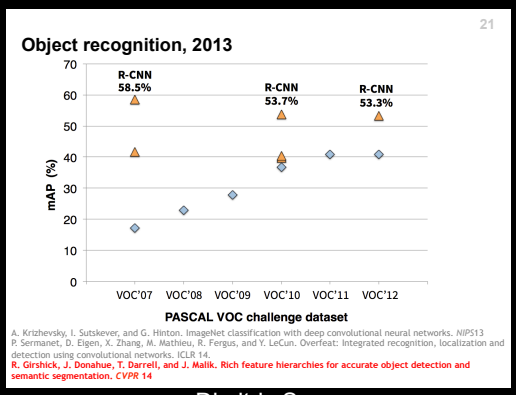
Introduction to Deep Learning

State of the Art CNN's in Computer Vision



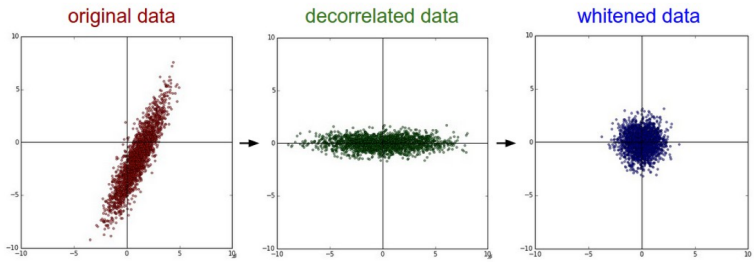
Dimitris Samaras
Stony Brook University

2

How to train Neural Networks

Step 1: Preprocess the data

In practice, you may also see **PCA** and **Whitening** of the data



Slide credit:Fei-Fei Li

3

How to train Neural Networks

- **Step 2: Choose the architecture:**
  - How many Layers? How many nodes?
  - Conv or fully connected
- **Step 3: Initialize well**
  - set weights to small random numbers
  - set biases to zero
  - Double check that the loss is reasonable: (by trying different reg. values

Slide credit:Fei-Fei Li

4

How to train Neural Networks

- **Step 4: Let's try to train:**
  - start with small regularization and find learning rate that makes the loss go down.
  - **loss not going down:**
  - learning rate too low
  - **loss exploding:**
  - learning rate too high
- **Step 5: Cross-validation strategy:**
  - **coarse -> fine** cross-validation in stages
  - **First stage**: only a few epochs to get rough idea of what params work
    Second stage: longer running time, finer search
    ... (repeat as necessary)
  - Tip for detecting explosions in the solver:
    If the cost is ever > 3 * original cost, break out early

Slide credit:Fei-Fei Li

5

How to train Neural Networks

Normally you can't afford a huge computational budget for expensive cross-validations.
Need to rely more on intuitions and visualizations...

Visualizations to play with:

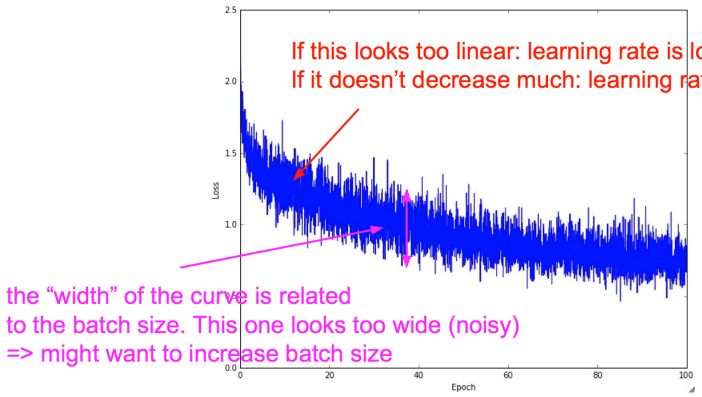
- **loss** function
- validation and training **accuracy**
- min,max,std for **values and updates**, (and monitor their ratio)
- **first-layer visualization** of weights (if working with images)

Seemingly unrelated: **Model Ensembles**
- One way to *always* improve final accuracy:
take several trained models and average their predictions

Slide credit:Fei-Fei Li

6

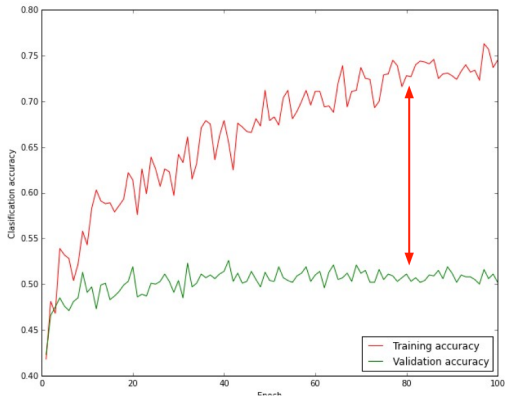
Monitor and visualize the loss curve



Slide credit:Fei-Fei Li

7

Monitor and visualize the accuracy:



