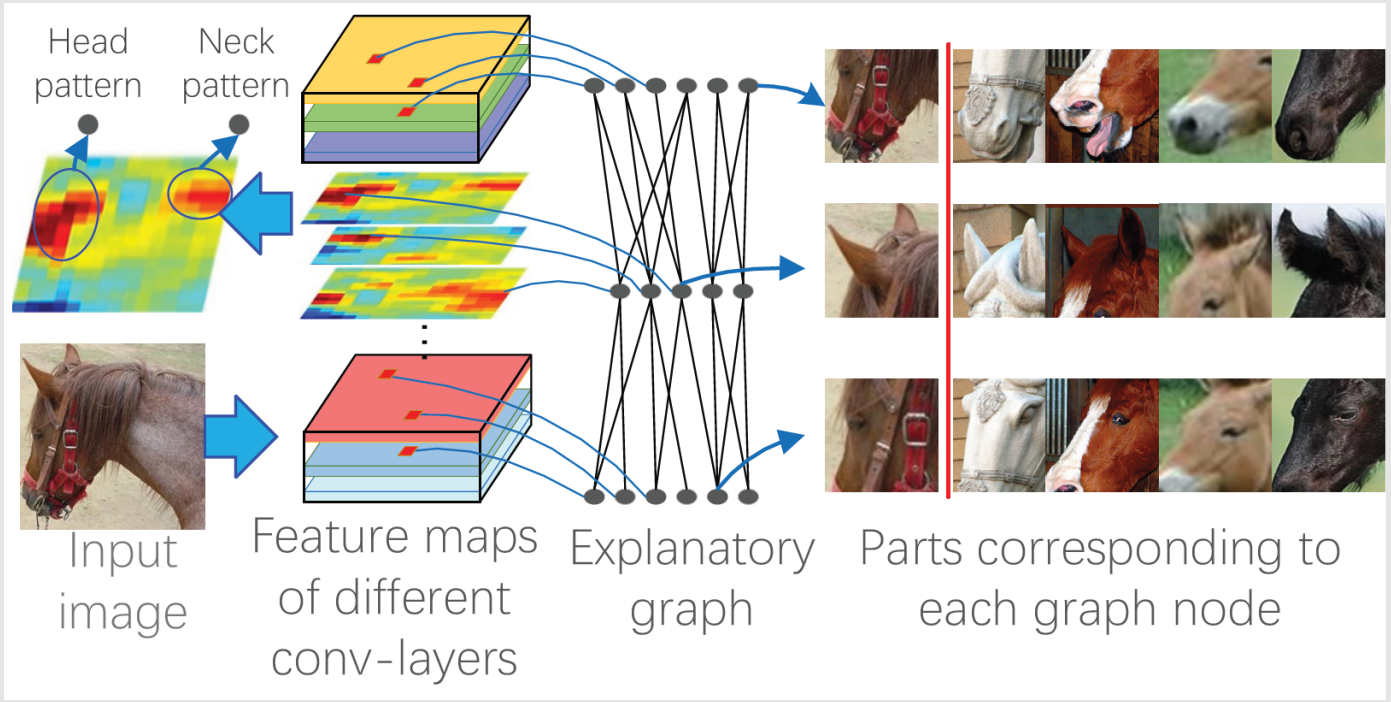
graphical model to represent the semantic hierarchy hidden inside a CNN.

- How many types of visual patterns are memorized by each convolutional filter of the CNN?
- Which patterns are co-activated to describe an object part?
- What is the spatial relationship between two co-activated patterns?

conv-layer represents a mixture of patterns

In each sub-feature, the filter is activated by various part patterns in an image. This makes it difficult to understand the semantic meaning of a filter.

