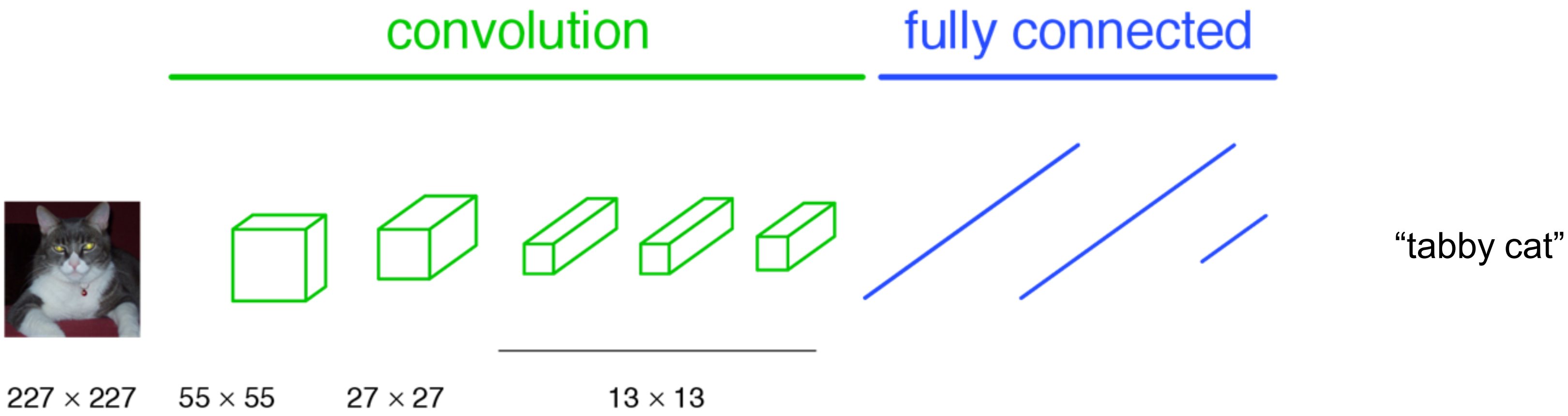
Structured Predictions with Deep Learning

James Hays

Recap of previous lecture

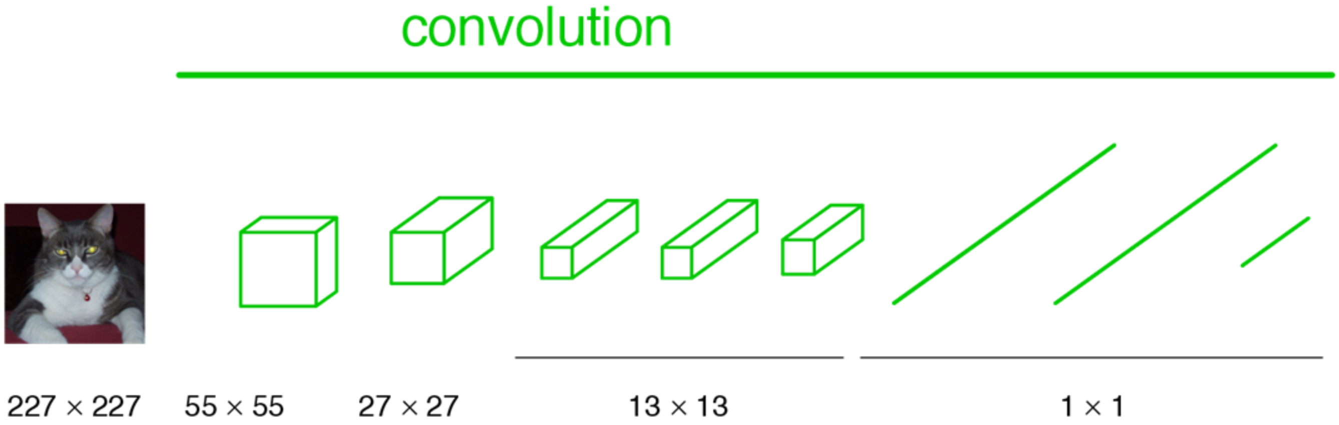
- COCO dataset. Instance segmentation of 80 categories. Keypoints + Language + other annotations, as well.
- Deeper deep models
    - VGG networks
    - GoogLeNet built from “Inception” modules
    - ResNet
- Deeper networks seem to work better than the equivalent shallow network with the same number of parameters, but they aren’t trivial to train.

