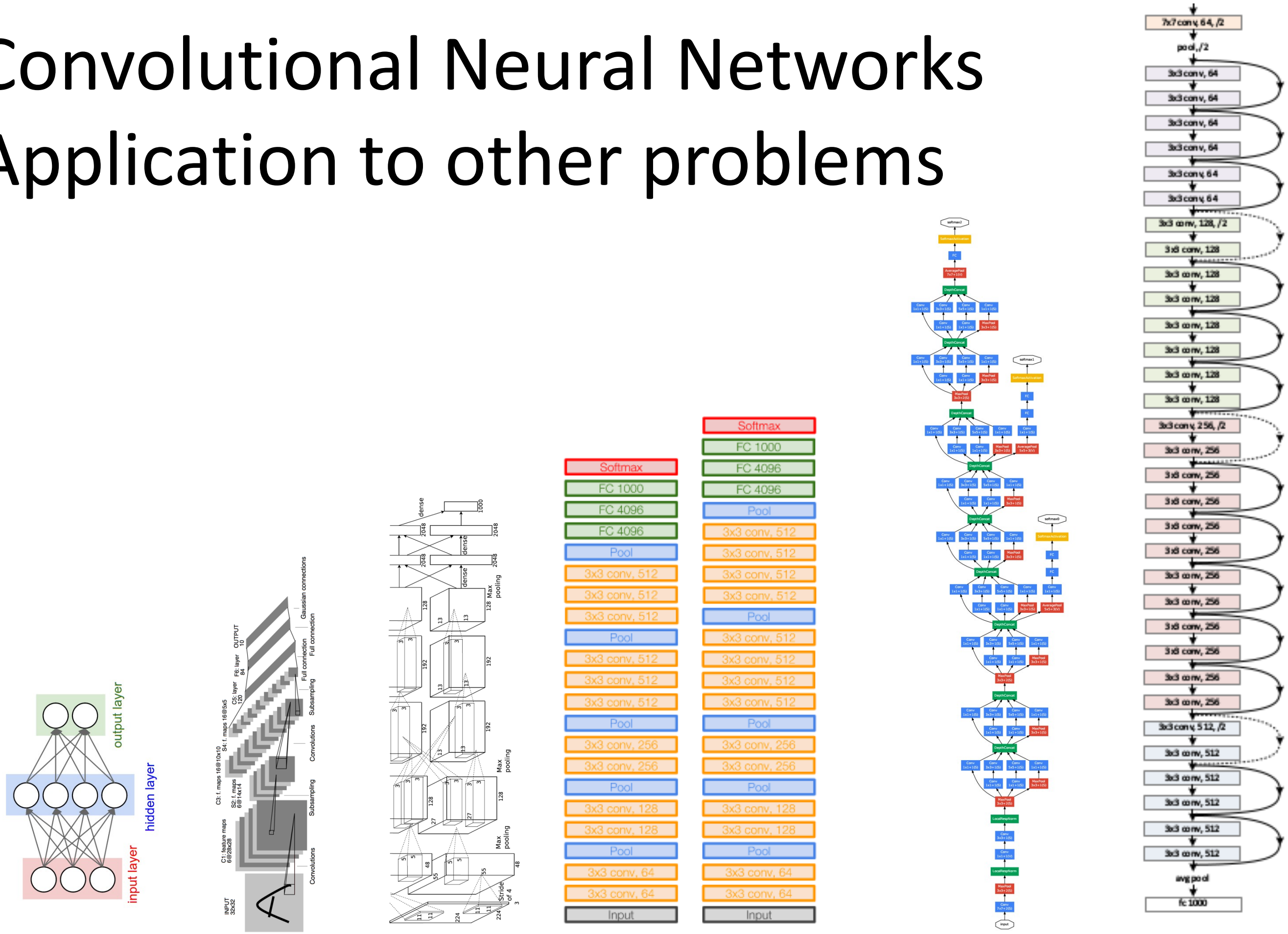
CSCI 497P/597P: Computer Vision

Convolutional Neural Networks
Application to other problems



Announcements

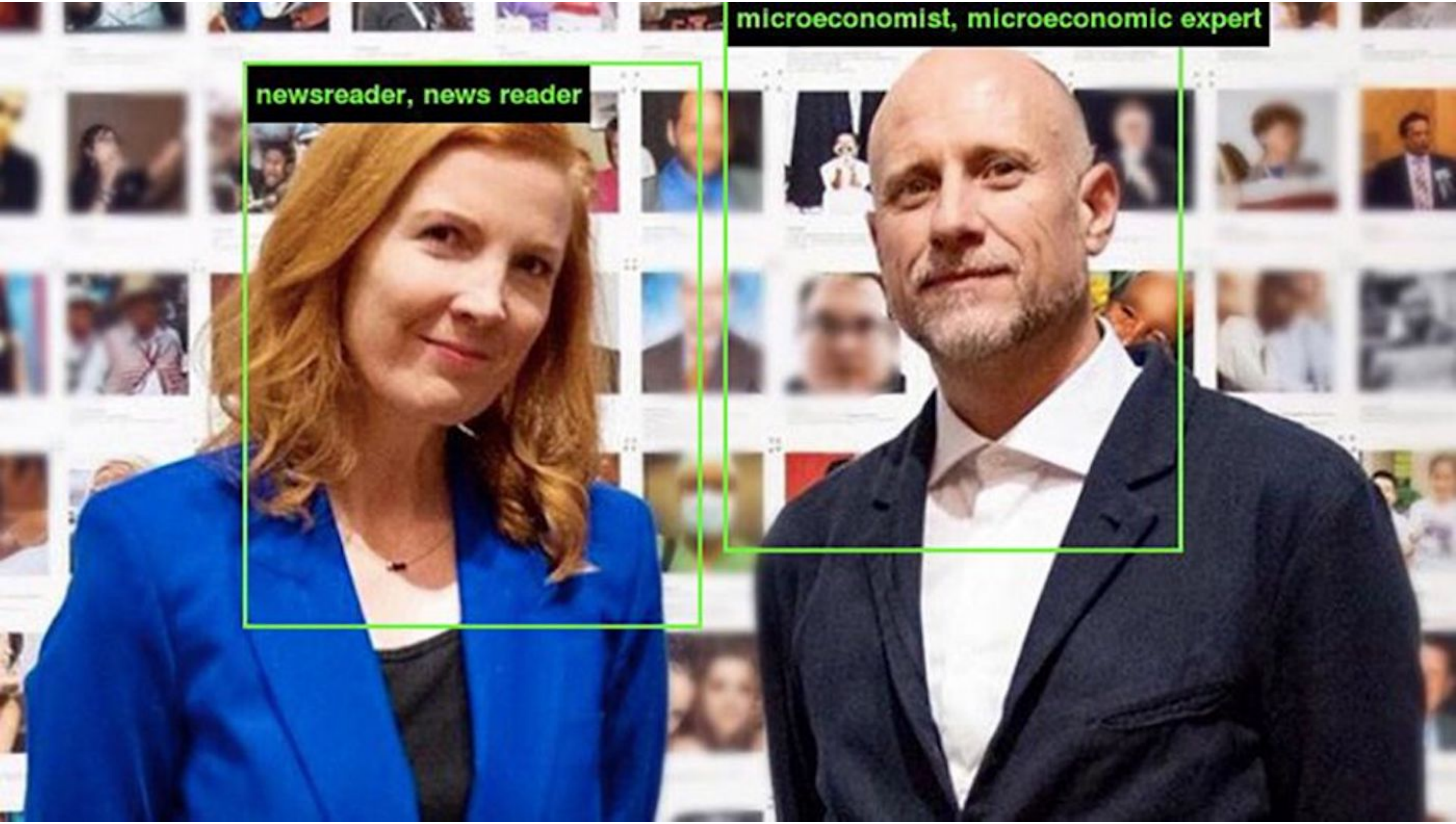
- Midterm and P3 grades are out
- Final Reminder: Friday is the last day to opt into letter grades for 497P students.

And so on and so forth…

- So we've beat the crap out of ImageNet… what now?
  - Can we do image classification on other datasets?
  - **Can we do things other than image classification?**

But first…

- A brief note about datasets.
- ImageNet is a collection of images with labels
  - The 1000 classes used for evaluation are a tiny subset of the tags available.
  - The labels were produced by humans.



AI Researcher

Artist

Lil Uzi Hurt at Home
@lostblackboy

No matter what kind of image I upload, ImageNet Roulette, which categorizes people based on an AI that knows 2500 tags, only sees me as Black, Black African, Negroid or Negro.

Some of the other possible tags, for example, are “Doctor,” “Parent” or “Handsome.”