Explanatory Graph



- Graph layer corresponds to a specific conv-layer
- Each node consistently represents the same object part through different images (multiple node → single filter)
- Each edge encodes the co-activation relationship and the spatial relationship between two nodes in adjacent layers
- E.G as a compression of feature maps of conv-layers

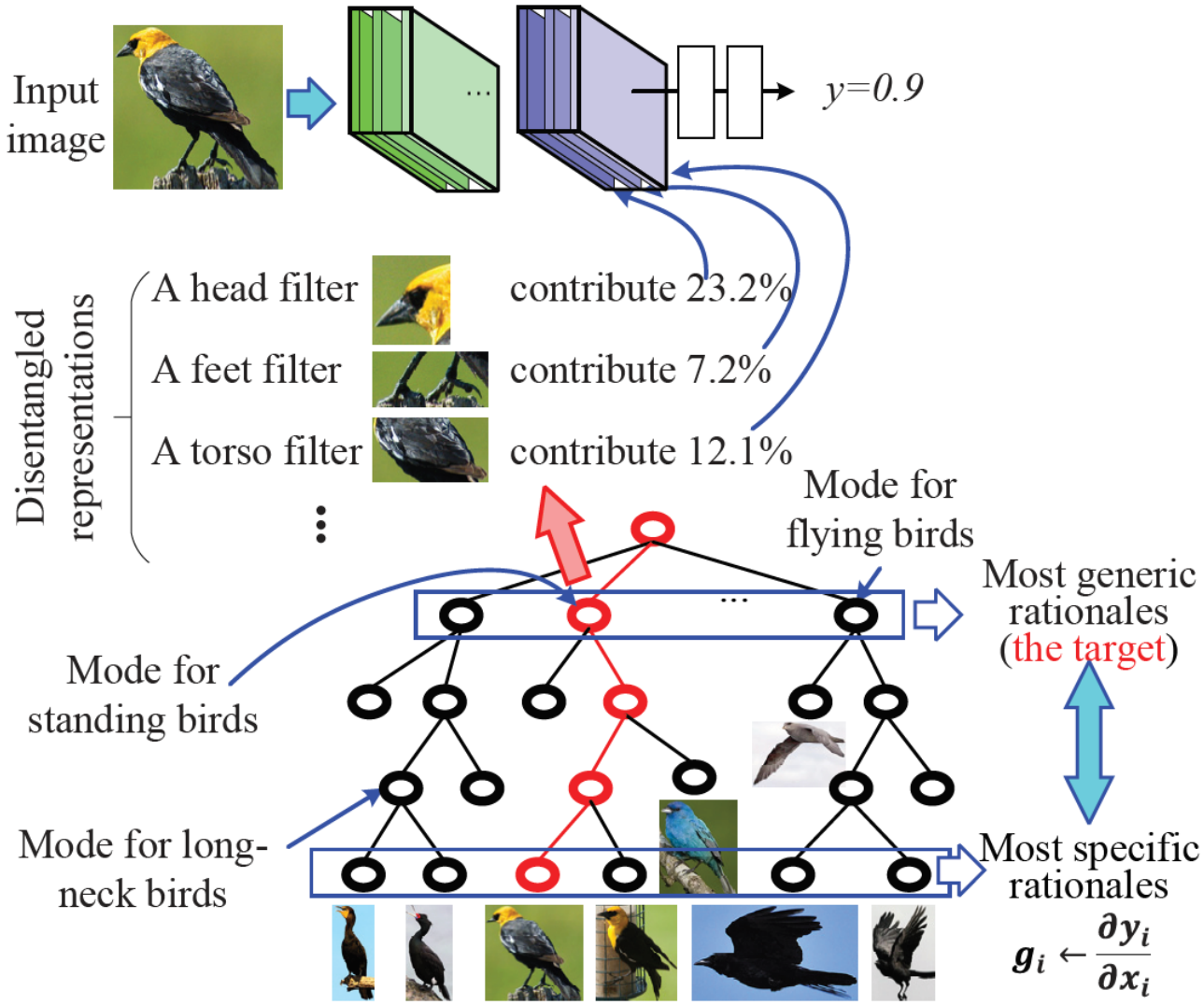
Disentangling CNN representations into decision trees

Explain the logic for each CNN prediction

- Which filters in a conv-layer are used for the prediction?
- How much filters contribute to the prediction?

Decision Tree

- Given an input image, the decision tree infers a parse tree (red lines) to quantitatively analyze rationales for the CNN prediction
- which object parts (or filters) are used for prediction and how much an object part (or filter) contributes to the prediction.