a classification network



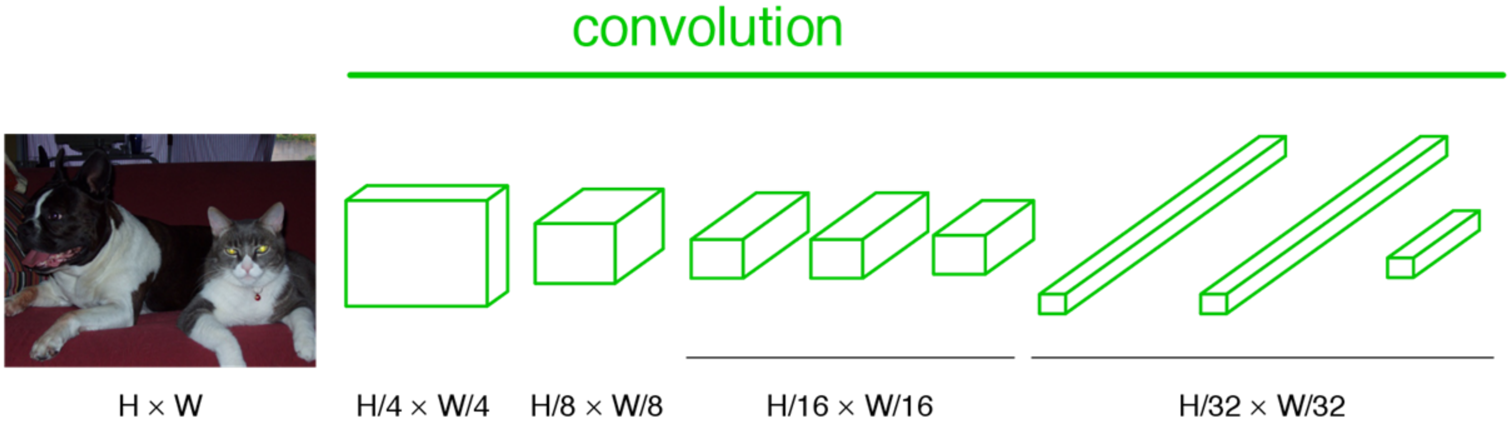
Fully Convolutional Networks for Semantic Segmentation.
Jon Long, Evan Shelhamer, Trevor Darrell. CVPR 2015

becoming fully convolutional

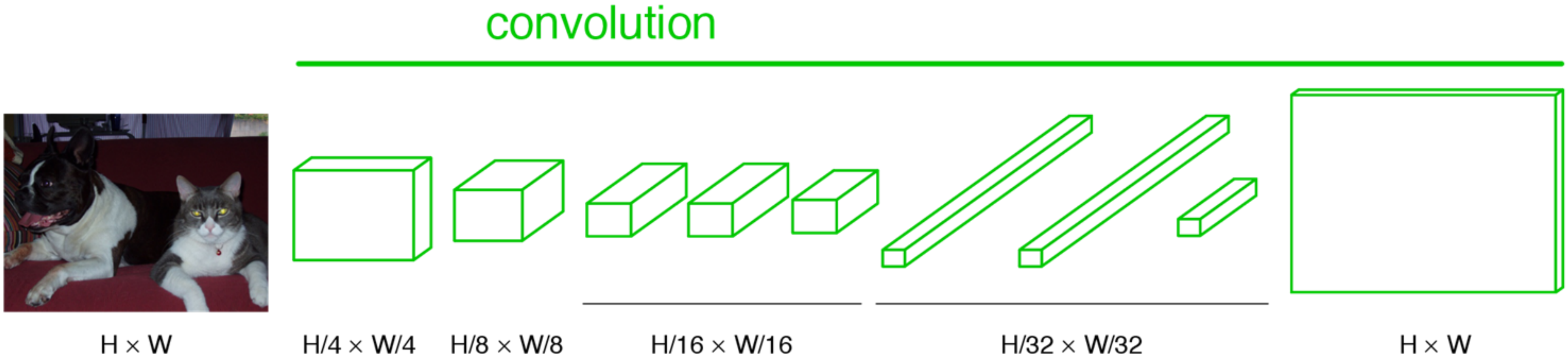


Note: “Fully Convolutional” and “Fully Connected” aren’t the same thing.
They’re almost opposites, in fact.

