Fast R-CNN

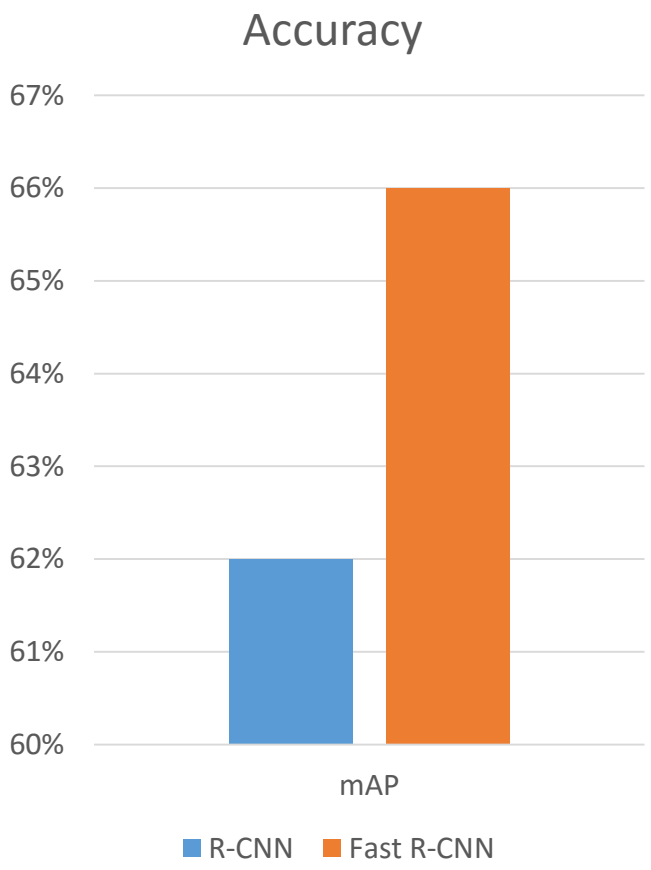
Author: Ross Girshick

Speaker: Charlie Liu

Date: Oct, 13th

Girshick, R. (2015). Fast R-CNN. arXiv preprint arXiv:1504.08083.

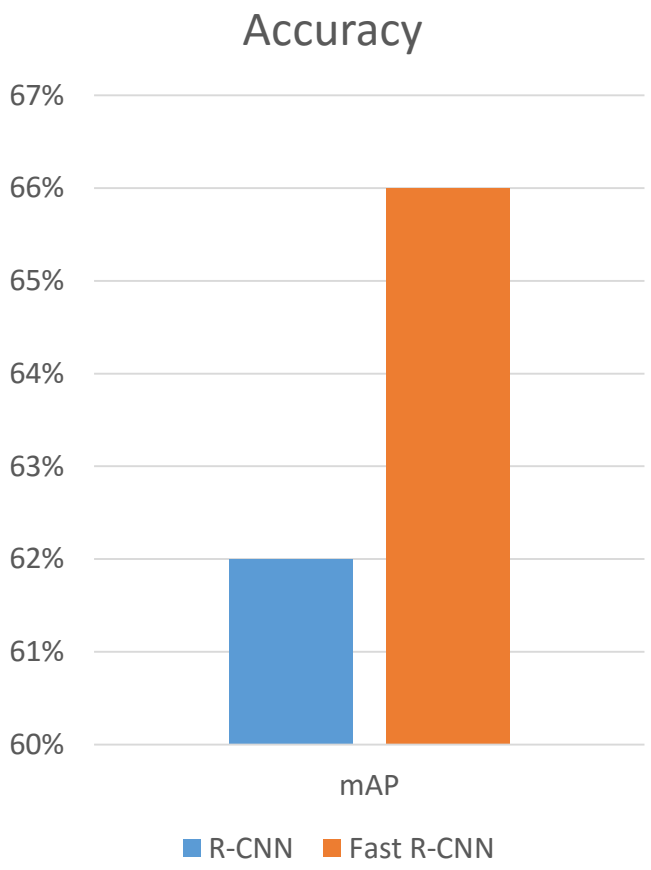
Result[1]