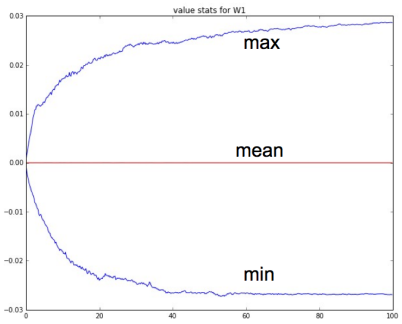
big gap = overfitting
=> increase regularization strength

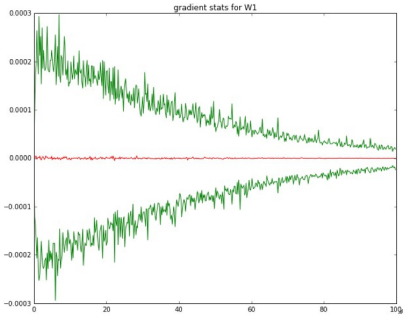
no gap
=> increase model capacity

Slide credit:Fei-Fei Li

8

Track the ratio of weight updates / weight magnitudes:



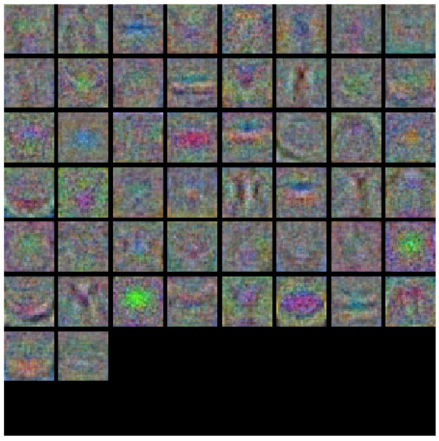


ratio between the values and updates: ~ 0.0002 / 0.02 = 0.01 (about okay)
want this to be somewhere around 0.01 - 0.001 or so

Slide credit:Fei-Fei Li

Visualizing first-layer weights:

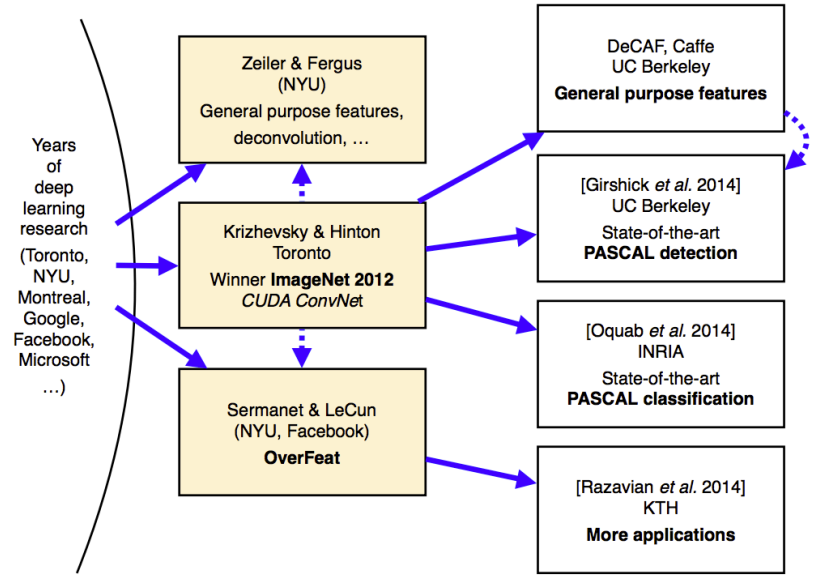
Noisy weights => Regularization maybe not strong enough



Slide credit:Fei-Fei Li

11

CNNs in vision

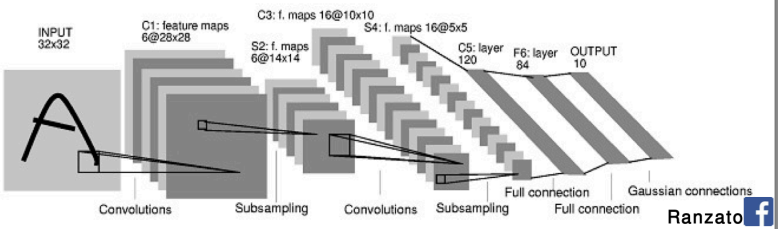


Slide: A. Vedaldi

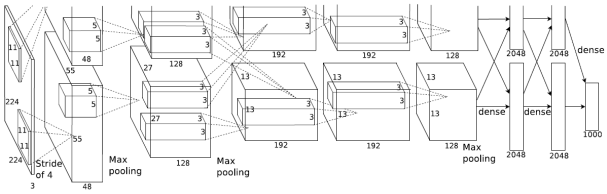
12

Famous CNNs

http://yann.lecun.com/exdb/lenet/



Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, november 1998



A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. *NIPS*13

13

Latest & greatest CNNs (deeper and deeper)