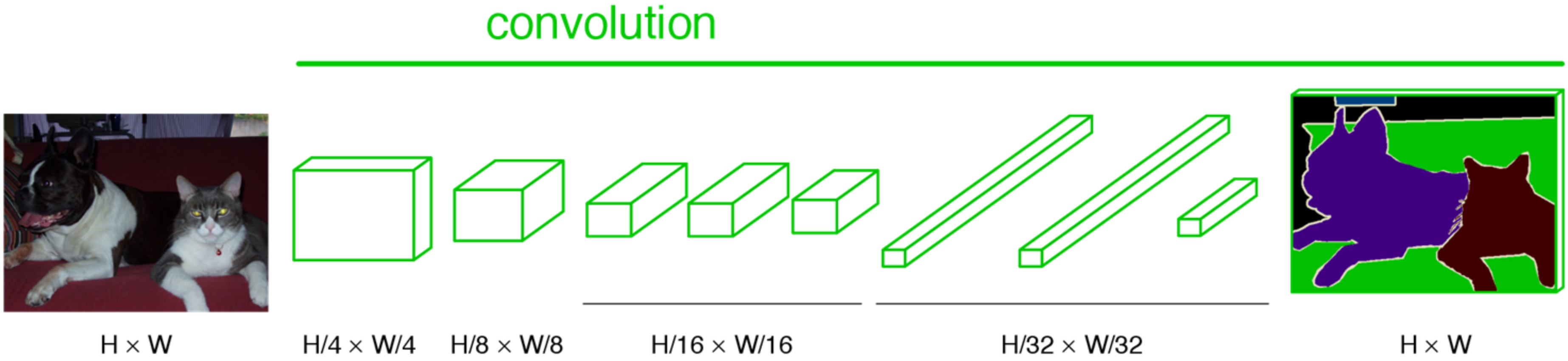
becoming fully convolutional



upsampling output



end-to-end, pixels-to-pixels network



What if we want other types of outputs?

- Easy: Predict any number of labels (with classification, there will be just one best answer, but for other labels like attributes dozens could be appropriate for an image)