511 5:08 PM - Sep 17, 2019 · Brooklyn, NY

184 people are talking about this

- **The viral selfie app ImageNet Roulette seemed fun – until it called me a racist slur**
  The Guardian, September 2019
  https://www.theguardian.com/technology/2019/sep/17/imagenet-roulette-asian-racist-slur-selfie

600,000 Images Removed from AI Database After Art Project Exposes Racist Bias

The image tagging system that went viral on social media was part of artist Trevor Paglen and AI researcher Kate Crawford's attempts to publicize how prejudiced technology can be.

https://hyperallergic.com/518822/600000-images-removed-from-ai-database-after-art-project-exposes-racist-bias/

Dataset bias

LFW [Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Huang et al.]	77.5% male 83.5% white
IJB-A [Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark. Klare et al.]	79.6% lighter-skinned
Adience [Age and gender classification using convolutional neural networks. Levi and Hassner.]	86.2% lighter-skinned

[Buolamwini and Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification]

Slide: Timnit Gebru, Emily Denton

Error Rate(1-PPV) By Female x Skin Type

	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
	1.7%	1.1%	3.3%	0%	23.2%	25.0%
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%

Buolamwini & Gebru FAT* 2018, Slides from Joy Buolamwini

Slide: Timnit Gebru, Emily Denton

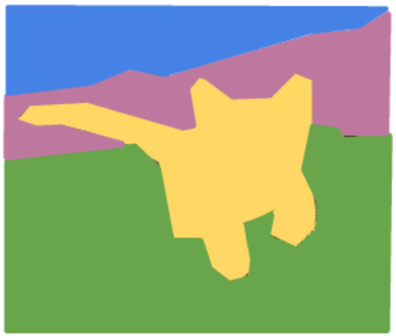
See also

- Writeup on ImageNet Bias: https://www.excavating.ai/

- ACM Conference on Fairness, Accountability, and Transparency: https://facctconference.org/index.html

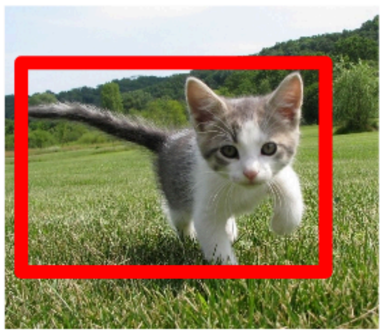
Other Computer Vision Tasks

Semantic Segmentation



GRASS, CAT, TREE, SKY

Classification + Localization

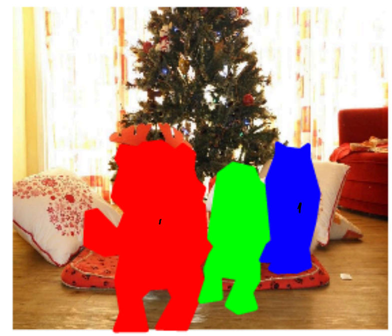


CAT

Object Detection

DOG, DOG, CAT

Instance Segmentation



DOG, DOG, CAT

No objects, just pixels

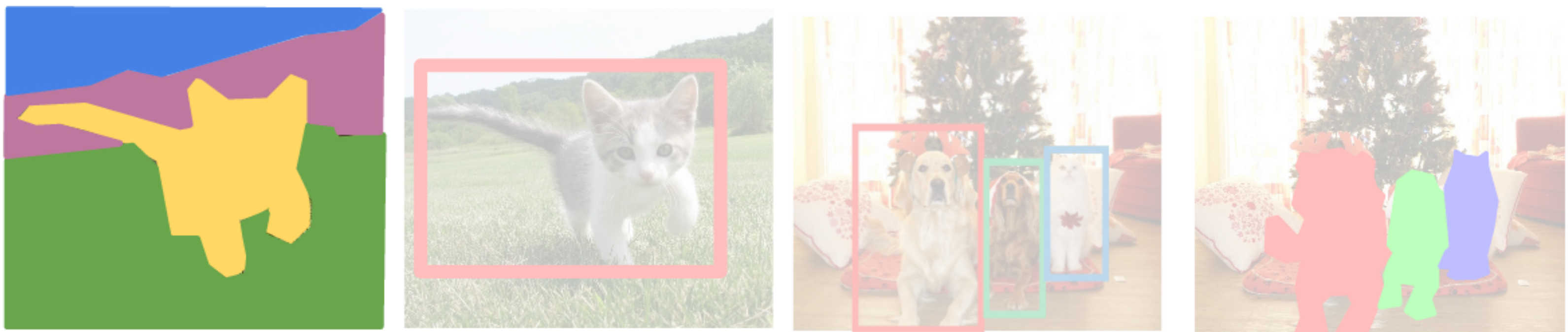
Single Object

Multiple Object

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Other Computer Vision Tasks

Semantic Segmentation | Classification + Localization | Object Detection | Instance Segmentation



GRASS, CAT, TREE, SKY | CAT | DOG, DOG, CAT | DOG, DOG, CAT

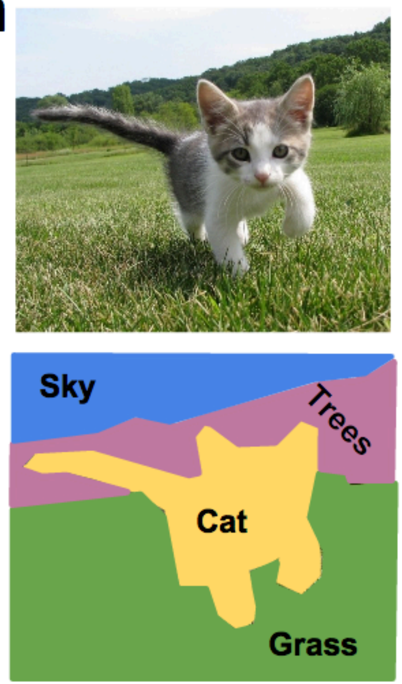
No objects, just pixels | Single Object | Multiple Object

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

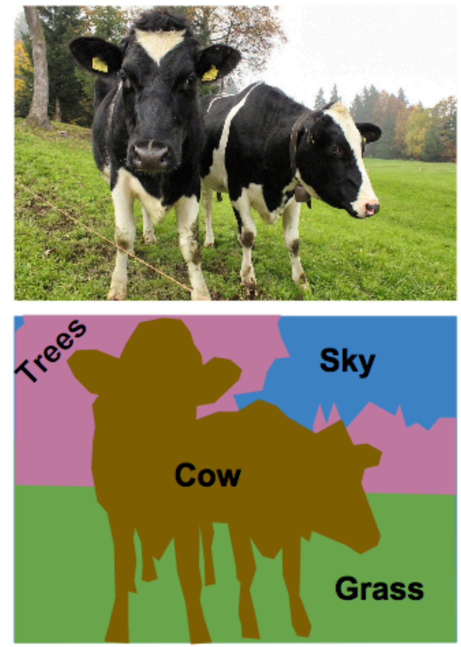
Semantic Segmentation

Label each pixel in the image with a category label

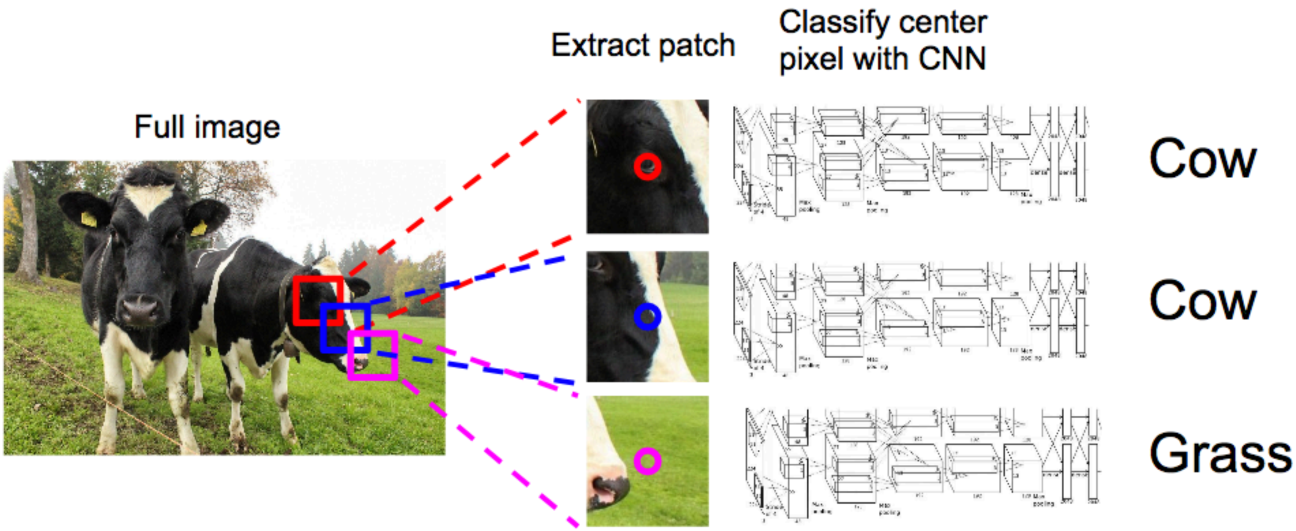
Don't differentiate instances, only care about pixels



This image is CC0 public domain



Semantic Segmentation Idea: Sliding Window



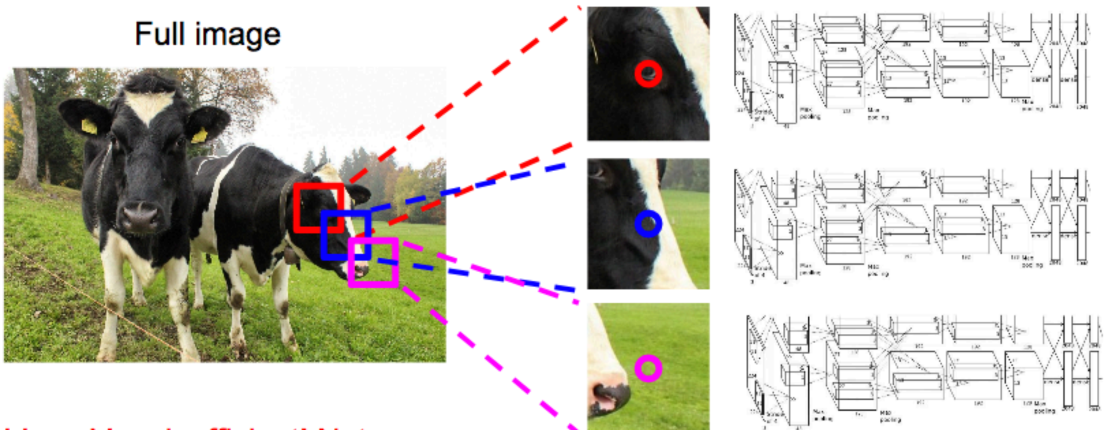
Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation Idea: Sliding Window

