CS698U: Topics in Computer Vision

Jan—May 2017

Lecture 5

Gaurav Sharma

Indian Institute of Technology Kanpur

www.grvsharma.com

Gradient Descent

$$\min_{\theta} J(\theta)$$

Objective function
= Empirical loss + Regularization

$$\theta = \theta - \eta \nabla_{\theta} J(\theta)$$

Move in the direction
of negative gradient

Remember: Data X is implicit in the objective/gradient

Batch Gradient Descent

$$\min_{\theta} J(\theta)$$

Objective function
= Empirical loss + Regularization

$$\theta = \theta - \eta \nabla_{\theta} J(\theta; X)$$

All the data is used

Might be making redundant computations for similar examples

Stochastic Gradient Descent

$$\min_{\theta} J(\theta)$$

Objective function
= Empirical loss + Regularization

$$\theta = \theta - \eta \nabla_{\theta} J(\theta; x_i, y_i)$$

Gradient with one example

Updates may have a large variance and the convergence would be very jittery and take a long time

Mini-Batch Gradient Descent

$$\min_{\theta} J(\theta)$$

Objective function
= Empirical loss + Regularization

$$\theta = \theta - \eta \nabla_\theta J(\theta; \{x_i, y_i\}_{i \in \mathcal{I}})$$

Gradient with a set of sampled examples

Compromise between batch and SGD,
Convergence may be smoother than stochastic

Variants of GD

- Remember a part of the previous direction
- t indexes the number of iterations
- d is direction vector

$$\theta = \theta - d_t$$

GD with Momentum

- Remember a part of the previous direction
- t indexes the number of iterations
- d is direction vector

$$d_t = \alpha d_{t-1} + \eta \nabla_\theta J(\theta)$$

$$\theta = \theta - d_t$$

Qian, Ning. "On the momentum term in gradient descent learning algorithms." Neural networks 12.1 (1999): 145-151.

GD with Nesterov Momentum

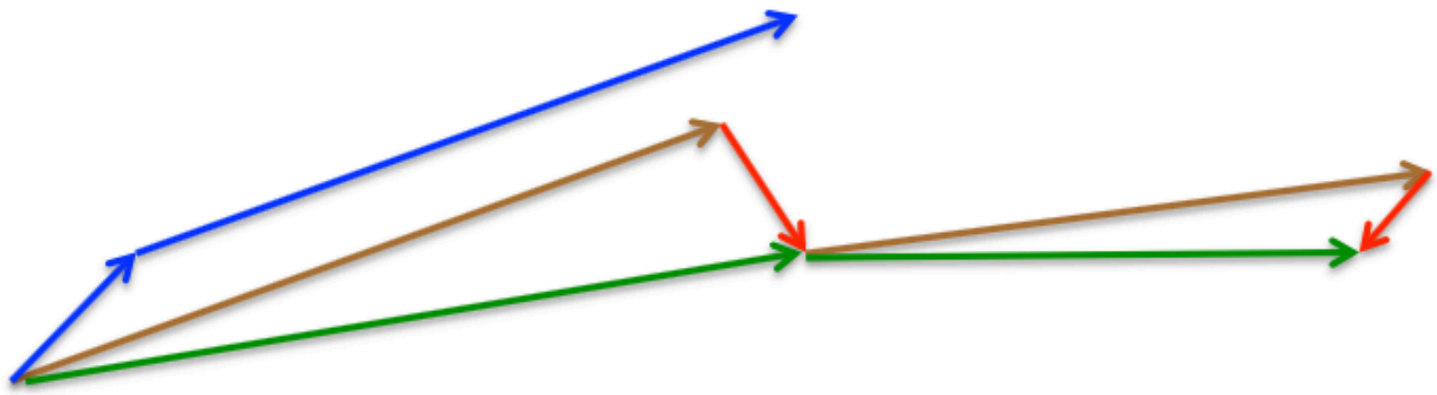
- Remember a part of the previous direction
- t indexes the number of iterations
- d is direction vector

$$d_t = \alpha d_{t-1} + \eta \nabla_\theta J(\theta - \alpha d_{t-1})$$

$$\theta = \theta - d_t$$

Nesterov, Yurii. "A method for unconstrained convex minimization problem with the rate of convergence O (1/k2)." Doklady an SSSR. Vol. 269. No. 3. 1983.