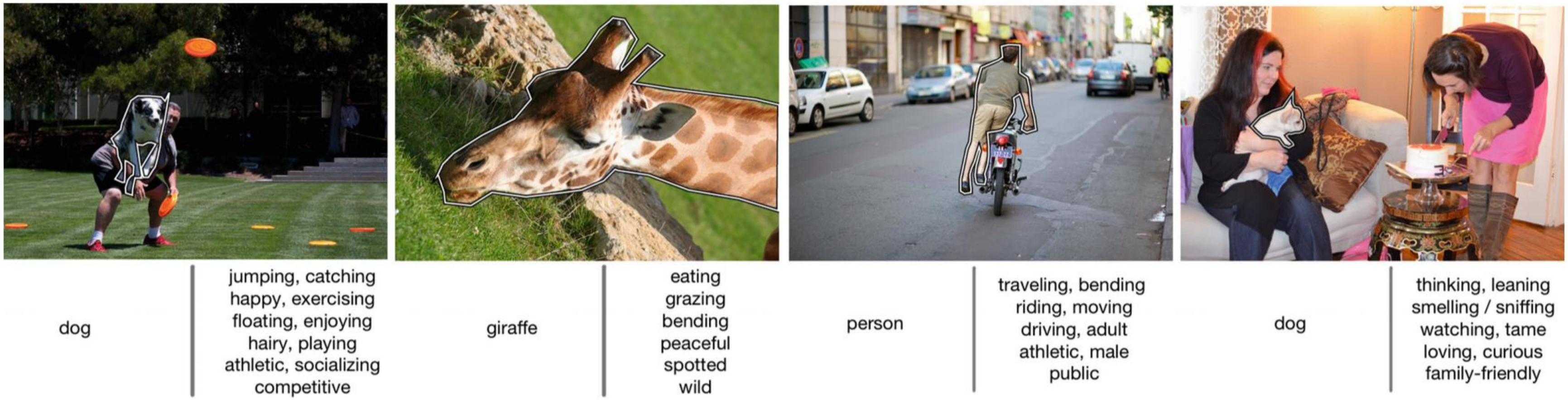
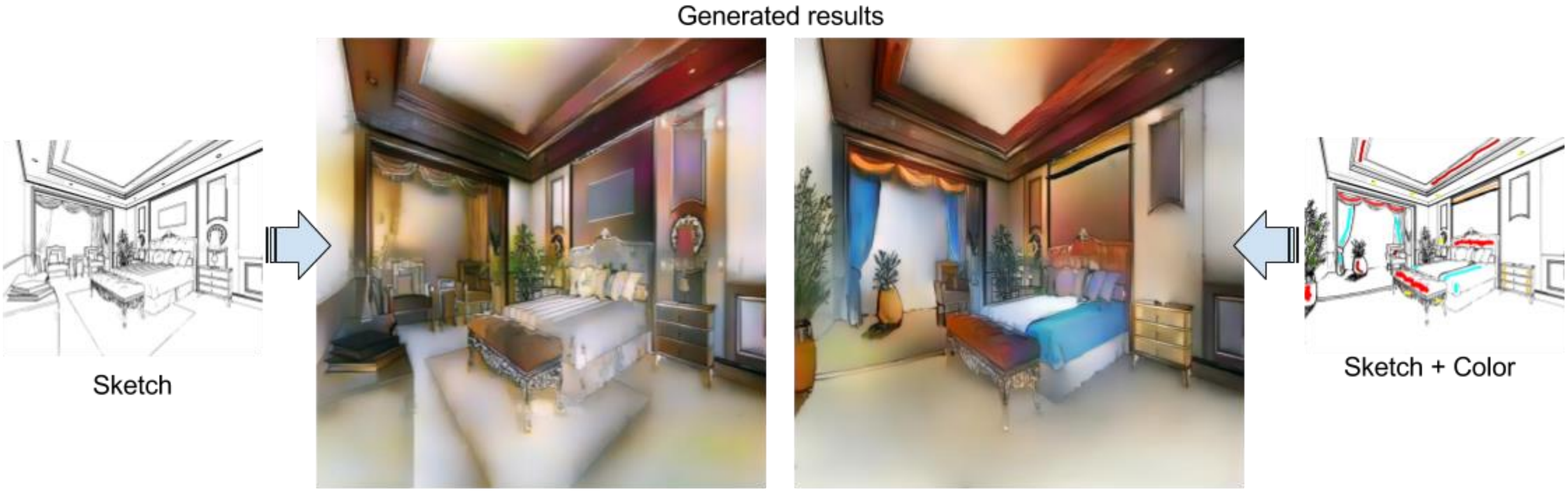


Fig. 1. *Examples from COCO Attributes.* In the figure above, images from the COCO dataset are shown with one object outlined in white. Under the image, the COCO object label is listed on the left, and the COCO Attribute labels are listed on the right. The COCO Attributes labels give a rich and detailed description of the context of the object.

What if we want other types of outputs?

- Easy*: Predict any fixed dimensional output, whether a feature (embedding networks) or an image.



Scribbler: Controlling Deep Image Synthesis with Sketch and Color.
Sangkloy, Lu, Chen Yu, and Hays. CVPR 2017

*easy to design an architecture. Not necessarily easy to get working.

What if we want other types of outputs?