Full image

Extract patch

Classify center pixel with CNN



Cow

Cow

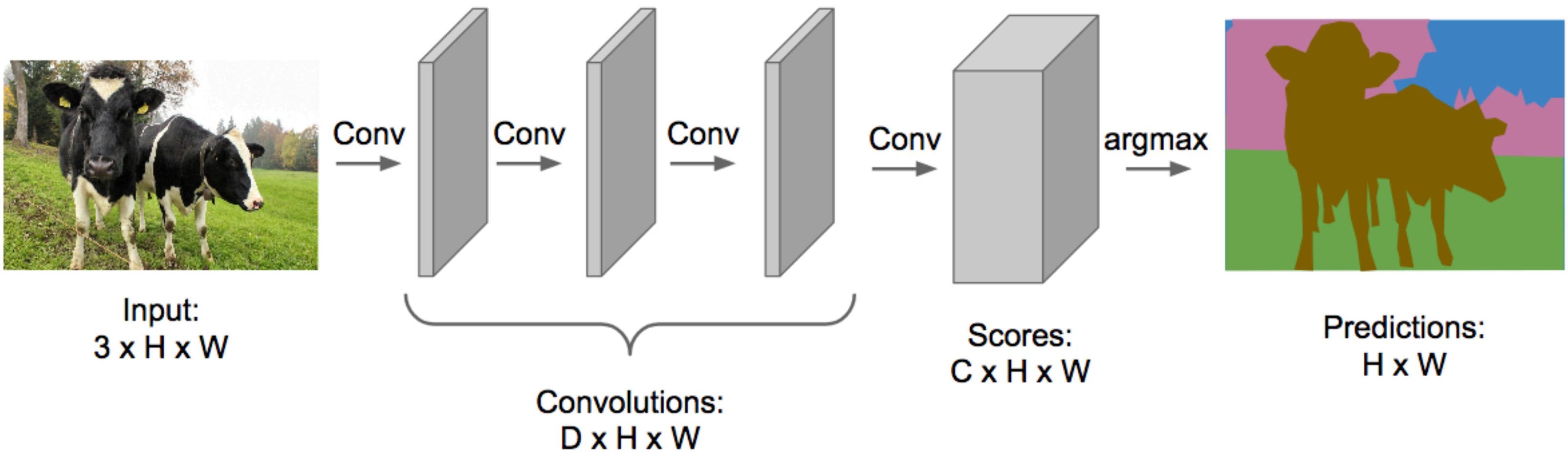
Grass

Problem: Very inefficient! Not reusing shared features between overlapping patches

Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

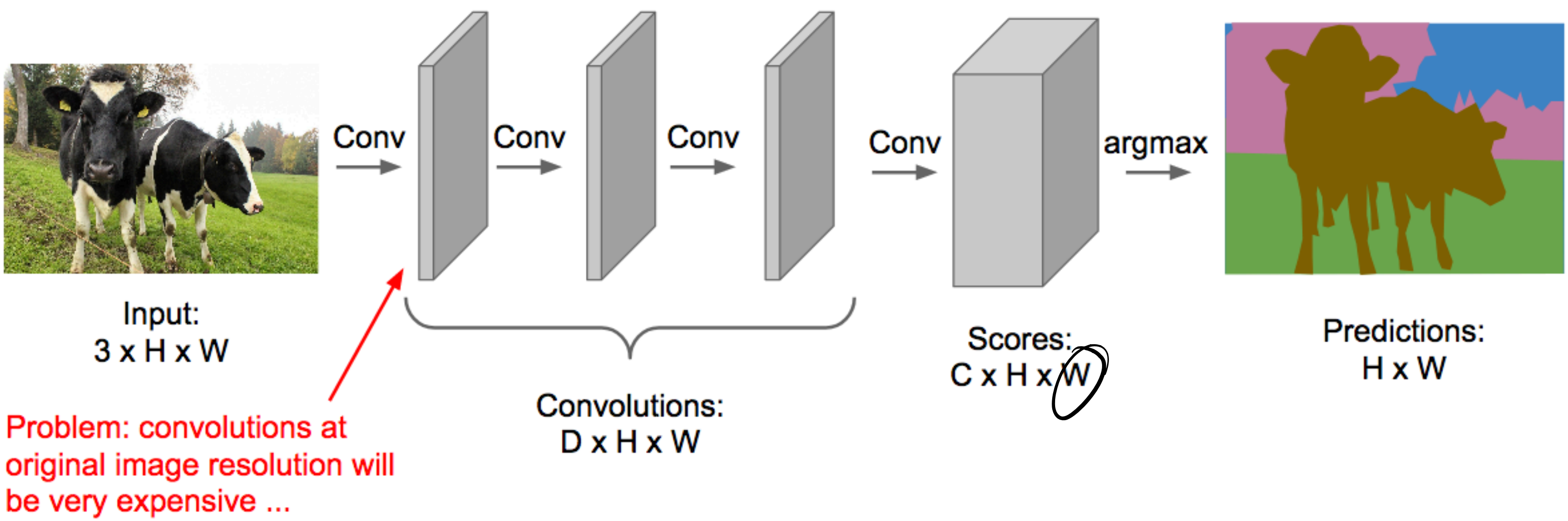
Semantic Segmentation Idea: Fully Convolutional

Design a network as a bunch of convolutional layers
to make predictions for pixels all at once!



Semantic Segmentation Idea: Fully Convolutional

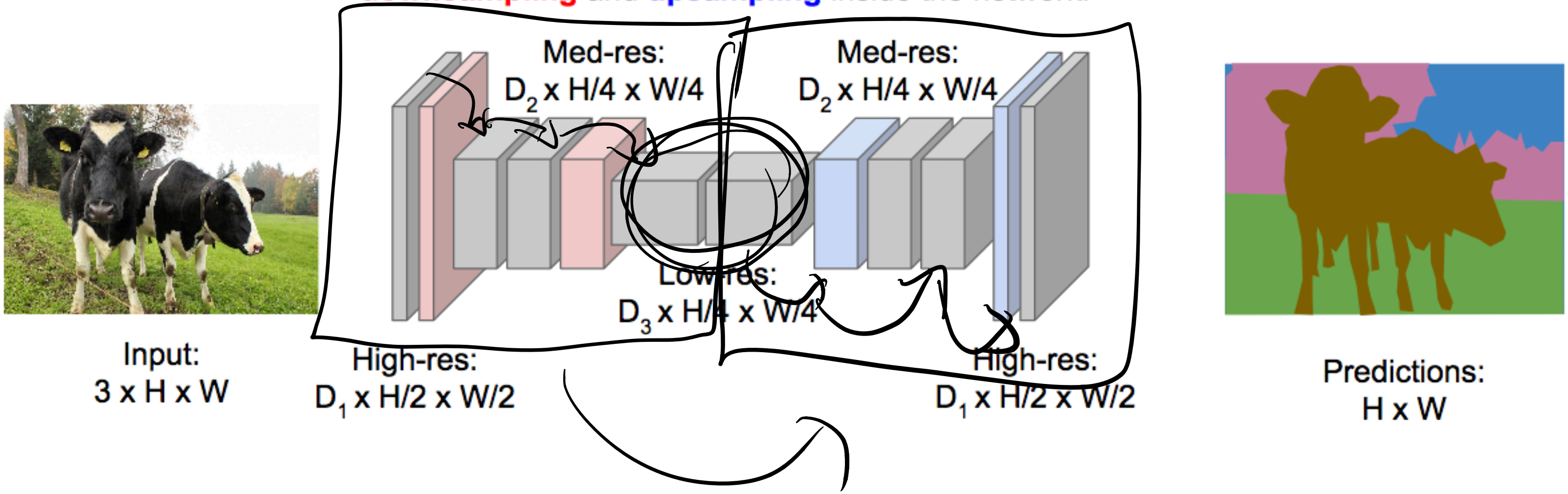
Design a network as a bunch of convolutional layers
to make predictions for pixels all at once!



Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Semantic Segmentation Idea: Fully Convolutional

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!



Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

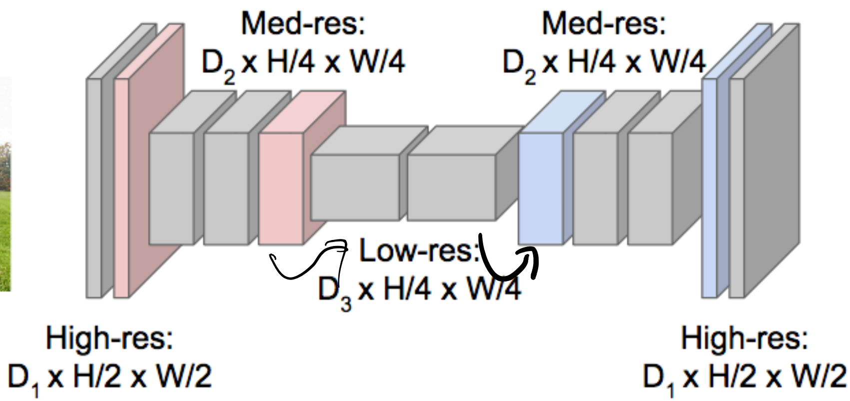
Semantic Segmentation Idea: Fully Convolutional

Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling: ???