Std vs Nesterov Momentum



brown vector = jump, red vector = correction, green vector = accumulated gradient

blue vectors = standard momentum

- Momentum = (weighted) add previous accumulated gradient to current gradient and make a move
- Nesterov = make a move from previous accumulated gradient then correct based on current grad.

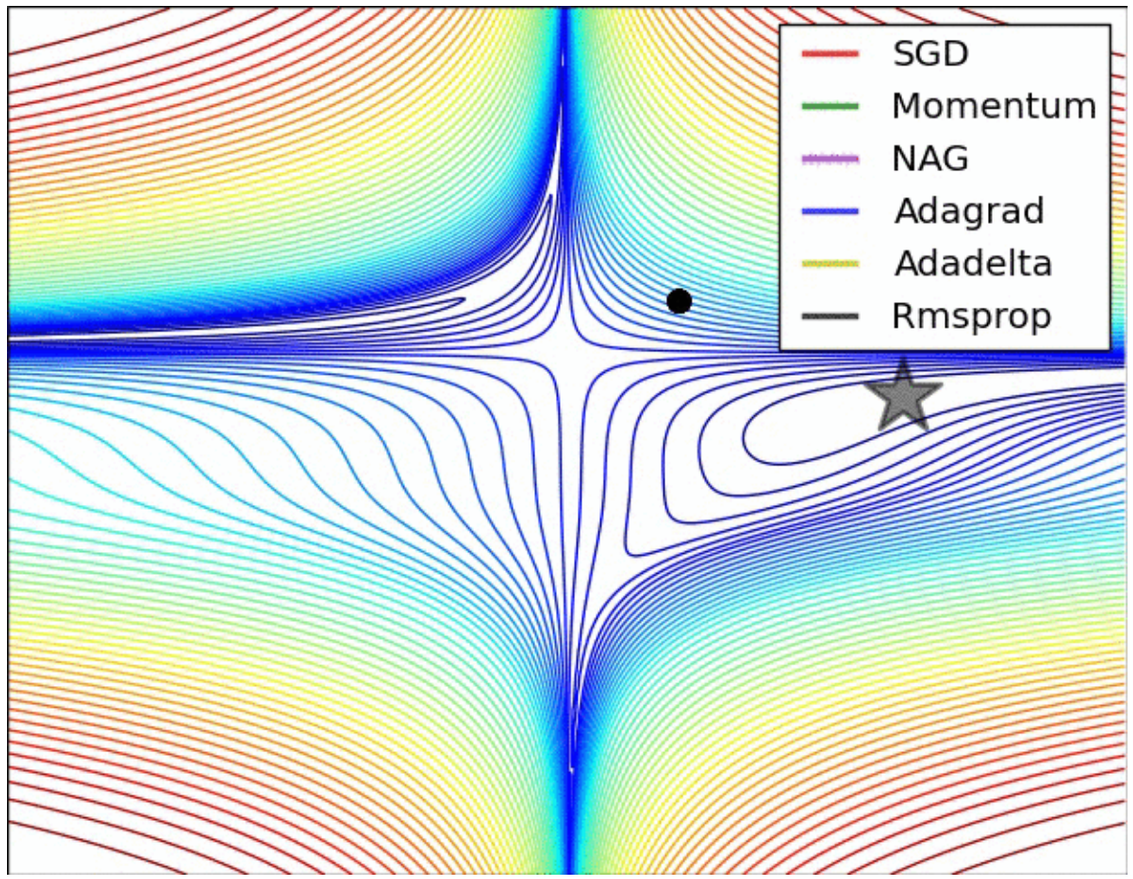
Image from G. Hinton's lecture at http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