- Hard: Outputs with varying dimensionality or cardinality
  - A natural language image caption
  - An arbitrary number of human keypoints (17 points each)
  - An arbitrary number of bounding boxes (4 parameters each)
- Today we will examine state-of-the-art methods for keypoint prediction and object detection

Convolutional Pose Machines

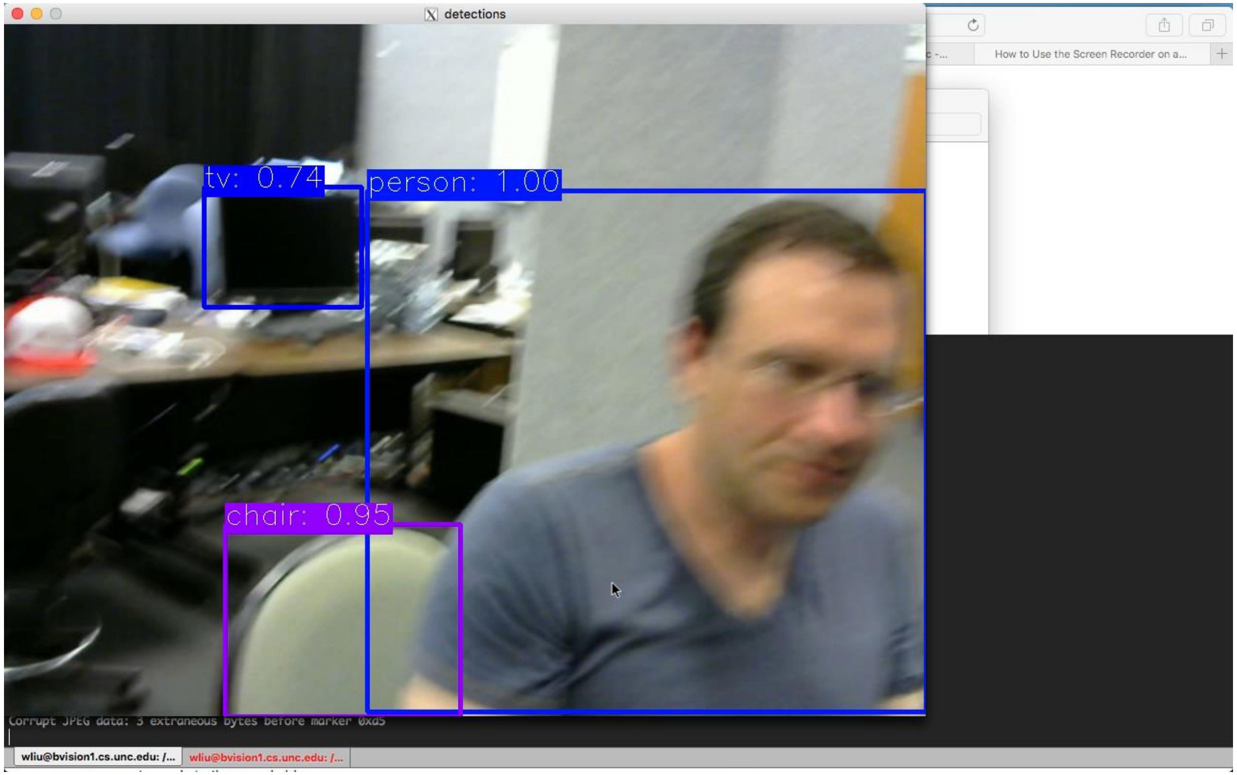
- Variant of Convolutional Pose Machines that won the inaugural COCO keypoint challenge.
- http://image-net.org/challenges/talks/2016/Multi-person%20pose%20estimation-CMU.pdf
- Videos: https://www.youtube.com/playlist?list=PLNh5A7HtLRcpsMfvyG0DED-Dr4zW5Lpcg

Multi-Person Pose Estimation using Part Affinity Fields

Zhe Cao, Shih-En Wei, Tomas Simon, Yaser Sheikh
Carnegie Mellon University

SSDBox

- Object Detector that is very nearly state-of-the-art accuracy and very, very fast
- http://www.cs.unc.edu/~wliu/papers/ssd_eccv2016_slide.pdf


detections
How to Use the Screen Recorder on a...
tv: 0.74
person: 1.00
chair: 0.95
wliu@bvision1.cs.unc.edu: /...
wliu@bvision1.cs.unc.edu: /...

```
Corrupt JPEG data: 3 extraneous bytes before marker 0xd5
```