Paper Agenda

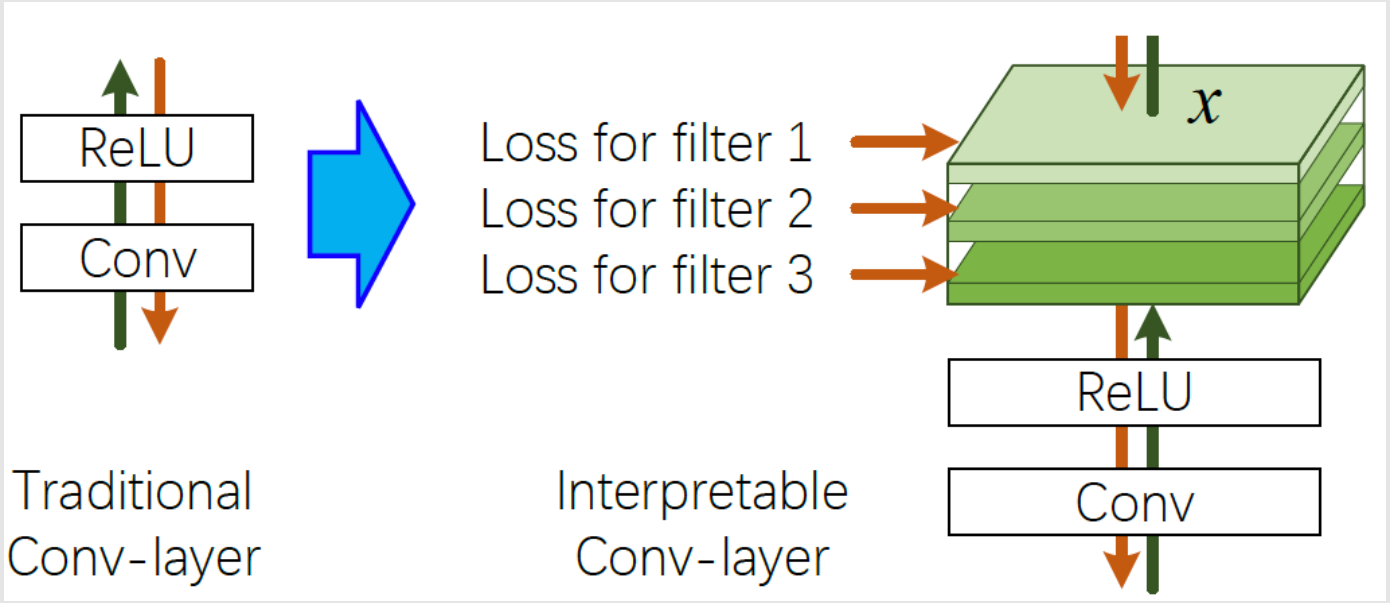
- Visualization of CNN representations in intermediate network layers
- Diagnosis of CNN representations
- Disentanglement of “the mixture of patterns” encoded in each filter of CNNs
- **Building explainable models**
- Semantic-level middle-to-end learning via human computer interaction

Building Explainable Models

neural networks where representations in middle layers are no longer a black box but have clear semantic meanings