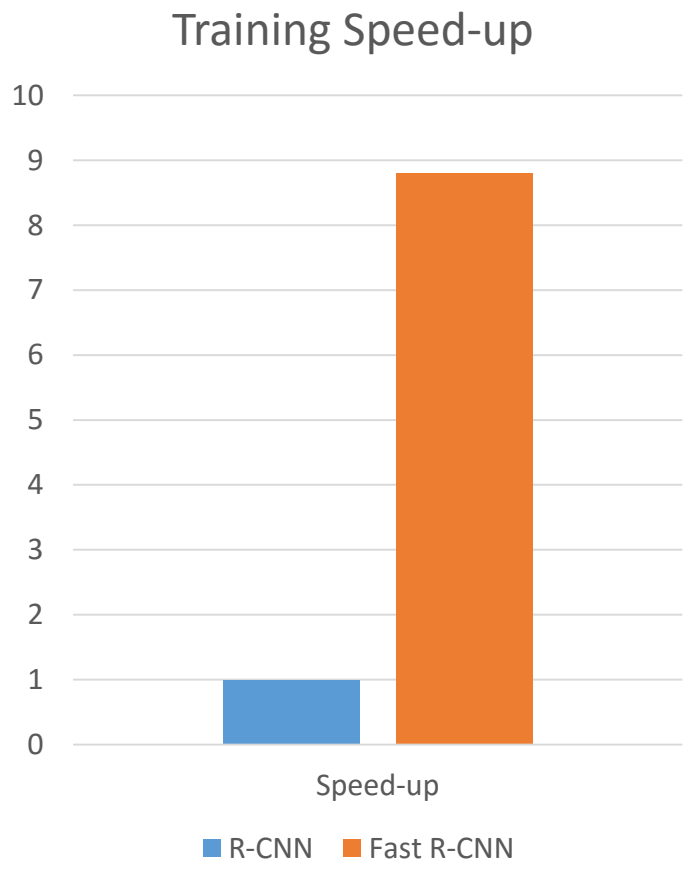


Girshick. Fast R-CNN
Girshick et. Al. Rich feature hierarchies for accurate object detection and semantic segmentation

1. Dataset: PASCAL VOC 2012

Result[1]