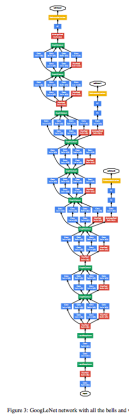


Figure 3: GoogLeNet network with all the bells and whistles

GoogLeNet
C. Szegedy et al.
Going deeper with convolutions
arXiv technical report, 2014

Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations incre from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). convolutional layer parameters are denoted as "conv(receptive field size)-(number of channe The ReLU activation function is not shown for brevity.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

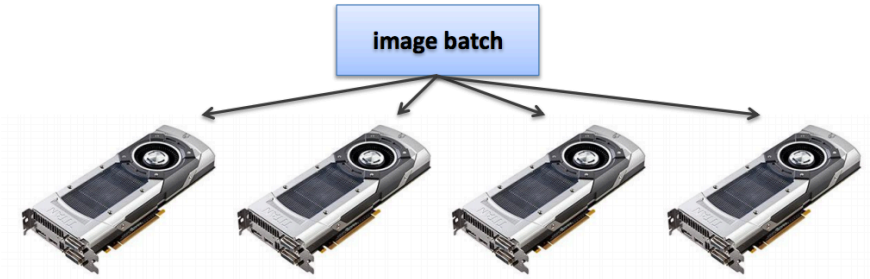
Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

VGG net (Oxford):
K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv, 2014

14

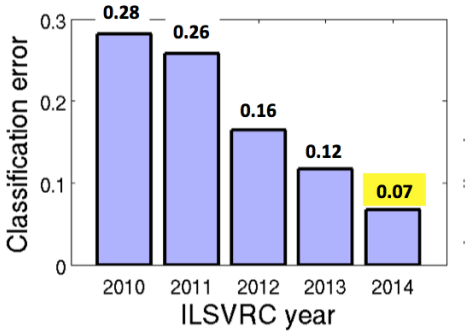
Really large CNNs
From VGG's network presentation:

- Heavily-modified Caffe C++ toolbox
- Multiple GPU support
  - 4 x NVIDIA Titan, off-the-shelf workstation
  - data parallelism for training and testing
  - ~3.75 times speed-up, 2-3 weeks for training



16

Imagenet top-5 error rate: 36%-> 18% (2012) -> 6% (2014)



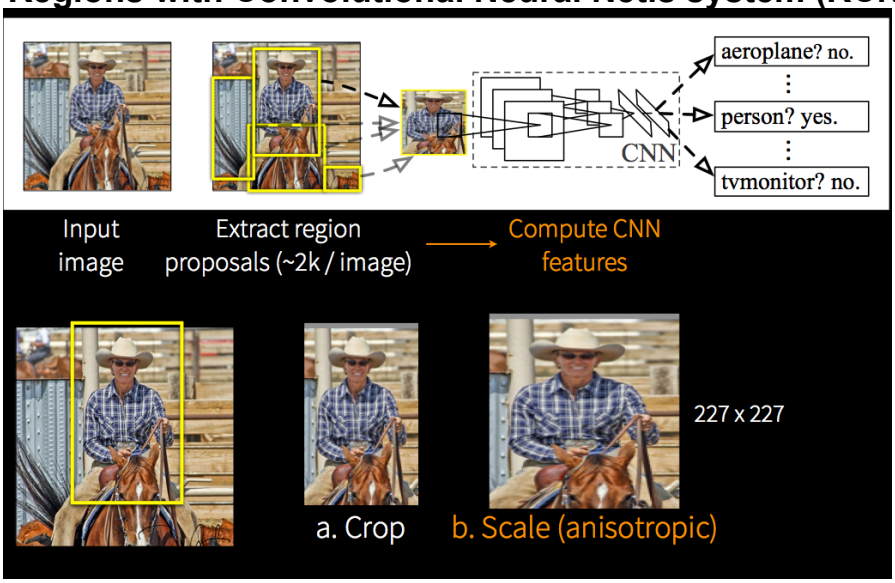
1.7x reduction in classification error since last year

4.2x reduction in classification error since 2010

http://www.image-net.org/challenges/LSVRC/2014/eccv2014

24

Regions with Convolutional Neural Net.s system (RCNN)



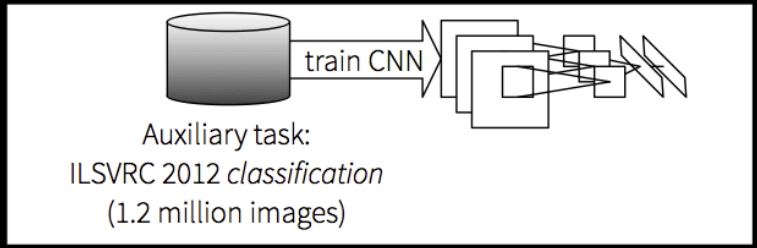
R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

25

Regions with Convolutional Neural Net.s system (RCNN)