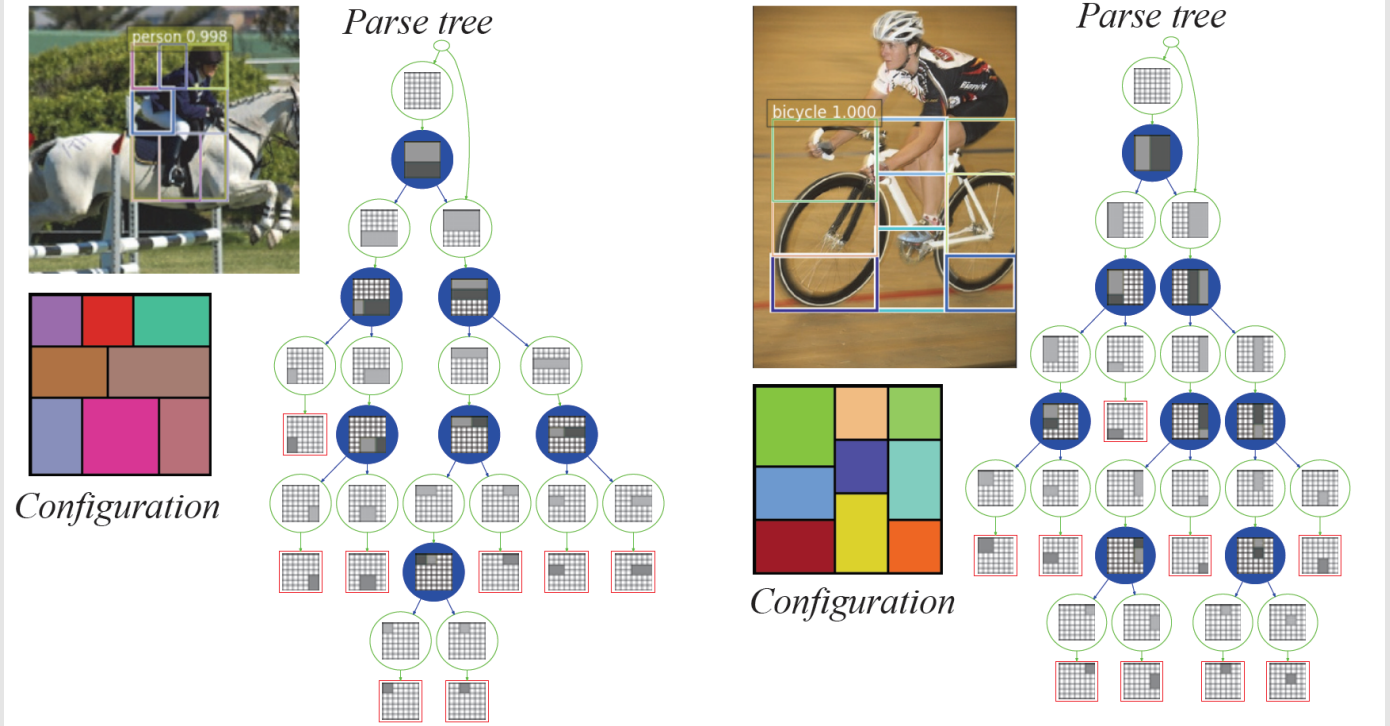
- Interpretable convolutional neural networks
- Interpretable R-CNN
- Capsule networks
- InfoGAN, Information maximizing generative adversarial nets

Interpretable convolutional neural networks



- Adding a loss to each filter in the conv-layers
- The loss is used to regularize the feature map towards the representation of a specific object part.
- Each filter in the conv-layer is assigned to a certain category.
  - If the input image belongs to the target category then the loss expects the filter's feature map to match a template well
  - Otherwise, the filter needs to remain inactivated

Interpretable R-CNN



- And-Or graph (AOG), to model latent configurations of objects
- During the detection process, a bounding box is interpreted as the best parse tree derived from the AOG on-the-fly
- Detects object bounding boxes
- Determines the latent parse tree and part configurations of objects
  - Qualitatively extractive rationale in detection

Capsule networks

- Capsules:
  - substitute for traditional neural units to construct a capsule network
- capsule outputs:
  - an activity vector instead of a scalar.
- The length of the activity vector:
  - represents the activation strength of the capsule
- Orientation of the activity vector:
  - Encodes instantiation parameters
- Active capsules in the lower layer send messages to capsules in the adjacent higher layer.
- Iterative routing-by-agreement mechanism to assign higher weights with the low-layer capsules whose outputs better fit the instantiation parameters of the high-layer capsule

InfoGAN

- Maximizes the mutual information between certain dimensions of the latent representation and the image observation.
- The InfoGAN separates input variables of the generator into two types
  - Incompressible noise z and
  - Latent code c
- Learn the latent code c to encode certain semantic concepts in an unsupervised manner

Paper Agenda

- Visualization of CNN representations in intermediate network layers
- Diagnosis of CNN representations
- Disentanglement of “the mixture of patterns” encoded in each filter of CNNs
- Building explainable models
- **Semantic-level middle-to-end learning via human computer interaction**