Input:
$3 \times H \times W$

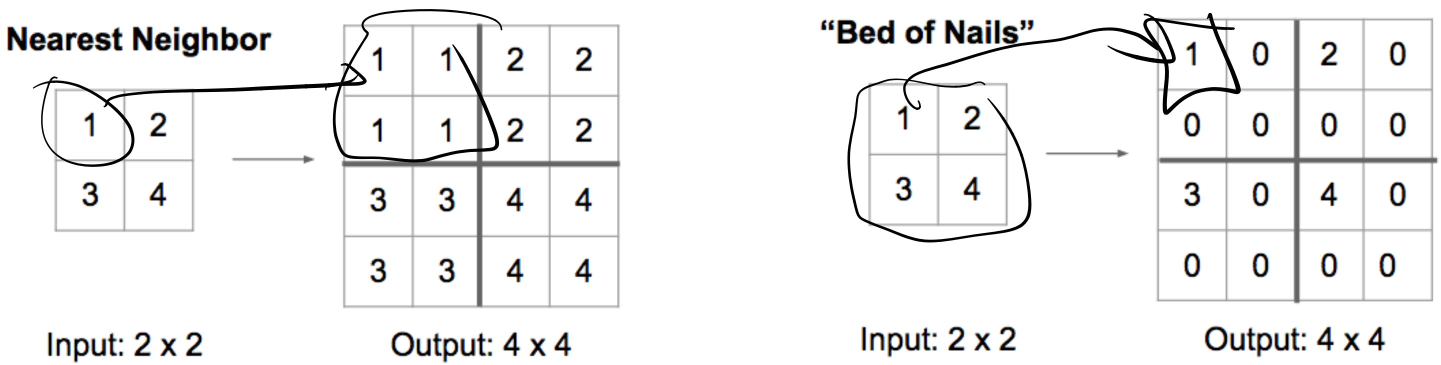


Predictions:
$H \times W$

Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

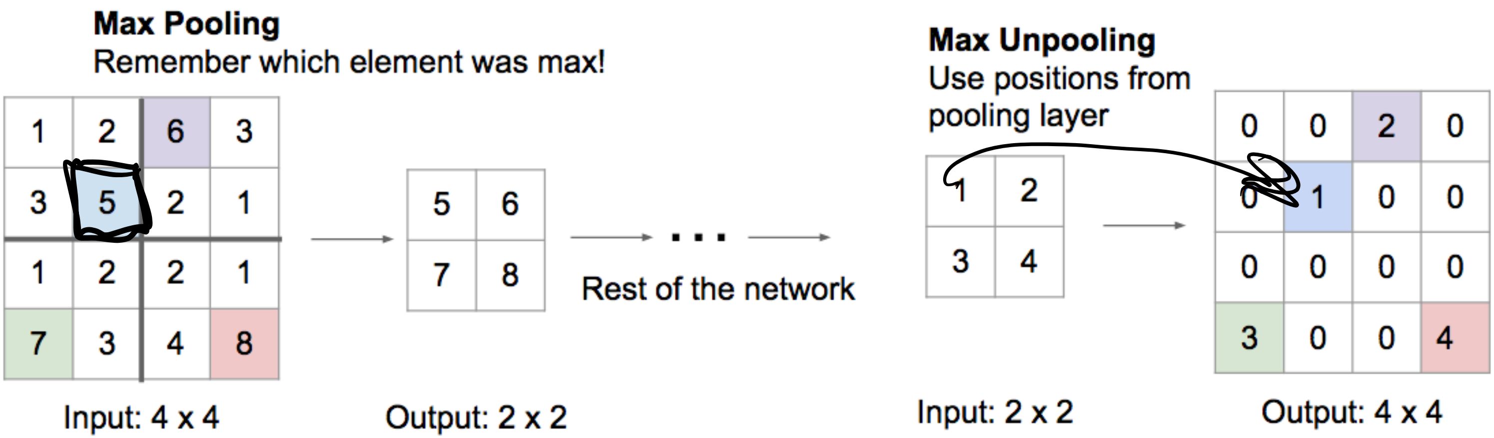
Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

In-Network upsampling: “Unpooling”

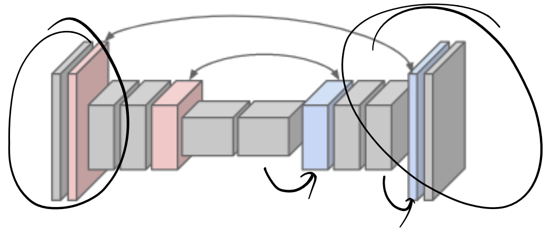


Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

In-Network upsampling: “Max Unpooling”



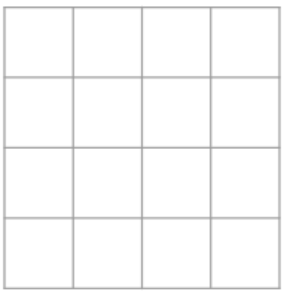
Corresponding pairs of downsampling and upsampling layers



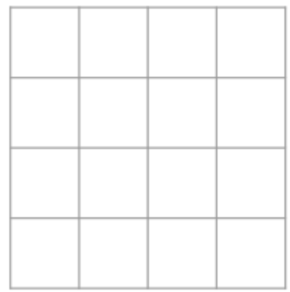
Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

Recall: Typical 3 x 3 convolution, stride 1 pad 1



Input: 4 x 4

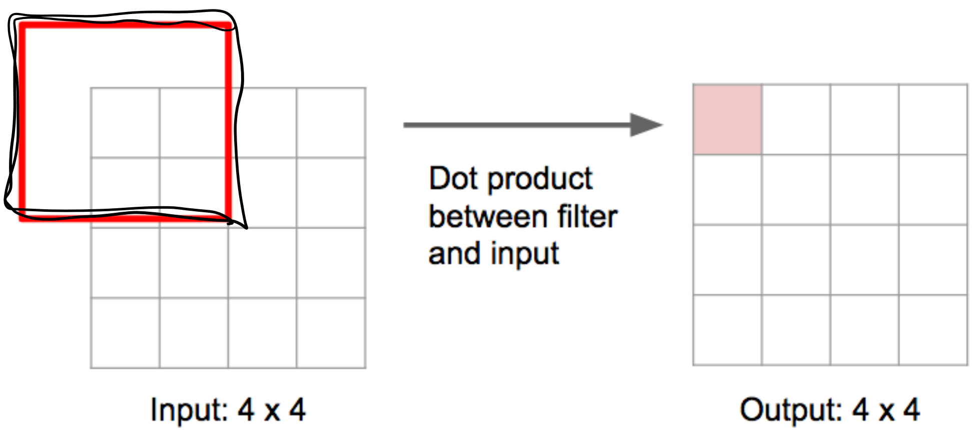


Output: 4 x 4

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

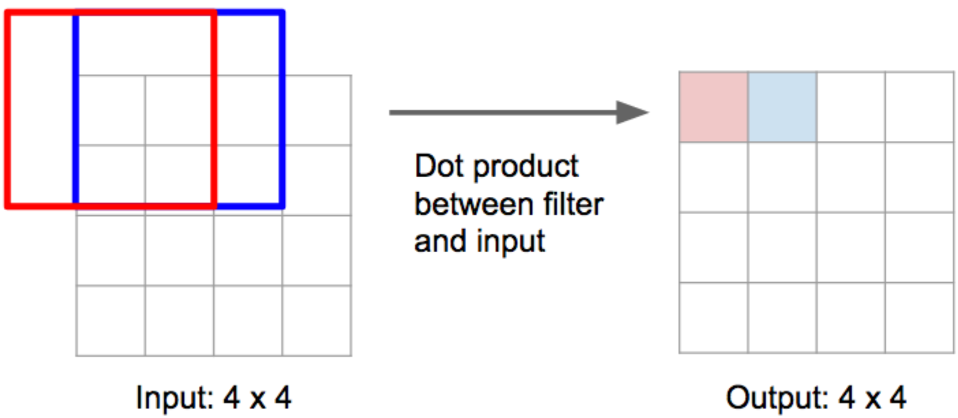
Recall: Normal 3 x 3 convolution, stride 1 pad 1



Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

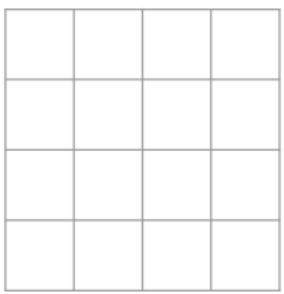
Recall: Normal 3 x 3 convolution, stride 1 pad 1



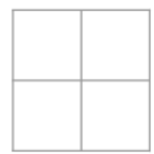
Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 2 pad 1



Input: 4 x 4

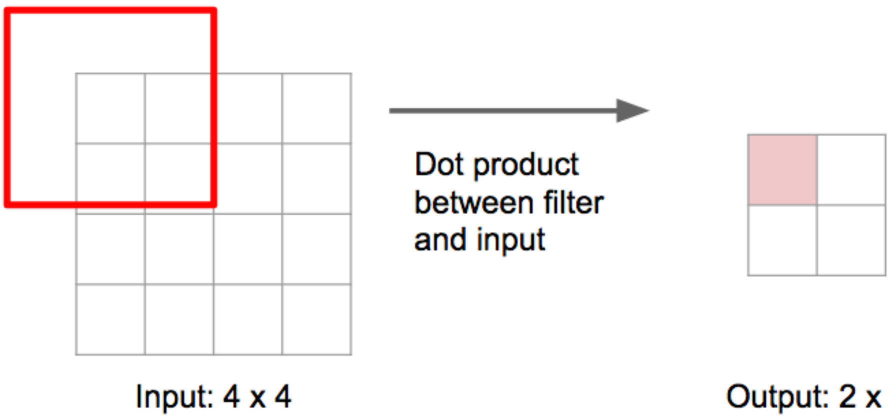


Output: 2 x 2

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

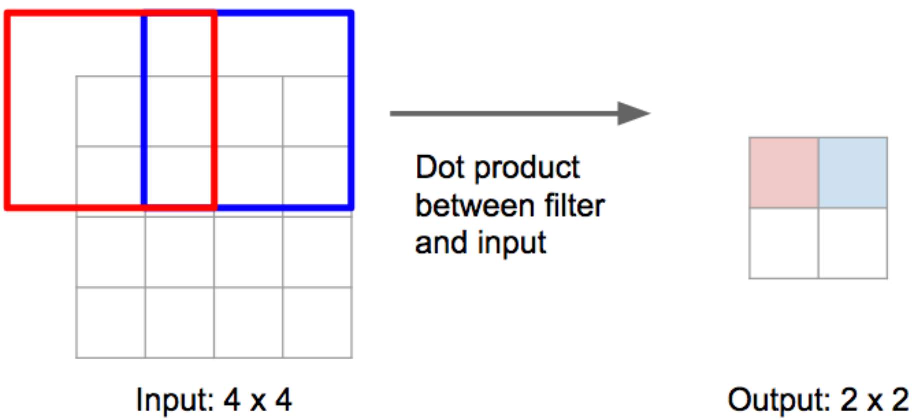
Recall: Normal 3 x 3 convolution, <u>stride 2</u> pad 1



Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 2 pad 1



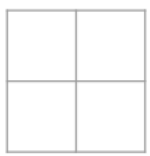
Filter moves 2 pixels in the input for every one pixel in the output

Stride gives ratio between movement in input and output

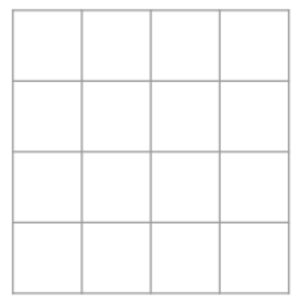
Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

3 x 3 **transpose** convolution, stride 2 pad 1



Input: 2 x 2

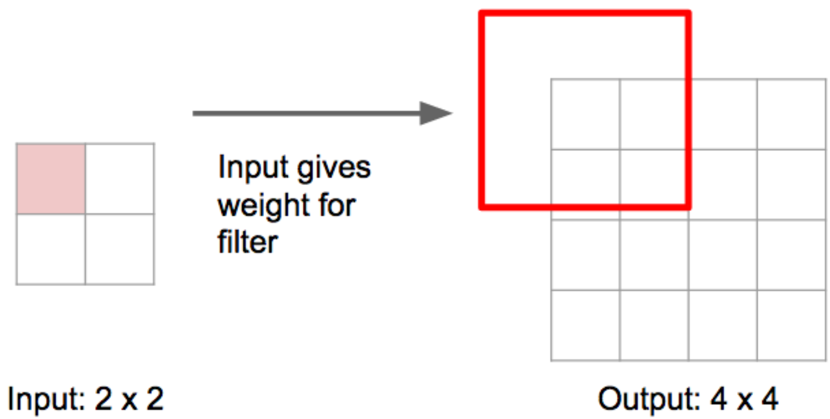


Output: 4 x 4

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

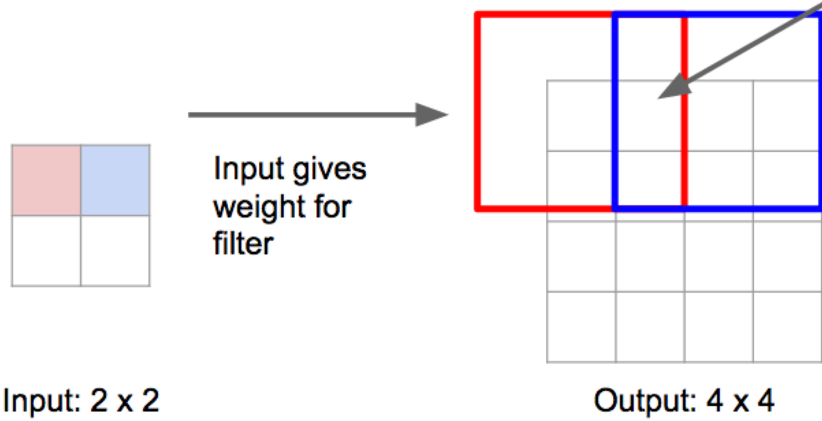
3 x 3 **transpose** convolution, stride 2 pad 1



Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: Transpose Convolution

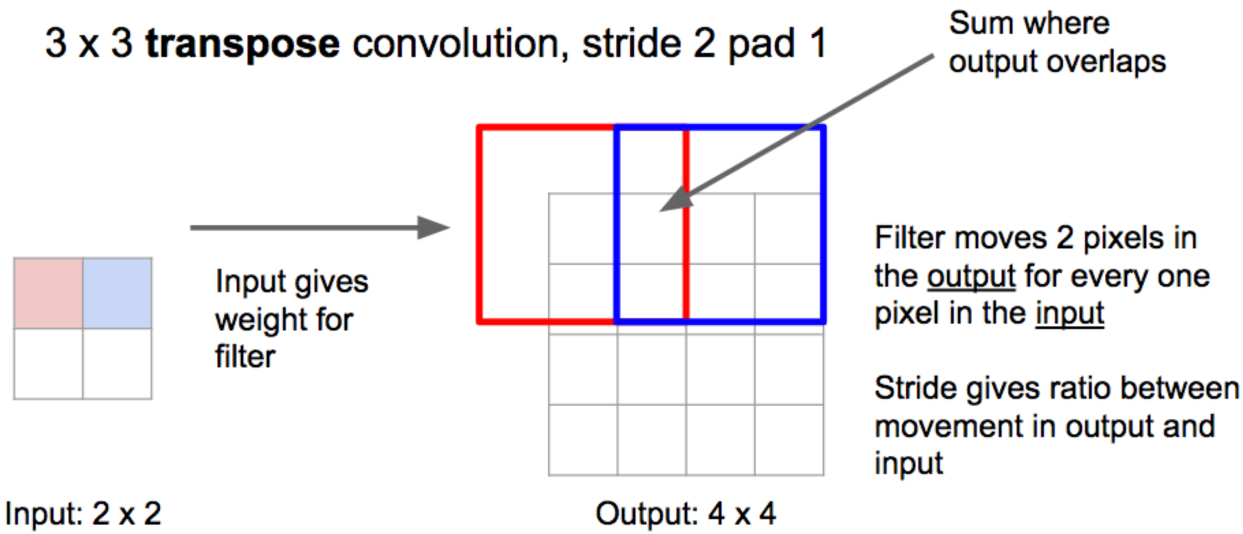
3 x 3 **transpose** convolution, stride 2 pad 1



Filter moves 2 pixels in the <u>output</u> for every one pixel in the <u>input</u>

Stride gives ratio between movement in output and input

Learnable Upsampling: Transpose Convolution

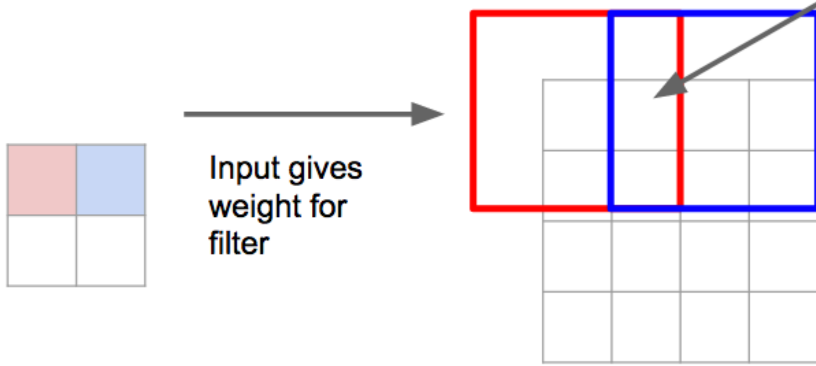


Learnable Upsampling: Transpose Convolution

Other names:
-Deconvolution (bad)
-Upconvolution
-Fractionally strided convolution
-Backward strided convolution

3 x 3 **transpose** convolution, stride 2 pad 1

Sum where output overlaps



Input: 2 x 2

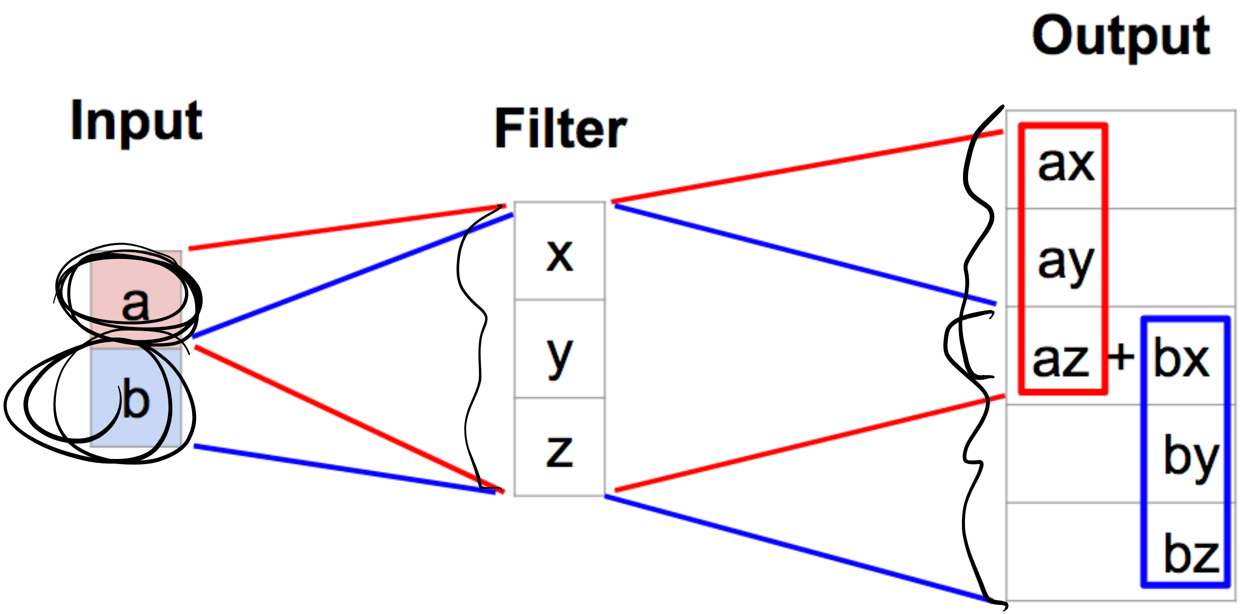
Output: 4 x 4

Filter moves 2 pixels in the <u>output</u> for every one pixel in the <u>input</u>