Supervised pre-training

Train a SuperVision CNN* for the 1000-way ILSVRC image classification task

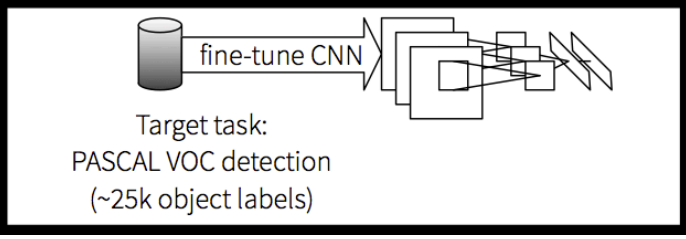


R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

Regions with Convolutional Neural Net.s system (RCNN)

Fine-tune the CNN for detection

Transfer the representation learned for ILSVRC classification to PASCAL (or ImageNet detection)



R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

Regions with Convolutional Neural Net.s system (RCNN)

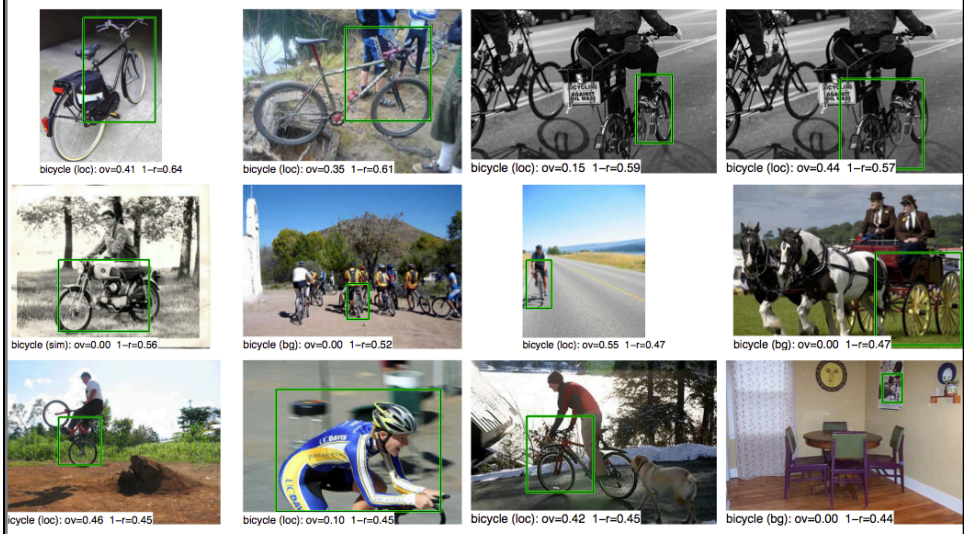
Train detection SVMs

(With the softmax classifier from fine-tuning mAP decreases from 54% to 51%)