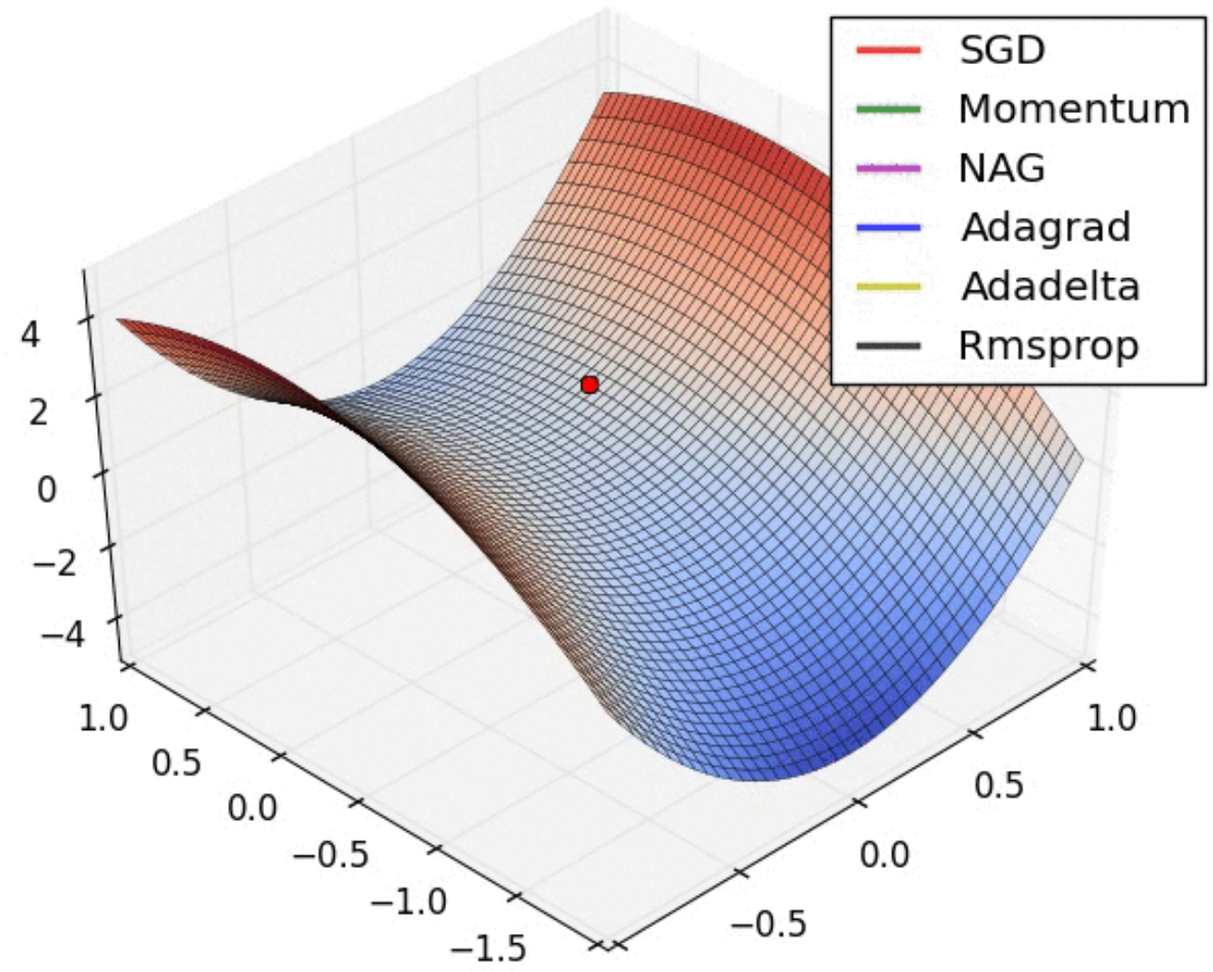
Adaptive Rates

- AdaGrad
- AdaDelta
- RMSprop

- ADAM
- Will cover in next lecture; assignment would require you to self study
- Sebastian Ruder's excellent blog http://sebastianruder.com/optimizing-gradient-descent/index.html

Learning algorithm matters !