Stride gives ratio between movement in output and input

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Learnable Upsampling: 1D Example



Output contains copies of the filter weighted by the input, summing at where at overlaps in the output

Need to crop one pixel from output to make output exactly 2x input

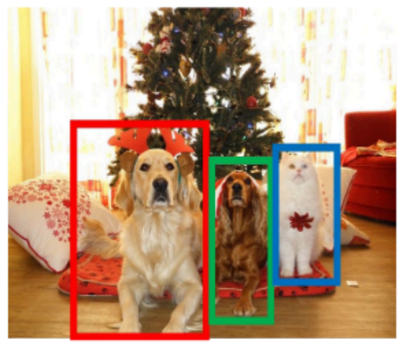
2D Object Detection

Semantic Segmentation

GRASS, CAT, TREE, SKY

No objects, just pixels

2D Object Detection



DOG, DOG, CAT

Object categories + 2D bounding boxes

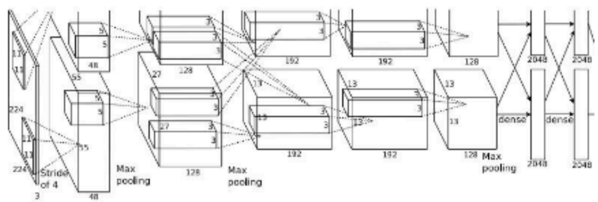
3D Object Detection

Car

Object categories + 3D bounding boxes

Object Detection as Classification: Sliding Window

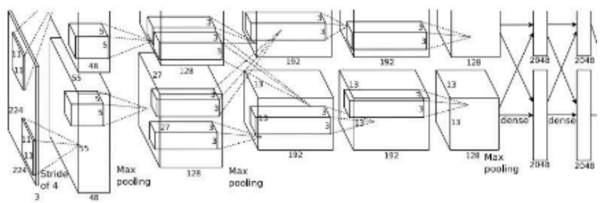
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? NO
Background? YES

Object Detection as Classification: Sliding Window

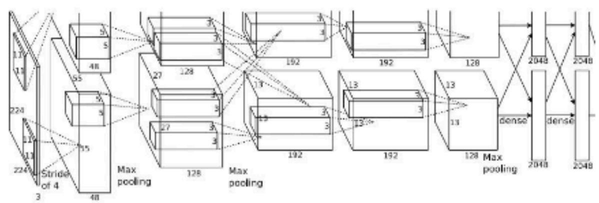
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

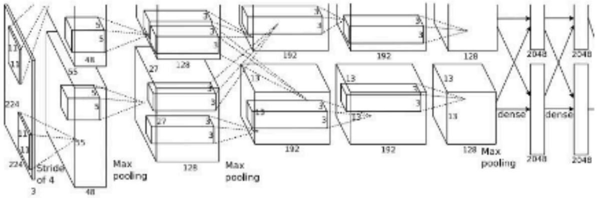


Dog? YES
Cat? NO
Background? NO

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Object Detection as Classification: Sliding Window

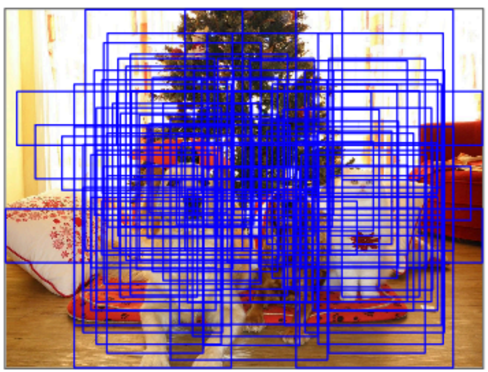
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

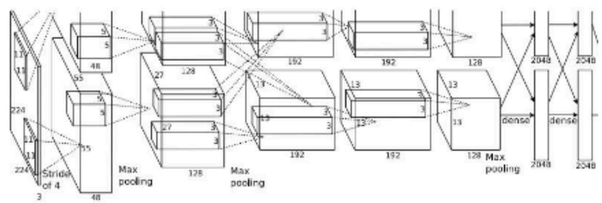


Dog? NO
Cat? YES
Background? NO

Object Detection as Classification: Sliding Window

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



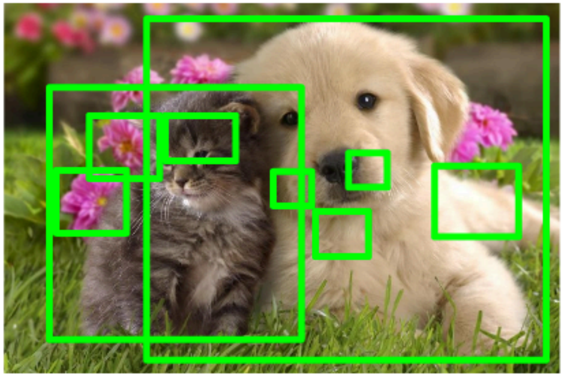


Dog? NO
Cat? YES
Background? NO

Problem: Need to apply CNN to huge number of locations, scales, and aspect ratios, very computationally expensive!

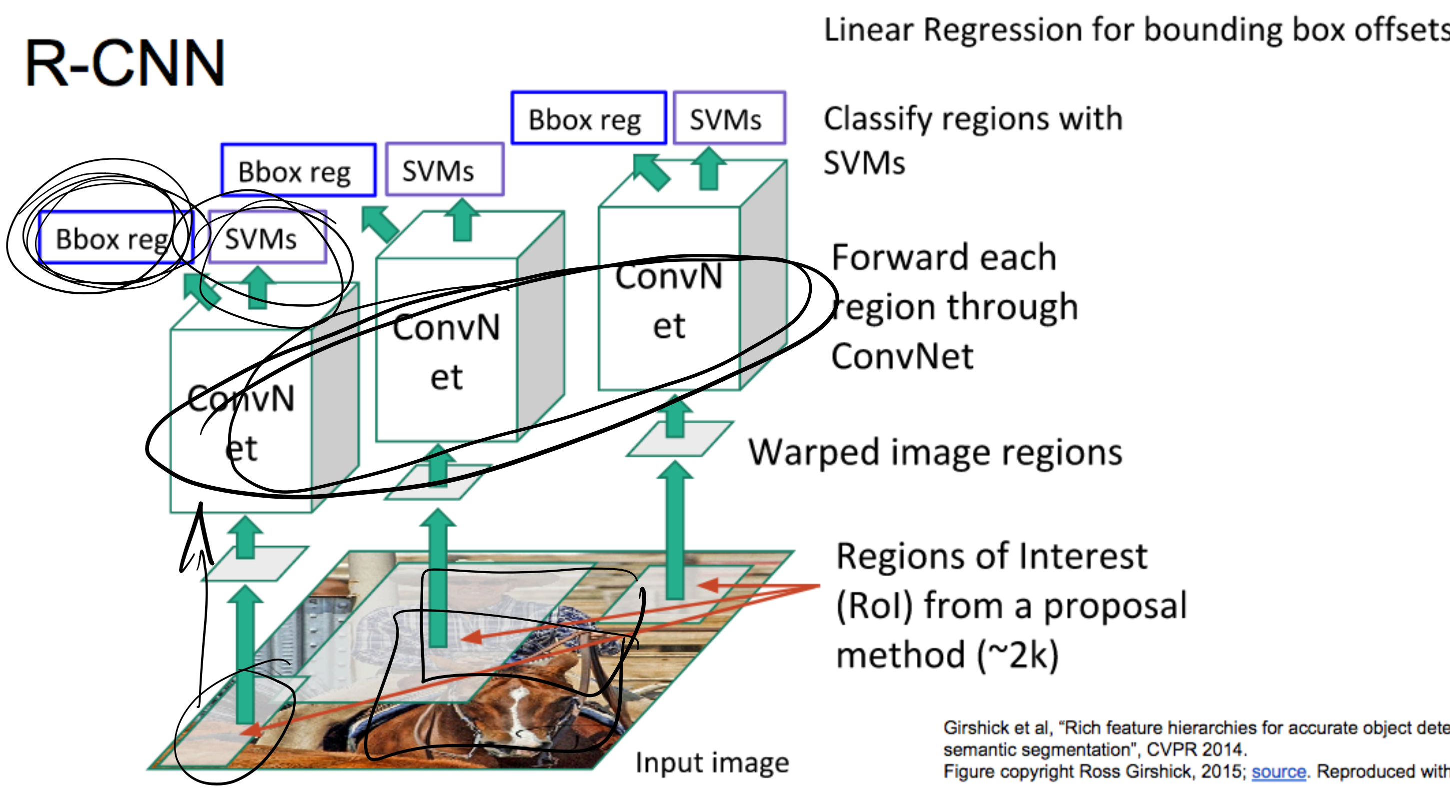
Region Proposals / Selective Search

- Find "blobby" image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 2000 region proposals in a few seconds on CPU