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

28

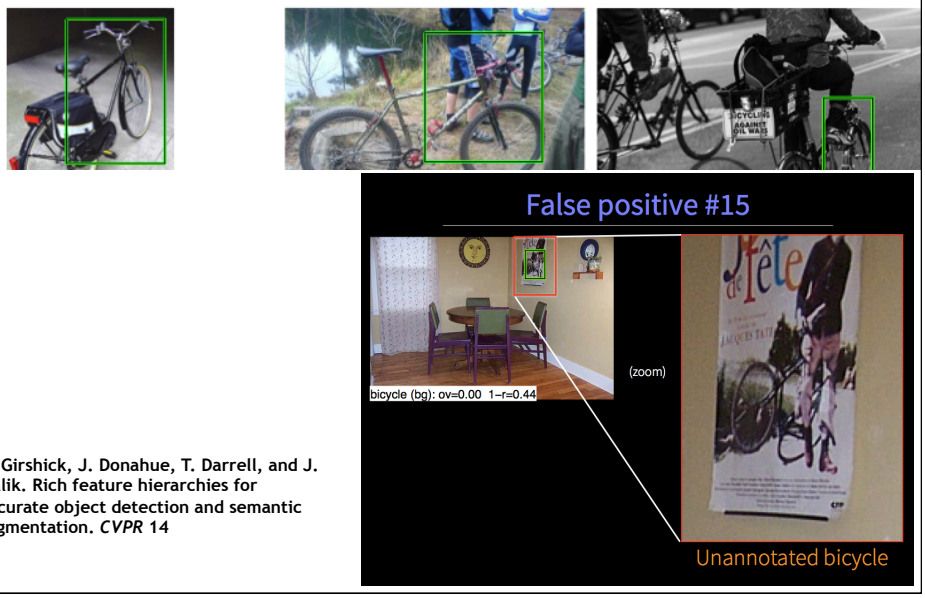
Detection results



R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

29

Failures: mostly localization errors

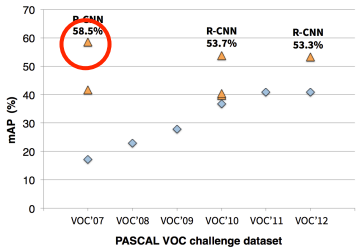


R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

Detection results

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN $pool_5$	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc_6	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc_7	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT $pool_5$	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
R-CNN FT fc_6	63.5	66.0	47.9	37.7	29.9	62.5	70.2	60.2	32.0	57.9	47.0	53.5	60.1	64.2	52.2	31.3	55.0	50.0	57.7	63.0	53.1
R-CNN FT fc_7	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN FT fc_7 BB	**68.1**	**72.8**	**56.8**	**43.0**	**36.8**	**66.3**	**74.2**	**67.6**	**34.4**	**63.5**	**54.5**	**61.2**	**69.1**	**68.6**	**58.7**	**33.4**	**62.9**	**51.1**	**62.5**	**64.8**	**58.5**
DPM v5 [20]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
DPM ST [28]	23.8	58.2	10.5	8.5	27.1	50.4	52.0	7.3	19.2	22.8	18.1	8.0	55.9	44.8	32.4	13.3	15.9	22.8	46.2	44.9	29.1
DPM HSC [31]	32.2	58.3	11.5	16.3	30.6	49.9	54.8	23.5	21.5	27.7	34.0	13.7	58.1	51.6	39.9	12.4	23.5	34.4	47.4	45.2	34.3

Table 2: Detection average precision (%) on VOC 2007 test. Rows 1-3 show R-CNN performance without fine-tuning. Rows 4-6 show results for the CNN pre-trained on ILSVRC 2012 and then fine-tuned (FT) on VOC 2007 trainval. Row 7 includes a simple bounding-box regression (BB) stage that reduces localization errors (Section C). Rows 8-10 present DPM methods as a strong baseline. The first uses only HOG, while the next two use different feature learning approaches to augment or replace HOG.



R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

Detection results

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN T-Net	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN T-Net BB	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5
R-CNN O-Net	71.6	73.5	58.1	42.2	39.4	70.7	76.0	74.5	38.7	71.0	56.9	74.5	67.9	69.6	59.3	**35.7**	62.1	64.0	66.5	**71.2**	62.2
R-CNN O-Net BB	**73.4**	**77.0**	**63.4**	**45.4**	**44.6**	**75.1**	**78.1**	**79.8**	**40.5**	**73.7**	**62.2**	**79.4**	**78.1**	**73.1**	**64.2**	35.6	**66.8**	**67.2**	**70.4**	71.1	**66.0**

Table 3: Detection average precision (%) on VOC 2007 test for two different CNN architectures. The first two rows are results from Table 2 using Krizhevsky et al.'s architecture (T-Net). Rows three and four use the recently proposed 16-layer architecture from Simonyan and Zisserman (O-Net) [43].

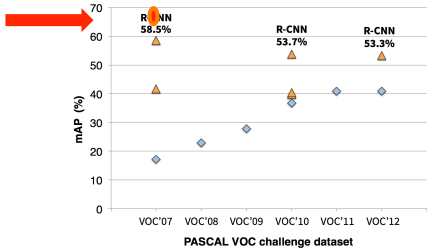
Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations incr from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channe The ReLU activation function is not shown for brevity.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

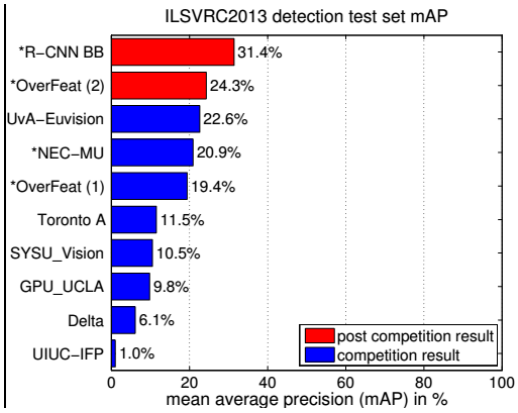
VGG net: K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv, 2014



R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR* 14

32

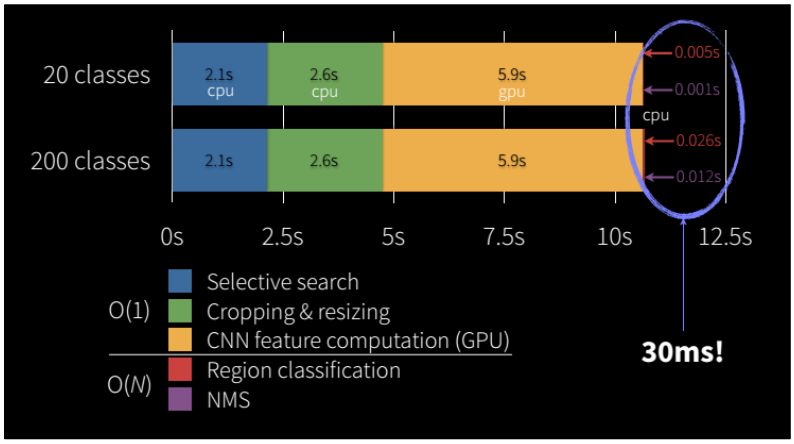
R-CNN: ILSVRC 2013 performance



33

R-CNN speed and

- R-CNN detection time/frame



34

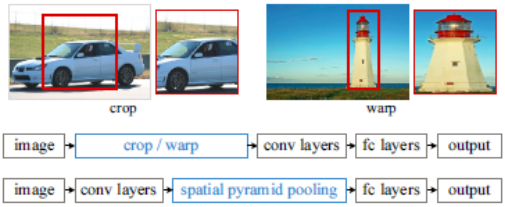
SPP-net = CNN + SPP

Kaiming He et al, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

"Classical" conv. NN" requires a fixed-size (e.g. 224224) input image:

- Need cropping or warping to transform original image to square shape
- This constraint is related to Fully-Connected layer ONLY

Idea: let's use Spatial Pooling Pyramid to transform any-shape image to 'fixed-length" feature vector.



http://research.microsoft.com/en-us/um/people/kahe/

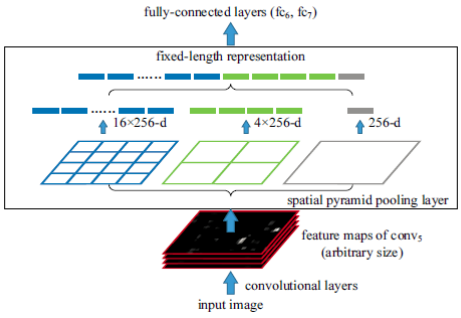
35

Spatial Pyramid Pooling

[pool3x3]
type=pool
pool=max
inputs=conv5
sizeX=5
stride=4

[pool2x2]
type=pool
pool=max
inputs=conv5
sizeX=7
stride=6

[pool1x1]
type=pool
pool=max
inputs=conv5
sizeX=13
stride=13



36

SPP-net training

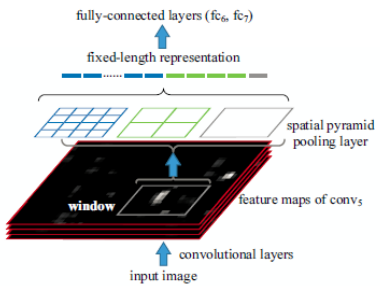
- Size augmentation:
  - Imagenet: 224x224 → 180x180
  - Horizontal flipping
  - Color altering
- Dropout with 2 last FC layers
- Learning rate:
  - Init lr= 0.01; divide by 10 when error plateau

37

SPP: Imagenet - Detection

1. Find 2000 windows candidate */~ R-CNN /*
2. extract the feature maps from the entire image only once (possibly at multiple scales) /~ Overfeat/.
3. Then apply the spatial pyramid pooling on each candidate window of the feature, which maps window to a fixed-length representation
4. Then 2 FC layers
5. SVM

~170x faster than R-CNN



38

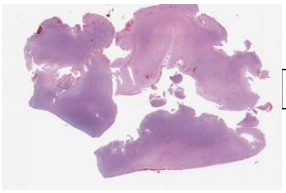
SPP-net: Imagenet classification

	method	test scale	test views	top-1 val	top-5 val
(a)	Krizhevsky *et al.* [16]	1	10	40.7	18.2
(b1)	Overfeat (fast) [24]	1	-	39.01	16.97
(b2)	Overfeat (fast) [24]	6	-	38.12	16.27
(b3)	Overfeat (big) [24]	4	-	35.74	14.18
(c1)	Howard (base) [15]	3	162	37.0	15.8
(c2)	Howard (high-res) [15]	3	162	36.8	16.2
(d1)	Zeiler & Fergus (ZF) (fast) [33]	1	10	38.4	16.5
(d2)	Zeiler & Fergus (ZF) (big) [33]	1	10	37.5	16.0
(e1)	our impl of ZF (fast)	1	10	35.99	14.76
(e2)	SPP-net$_4$, single-size trained	1	10	35.06	14.04
(e3)	SPP-net$_6$, single-size trained	1	10	34.98	14.14
(e4)	SPP-net$_6$, multi-size trained	1	10	34.60	13.64
(e5)	SPP-net$_6$, multi-size trained	1	8+2full	**34.16**	**13.57**

30

Automatic Glioma Classification

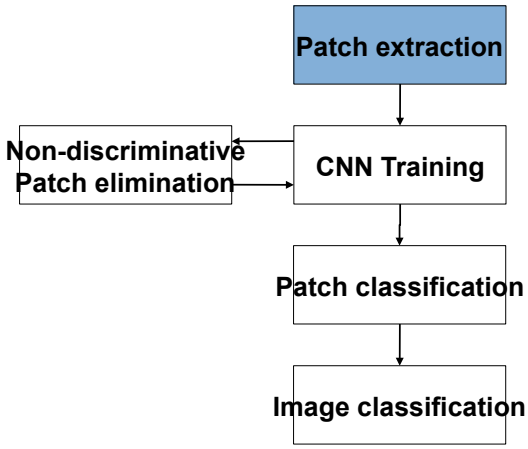
- Gliomas are the most common brain cancers.
- Better classification is critical to the development of targeted therapies.
- Microscopy images of tissue slides provide reach information.
- Glioma Tissue Example: http://cancer.digitalslidearchive.net/ select GBM or LGG patients.