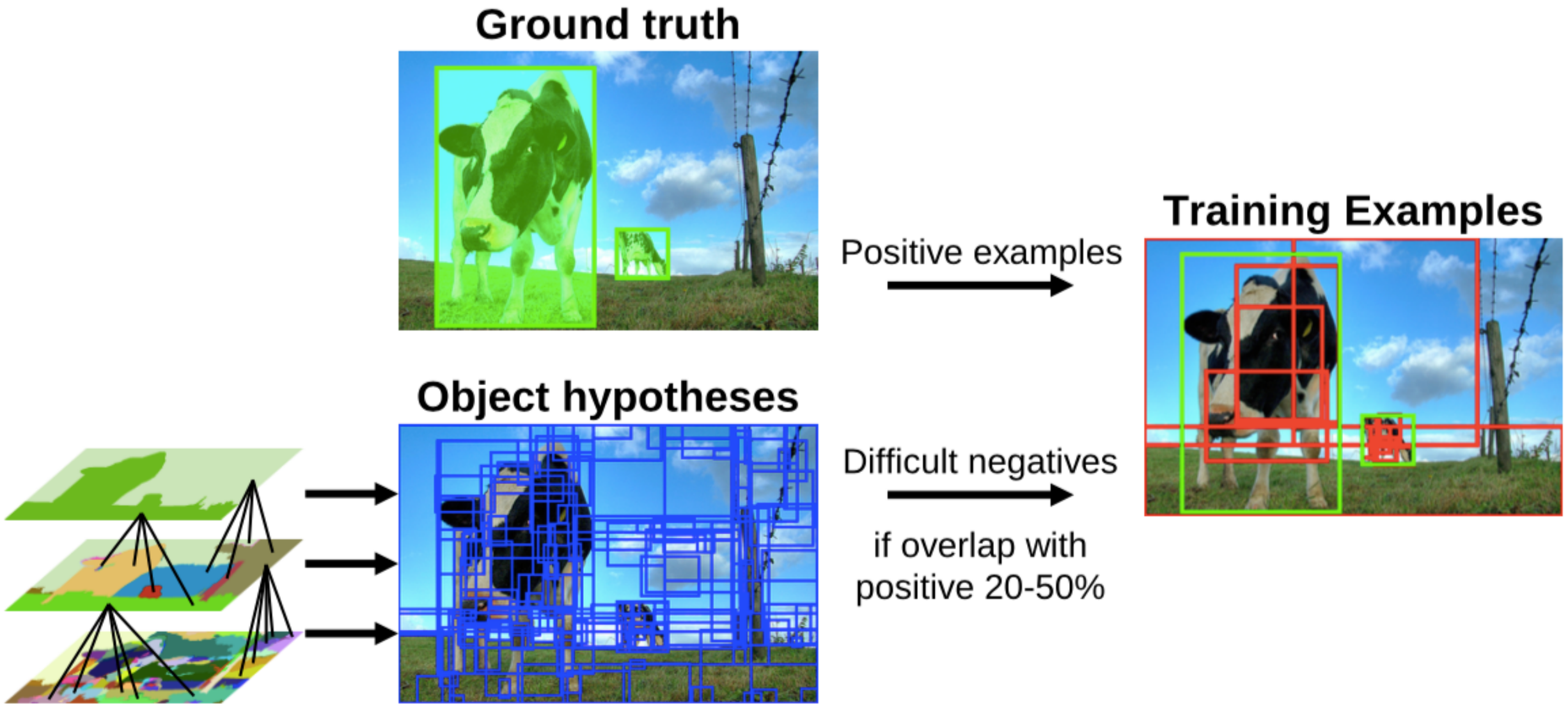




Images credit: Alec Radford (https://twitter.com/alecrad)