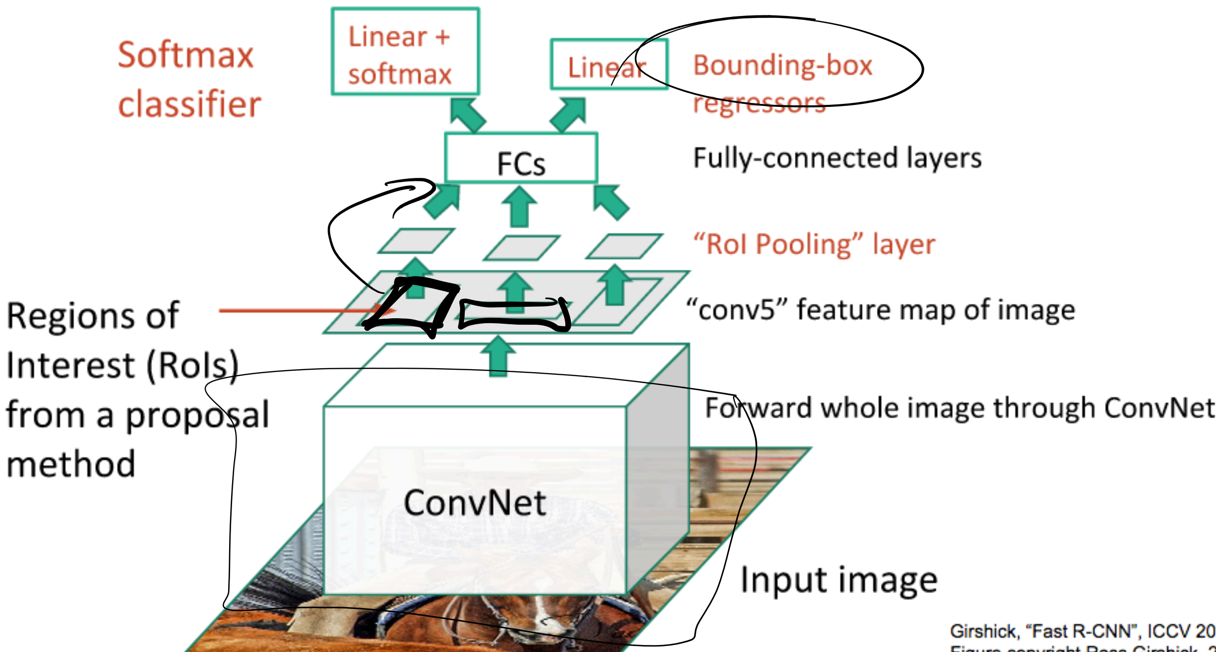
Alexe et al, "Measuring the objectness of image windows", TPAMI 2012
Uijlings et al, "Selective Search for Object Recognition", IJCV 2013
Cheng et al, "BING: Binarized normed gradients for objectness estimation at 300fps", CVPR 2014
Zitnick and Dollar, "Edge boxes: Locating object proposals from edges", ECCV 2014

R-CNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

Fast R-CNN



Girshick, "Fast R-CNN", ICCV 2015.

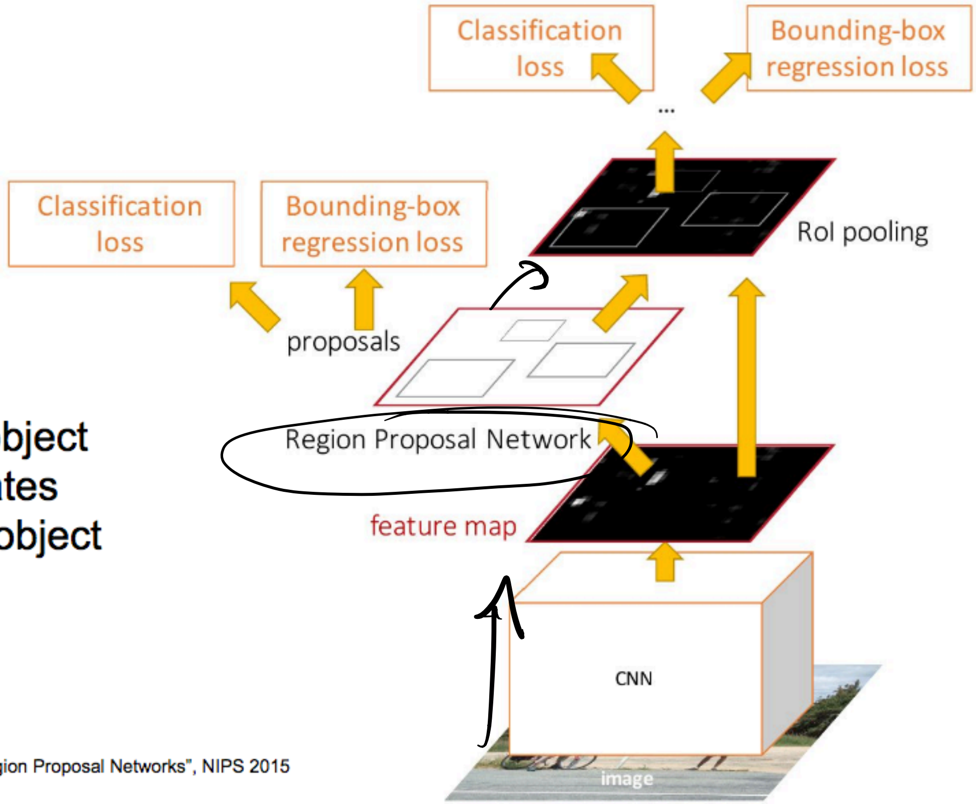
Faster R-CNN:

Make CNN do proposals!

Insert **Region Proposal Network (RPN)** to predict proposals from features

Jointly train with 4 losses:

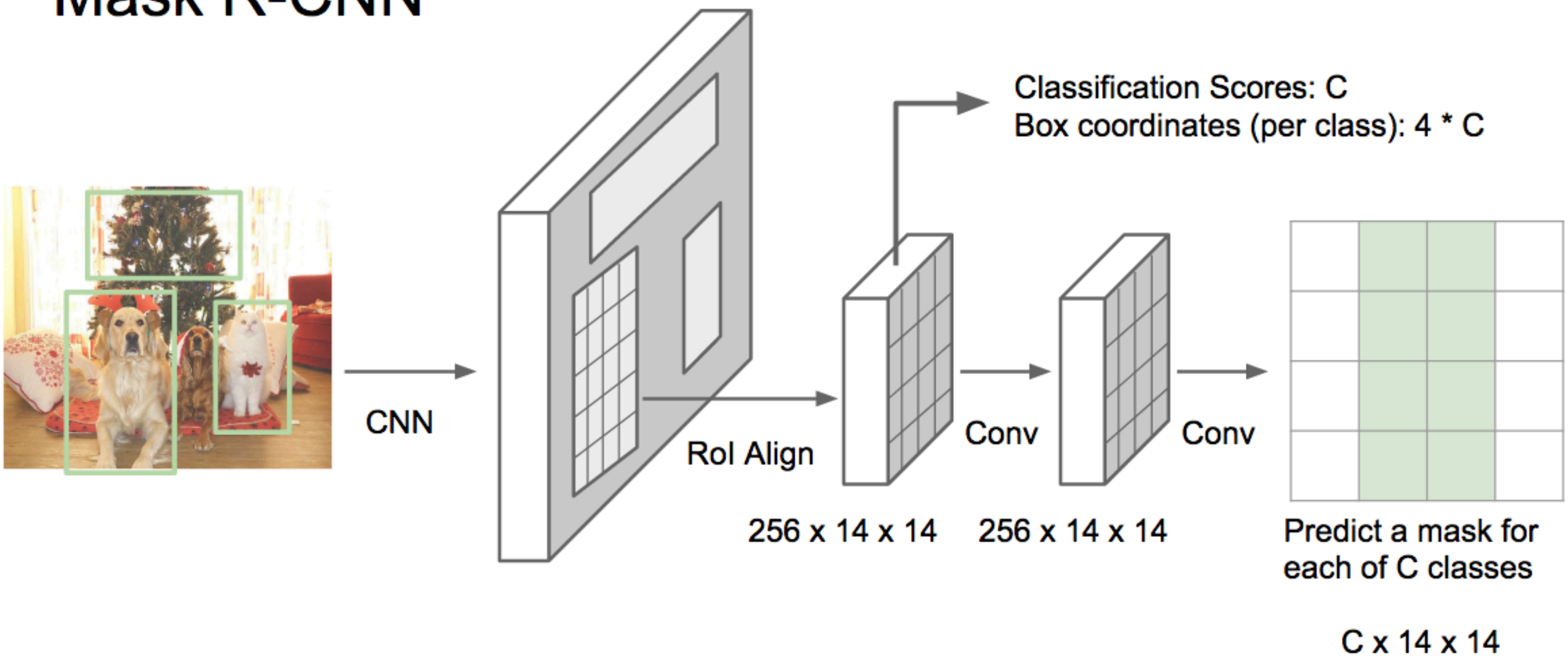
1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

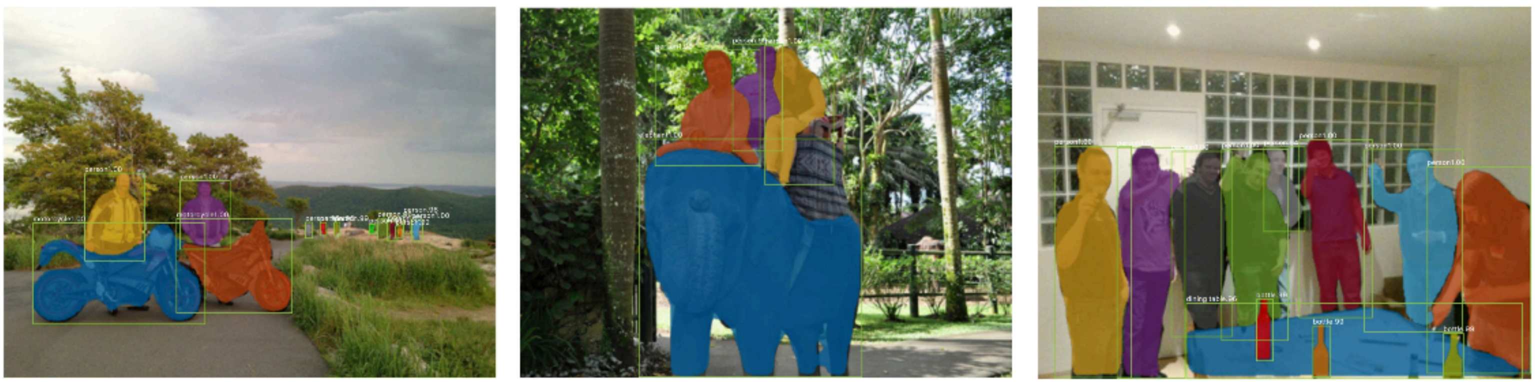
Mask R-CNN



He et al, "Mask R-CNN", arXiv 2017

Slide: Fei-Fei Li, Justin Johnson, & Serena Yeung

Mask R-CNN: Very Good Results!