Glioblastoma?
Oligodendroglioma?
Astrocytoma?
Oligoastrocytoma?
Anaplastic Astro?
Anaplastic Oligo?

31

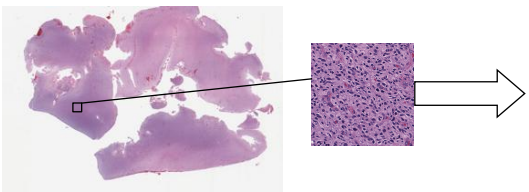
Pipeline summary



32

Too Large for CNNs

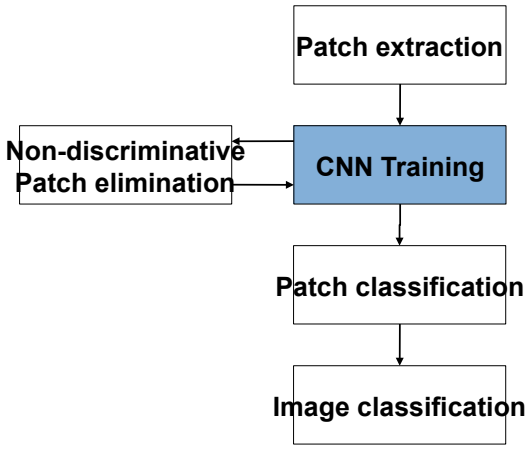
- The resolution of tissue images is really high (gigapixel), it is impossible to run CNNs on the whole images.
  - Pathologists cannot look at the whole image at the finest resolution at one time either. They scan through the image and make decisions based on regions.
  - The algorithm can do the same thing. A CNN is trained to classify image patches. Then the patch-level classification results are aggregated for image-level classification.



Glioblastoma?
Oligodendroglioma?
Astrocytoma?
Oligoastrocytoma?
Anaplastic Astro?

33

Pipeline summary



34

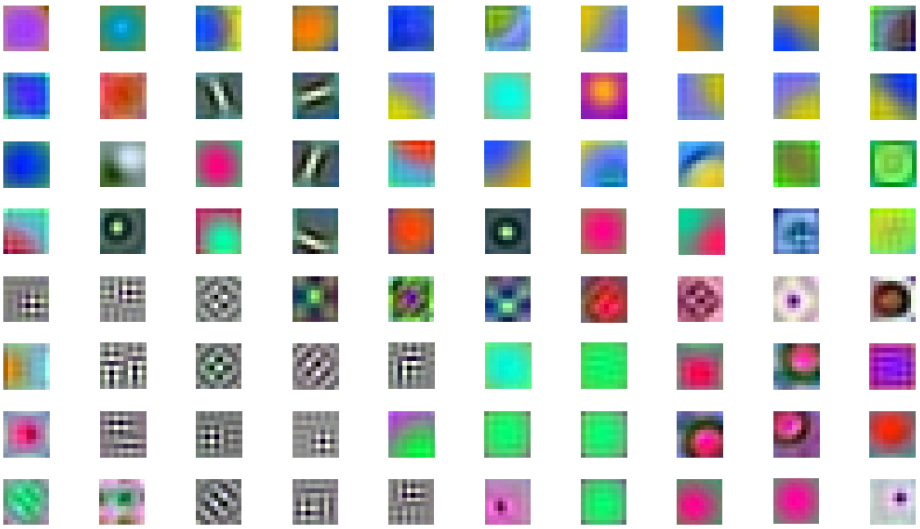
CNN Architecture

Layer	Filter size, stride	Output size
Input	-	$400 \times 400 \times 3$
Conv	10×10, 2	$196 \times 196 \times 80$
ReLU+LRN	-	$196 \times 196 \times 80$
Max-pool	6×6, 4	$49 \times 49 \times 80$
Conv	5×5, 1	$45 \times 45 \times 120$
ReLU+LRN	-	$45 \times 45 \times 120$
Max-pool	3×3, 2	$22 \times 22 \times 120$
Conv	3×3, 1	$20 \times 20 \times 160$
ReLU	-	$20 \times 20 \times 160$
Conv	3×3, 1	$18 \times 18 \times 200$
ReLU	-	$18 \times 18 \times 200$
Max-pool	3×3, 2	$9 \times 9 \times 200$
FC	-	320
ReLu+Drop	-	320
FC	-	320
ReLu+Drop	-	320
FC	-	6
Softmax	-	6

- **800,000 patches of size 500 x 500 are extracted from 1000 tissue images as input data.**
- **Patches are randomly rotated, flipped, cropped, and color-adjusted.**
- **A CNN with 4 convolutional layers are then applied on those patches.**
  - **ReLU: Rectified Linear Units.**
  - **LRN: Local Response Normalization.**
  - **FC: Fully Connected layer.**
  - **Drop: Dropout layer, probability = 0.5.**