Semantic-level middle-to-end learning via human computer interaction

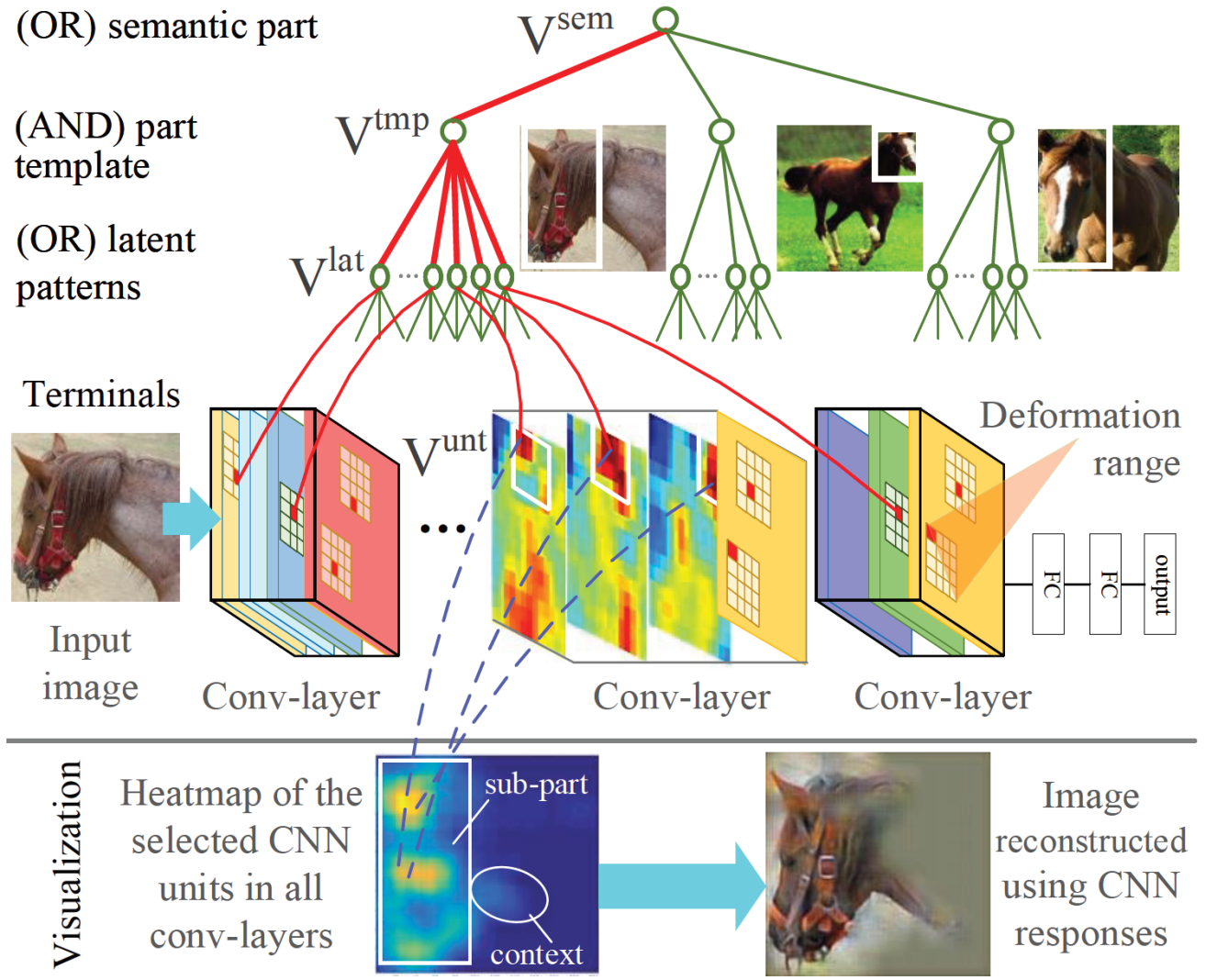
- Active question-answering for learning And-Or graphs
  - Extract a four-layer interpretable And-Or graph (AOG) to explain the semantic hierarchy hidden in a CNN
- Interactive manipulations of CNN patterns