State-of-the-art Object Detection Networks

Selective Search



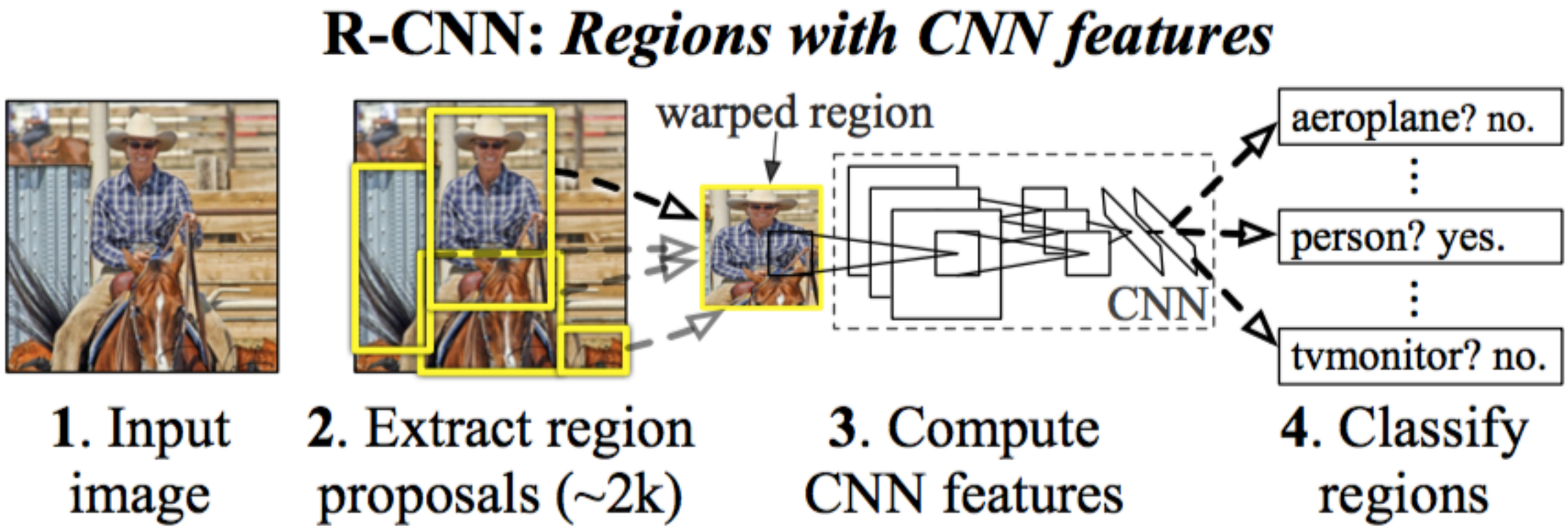
Uijlings et al., Selective Search for Object Recognition, IJCV 2013

Selective Search

- Multi scale
- Diversity (contour, texture, color …)
- Fast to compute

Uijlings et al., Selective Search for Object Recognition, IJCV 2013

R-CNN



- Use pre-trained CNNs on region proposals
- Region proposals = generic object detector

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

R-CNN

- Supervised pre-training
  - Use CNN pre-trained on ImageNet dataset
- Domain specific fine tuning
  - 1000 class classification to 21 class (20 classes + background)
  - Object proposals with IoU > 0.5 as positive
- Also for semantic segmentation

