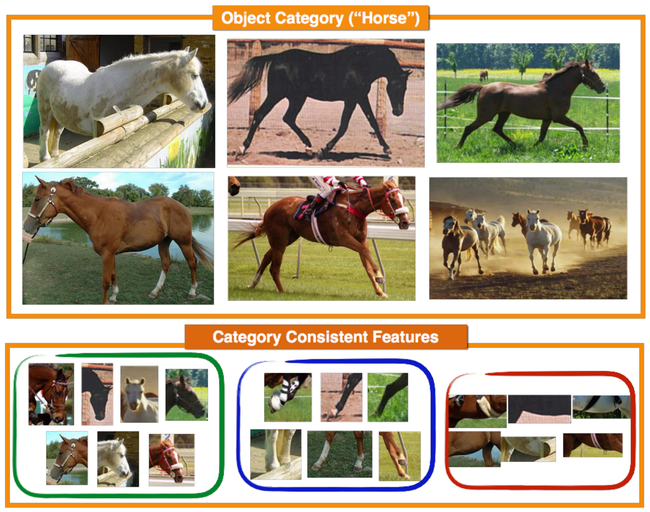
Object detection and segmentation for free using category-consistent CNN features

Hieu Le[1], Chen-Ping Yu[1], Dimitris Samaras[1] & Gregory Zelinsky[1,2]
[1]Department of Computer Science; [2]Department of Psychology; Stony Brook University

Introduction

What are category-consistent features?

- Introduced recently by Yu et al [1], the term *category-consistent feature* (CCF) refers to those features that are most representative [2] of an object category. They are the features of an object category that we learn from encounters with that category throughout our lives. What CCFs do you use to recognize a horse?



- CCFs are important because these features are the basis by which we interact with objects as part of our goal-directed behavior. For example, the method from Yu et al [1] generatively learned the visual CCFs for 64 object categories, then used them to predict how strongly eye gaze was guided to different target goals in the context of a search task, and how long it took to make category judgements about targets following their fixation.

Why build a new CCF model?

- There are various ways of selecting the most representative features from the rest, and there are many feature spaces to explore. There are also different behaviors to predict. These questions were addressed in a new CCF model.

1. **CNN Features**. The method from [1] used relatively simple SIFT and color histogram features, but newer and more biologically-plausible [3] methods exist to obtain far more powerful features to use as CCFs. One of our goals is to extract visual features from category exemplars using a *convolutional neural network (CNN).*

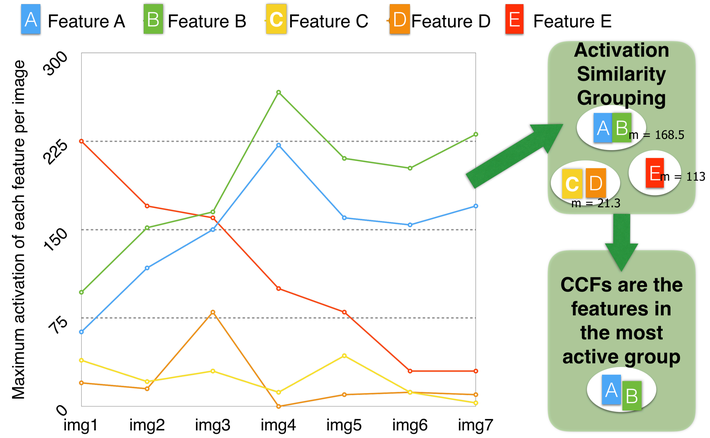
2. **CCF Feature Selection**. [1] also selected CCFs based on two criteria: CCFs must have high frequency (i.e., appear often among the category's exemplars) and low variability (i.e., appear consistently among the category's exemplars). Another goal will be to develop a method of selecting CCFs based instead on *feature activation similarity*.

3. **Object Segmentation**. Rather than using CCFs to predict target guidance and verification during search, a final goal will use CCF *co-occurrence* to localize or segment target objects in images by combining local clusters of CCFs into larger visual fragments (proto-objects).

Model	Features	CCF selection	Goals
Yu et al [1]	• SIFT and Color histogram • Bag-of-Words	• High activation • Low variability	Predict effect of category hierarchy on target guidance and verification in search
Ours	CNN features	• High activation • Activation similarity	To segment exemplars of target categories from images

CCF Selection

- Yu et al's [1] criterion of selecting CCFs based on feature frequency is reasonable, but their low-variability (high consistency) criterion is perhaps overly strong given the often extreme variability among category exemplars.
- Alternatively, we introduce a method based on feature activation similarity that groups features that are similarly active.