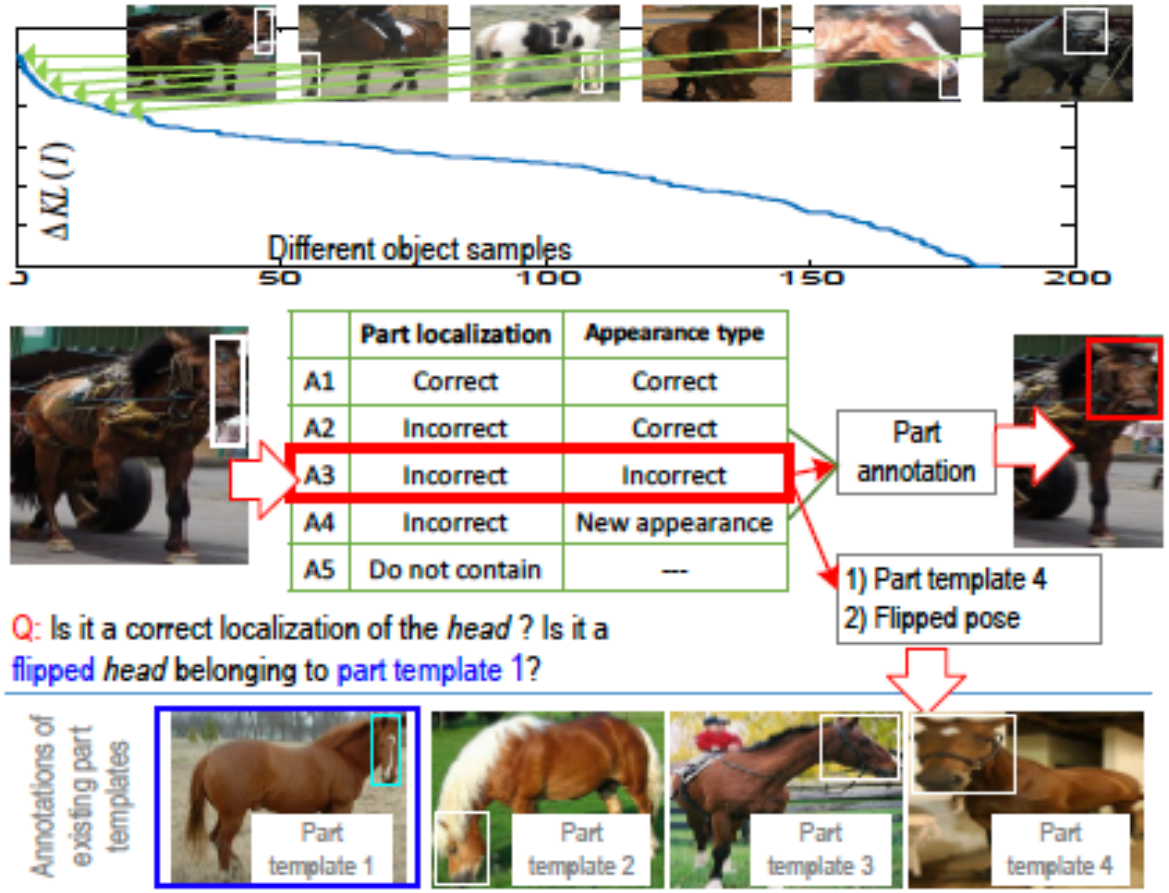
Active question-answering for learning And-Or graphs

four-layer semantics,

- semantic part (OR node),
- Part templates (AND nodes),
- latent patterns (OR nodes),
- neural units (terminal nodes) on feature maps.

To learn an AOG, allows the computer to actively identify and ask about objects, whose neural patterns cannot be explained by the current AOG





Active Question-Answering

- The current AOG is used to localize object parts among all the unannotated images
- Actively selects objects that cannot well fit the AOG
  - unexplained objects.
- Predicts the potential gain of asking about each unexplained object
- Determines the best sequence of questions
  - Template types? bounding boxes?
- Uses the answers to
  - Refine an existing part template
  - Mine latent patterns for new object-part templates
- Grow AOG branches

Interactive manipulations of CNN patterns

Transfer CNN patterns to model object parts

- Mines object part patterns from the CNN
- Part annotations on very few object images for supervision
  - Find bounding-box annotation of a part
- Mine latent patterns from conv-layers of the CNN
- AOG is used to organize all mined patterns
- Visualizes the mined latent patterns
- Asks people to remove latent patterns unrelated to the target part interactively
- People can simply prune incorrect latent patterns from AOG branches to refine the AOG