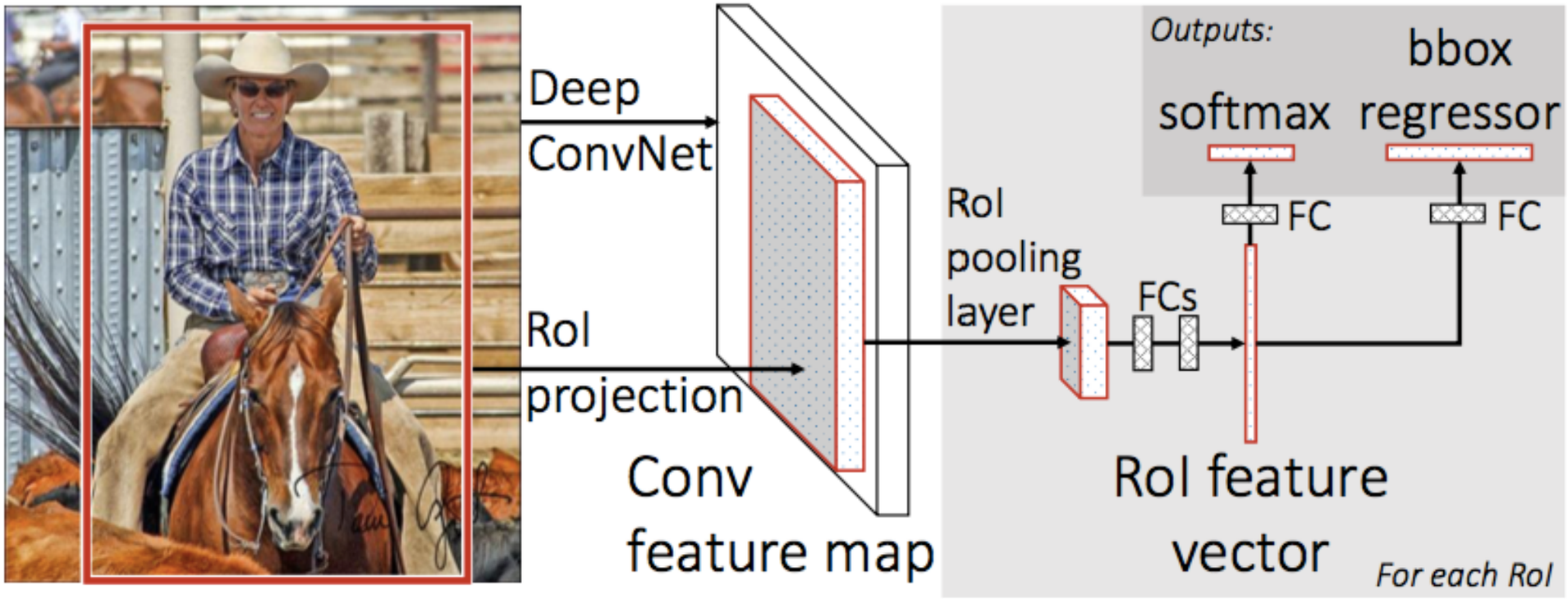
R-CNN results

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 [17]	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 [25]	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 [27]	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

34 to 58 mAP on Pascal Visual Object Challenge (VOC)

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

Fast R-CNN

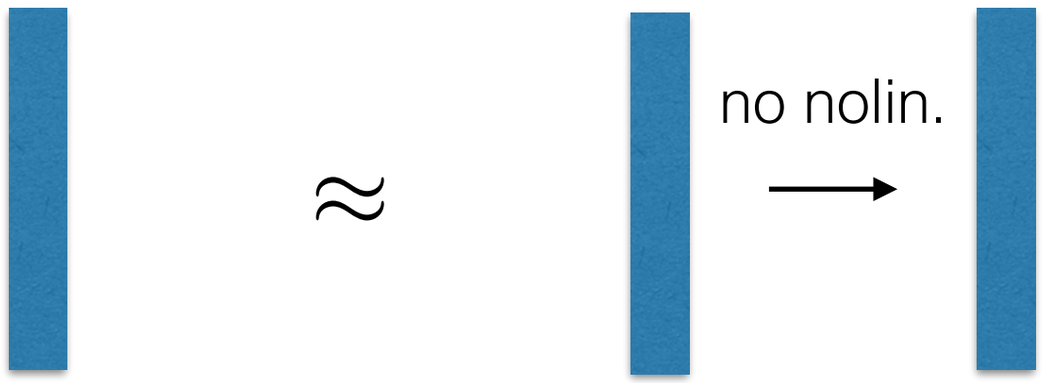


- No redundant (esp. conv.) computations cf. R-CNN
- Training — choose N images, R regions each
  - Typical N=2, R=128

Girshick, Fast R-CNN, ICCV 2015

Truncated SVD

$$W \approx U \Sigma_t V^\top$$



#params uv to $t(u+v)$

W $\quad$ $\Sigma_t V^\top$ $\quad$ U

- Detection — many RoI — time for FC dominates
- Truncated SVD to replace 1x FC to 2x FC layers
  - Keep first t left-singular values; truncate rest

Girshick, Fast R-CNN, ICCV 2015

Fast R-CNN results

	Fast R-CNN			R-CNN		
	S	**M**	**L**	**S**	**M**	**L**
train time (h)	**1.2**	2.0	9.5	22	28	84
train speedup	**18.3**×	14.0×	8.8×	1×	1×	1×
test rate (s/im)	0.10	0.15	0.32	9.8	12.1	47.0
▷ with SVD	**0.06**	0.08	0.22	-	-	-
test speedup	98×	80×	146×	1×	1×	1×
▷ with SVD	169×	150×	**213**×	-	-	-
VOC07 mAP	57.1	59.2	**66.9**	58.5	60.2	66.0
▷ with SVD	56.5	58.7	66.6	-	-	-

Girshick, Fast R-CNN, ICCV 2015