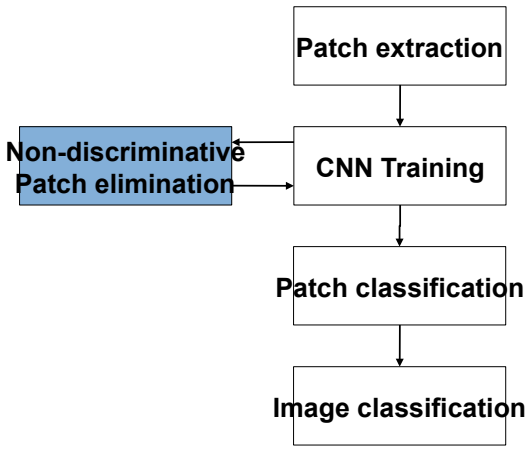
35

First Layer Filters Learnt



36

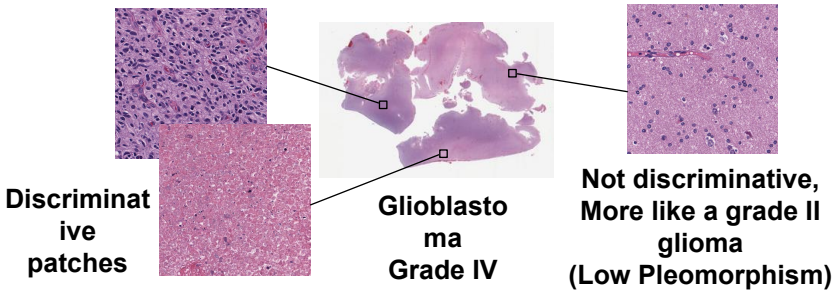
Pipeline summary



37

However, Patches May not be Discriminative

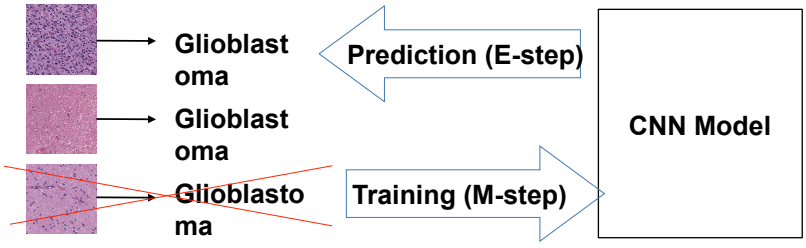
- A patch in an image of grade IV Glioblastoma may be a:
  - Low grade glioma tissue.
  - Healthy tissue.
- A patch in an image of mixed **Oligoastro**cytoma may be a:
  - **Oligo**dendroglioma or **Astro**cytoma tissue.
  - Healthy tissue.



38

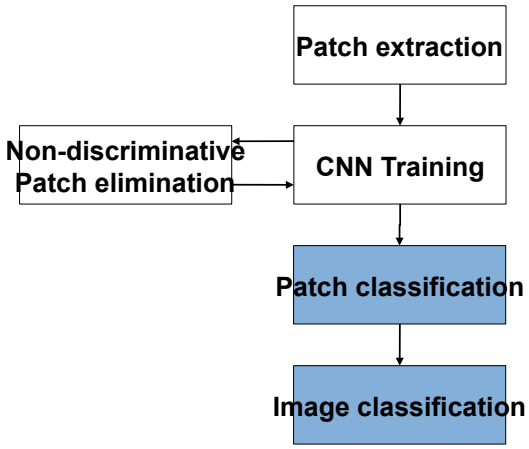
Non-discriminative patch elimination

- Non-discriminative patches are noise. They can be identified using Multiple Instance Learning techniques.
  - Modeling whether a patch is discriminative or not by a hidden variable.
  - Solve the hidden variables by EM.



39

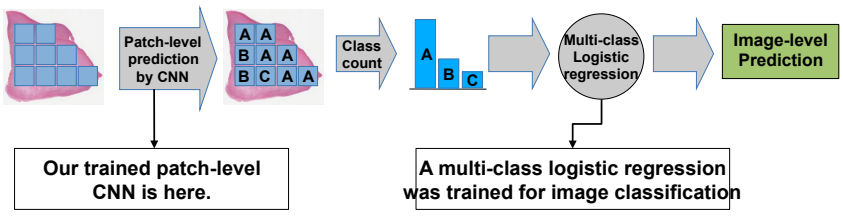
Pipeline summary



40

Given a New Image

- Extract patches (usually sparse sampling).
- Patch-level classification by CNN.
- Count the number of predicted patch-level classes.
- Image-level classification by multiclass logistic regression.



Glioma Image Classification Results

Methods	Accuracy
Morphology Features (Cooper etal. 2012) + SVM	0.629
Patch-level CNN + Voting	0.710
Multiple-instance CNN + LR	0.771
Inter-observer agreement (Coons etal. 1997)	0.7-0.8
Chance	0.513

Inter-observer agreement: the agreement between experienced pathologists on a similar dataset. The agreement increased from 0.7 to 0.8 after reviewing cases together.