He et al, "Mask R-CNN", arXiv 2017

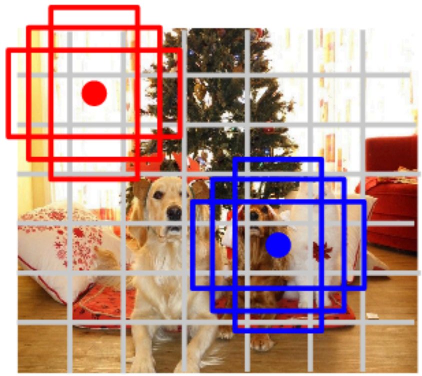
You Only Look Once

Single Shot Detection

Detection without Proposals: YOLO / SSD

Go from input image to tensor of scores with one big convolutional network!

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

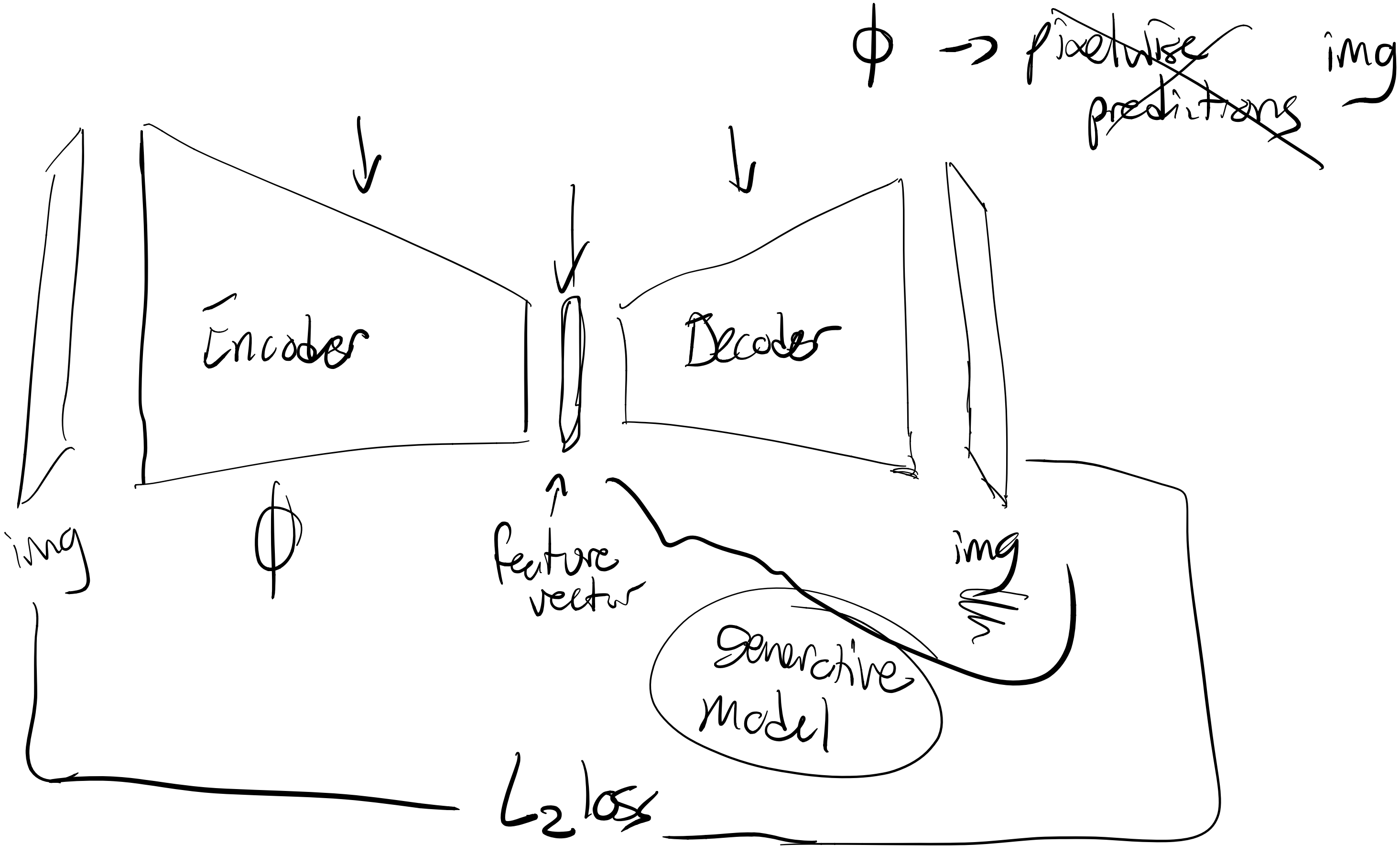
Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

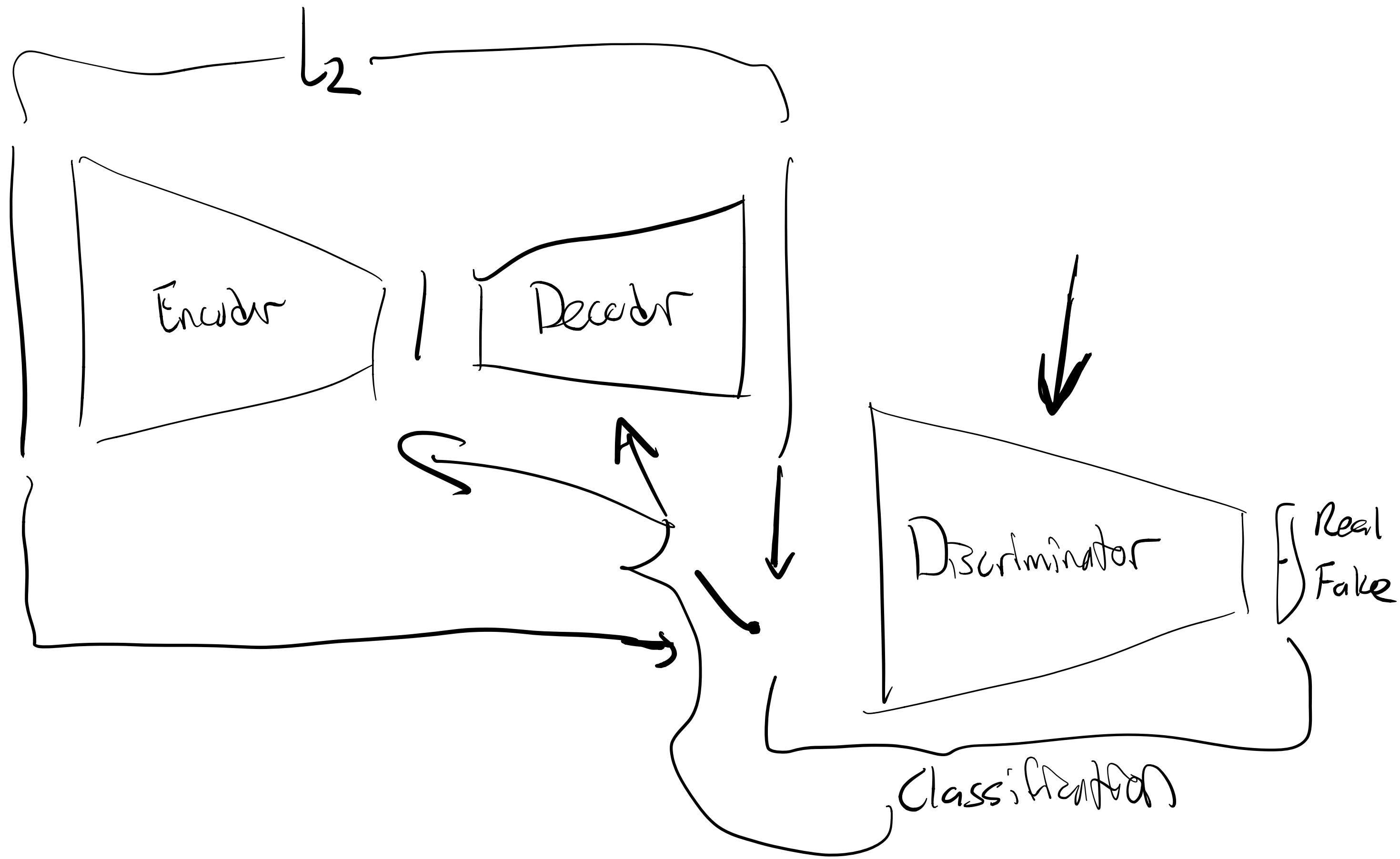
Output:
7 x 7 x (5 * B + C)

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016

Autoencoders and Generative Models



Generative Adversarial Networks



Other Problems

- Fine-grained recognition (e.g., dog/bird species)
- Instance segmentation
- Face detection and recognition
- Motion estimation
- Feature detection and description
- Depth estimation
- Novel view synthesis
- ...and many others