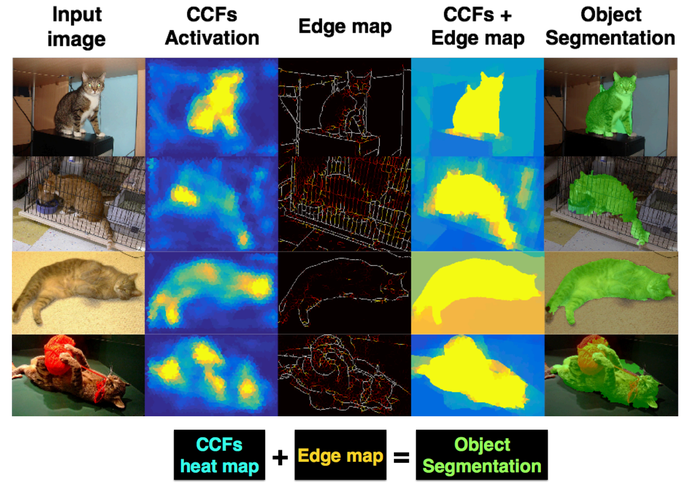


- CNN features are extracted from image exemplars of random categories. *VGG19* [7] was used in the present study, but our method is generic in that it can select CCFs from any layer of any CNN.
- Features are grouped using k-means (*L-p* distance) based on their similarity in activation values. Features (**A**,**B**) would be grouped, as would features (**C**,**D**). Feature (**E**) would remain ungrouped. Note that this method emphasizes interactions between features; features are grouped if they have similar activations across exemplars, regardless of whether these activations are strong or weak.
- CCFs are defined as the features (or feature singleton) in the group having the highest average activation.

What do our CCFs look like?

- Examples of CCFs selected for the "airplane" category from VOC 2012. All CCFs are from the last convolutional layer of VGG19.
- The first column shows the input exemplars, the other columns show the activated areas corresponding to the top 4 CCFs selected by our method.
- Note that the CCFs capture many of the representative features or parts of planes, such as the fuselage and wings, the landing gear, and the tail. Note also that the activated regions all fall within the spatial extent of the objects.

Object Segmentation

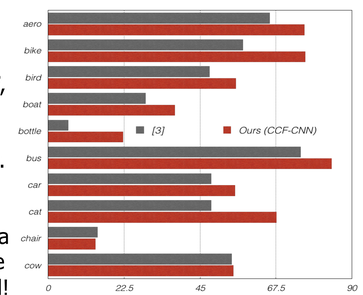


- Object Segmentation is the specification of an object's region in an image. Even with very good CNN features, this is a difficult and unsolved problem that humans can perform effortlessly.
- Our method for doing this propagates values in a CCF-CNN heat map, indicating the rough locations of object parts, throughout a closed boundary defined by a supervised edge detection method [8].

How well does it work?

- The CorLoc metric [4] is used to evaluate segmentation success. An object is considered correctly localized if a tight bounding box drawn around the predicted segment matches at least one human-annotated bounding box based on a 50% Intersection-over-Union (IoU) score. We evaluate our method on three commonly used datasets: VOC 2007 and 2012 [5], and the Object Discovery dataset [6]. Following [3], we also to test our method on the six held-out ImageNet subsets.

- Our method achieves state-of-the-art CorLoc scores for all three benchmarks: 41.97% for VOC 2007, 48.2% for VOC 2012, 86.3% for Object Discovery, and 60.95% for the held-out subset from ImageNet.
- As shown by the plot, our method also localizes many of the VOC 2012 object categories better than a competitor, despite not knowing the category of object to be segmented!



Conclusion

- Our CCF-CNN method, by exploiting feature activation similarity and CCF co-occurrence, was able to discover or "preattentively localize" categories of objects in scenes. This provides a hint at how people might perform this most basic of operations underlying object-based perception.

References

[1] Chen-Ping Yu, Justin T. Maxfield, Gregory J. Zelinsky (2016), "Searching for Category-Consistent Features A Computational Approach to Understanding Visual Category Representation". Psychological Science.
[2] Kahneman, Daniel; Tversky, Amos (1972). "Subjective probability: A judgment of representativeness". Cognitive Psychology. 3 (3): 430–454
[3] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Image colocalization by mimicking a good detector's confidence score distribution. In ECCV, 2016.
[4] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. International Journal of Computer Vision
[5] M. Everingham, S. Eslami, L. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. IJCV, 2015.
[6] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.
[8] Larry Zitnick Piotr Dollar. Structured forests for fast edge detection. In ICCV. International Conference on Computer Vision, December 2013.