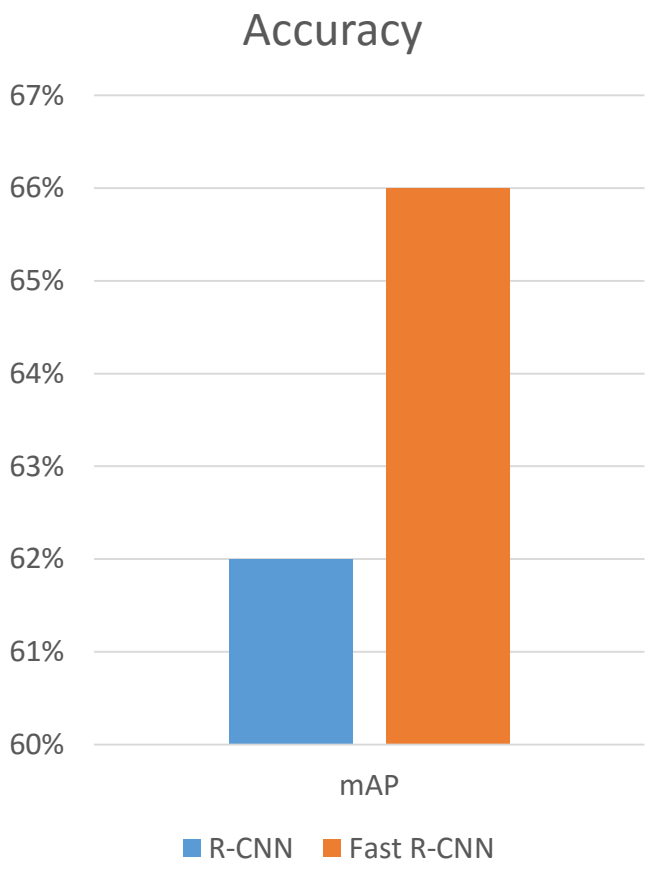


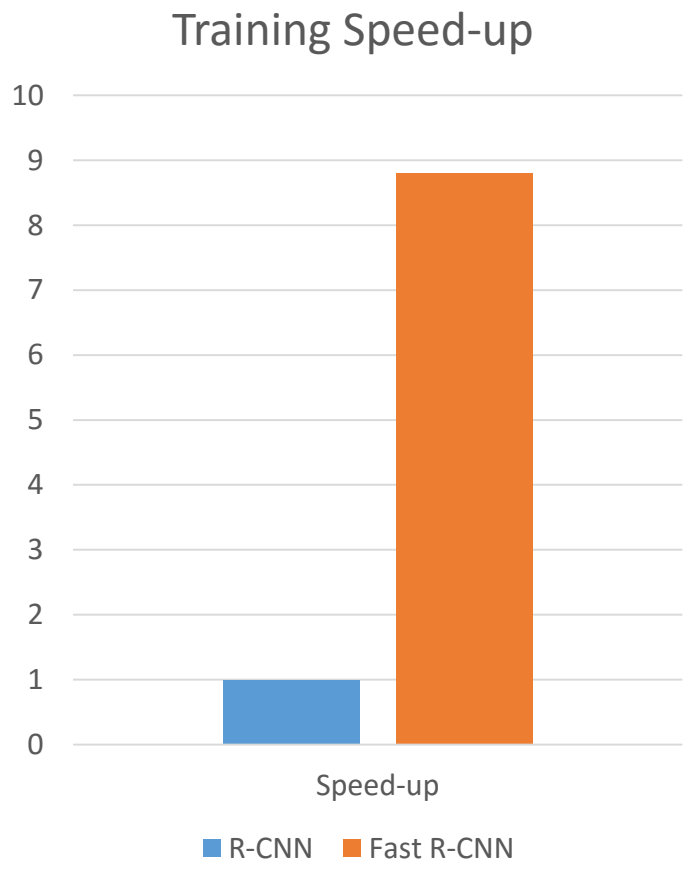


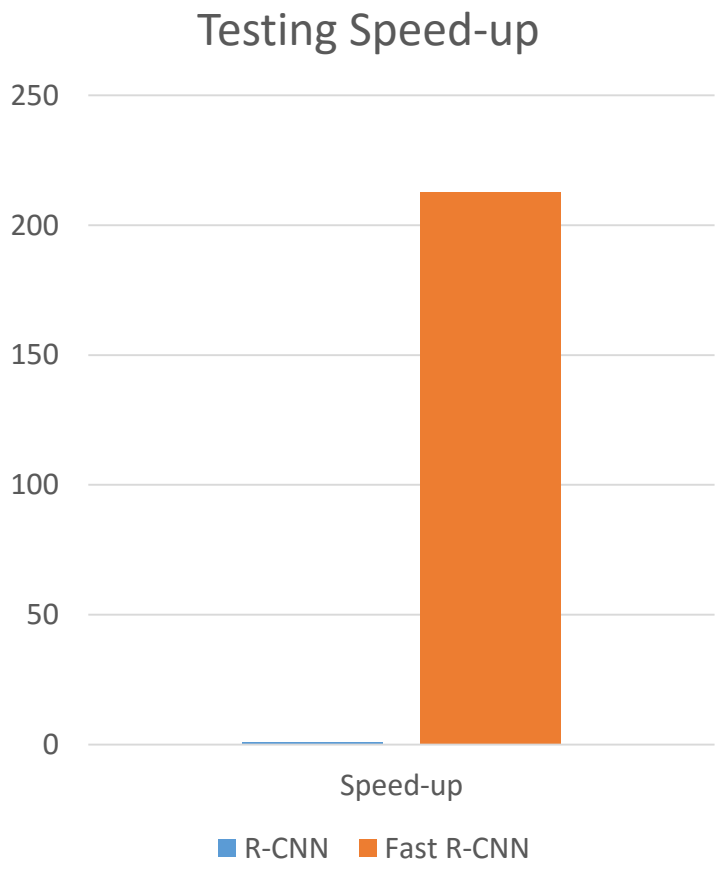
Girshick. Fast R-CNN
Girshick et. Al. Rich feature hierarchies for accurate object detection and semantic segmentation

1. Dataset: PASCAL VOC 2012

Result[1]







Girshick. Fast R-CNN
Girshick et. Al. Rich feature hierarchies for accurate object detection and semantic segmentation

1. Dataset: PASCAL VOC 2012

Object Detection after R-CNN

R-CNN

- 66.0% mAP[1]
- 1x Test Speed

SPPnet

- 63.1% mAP
- 24x than R-CNN

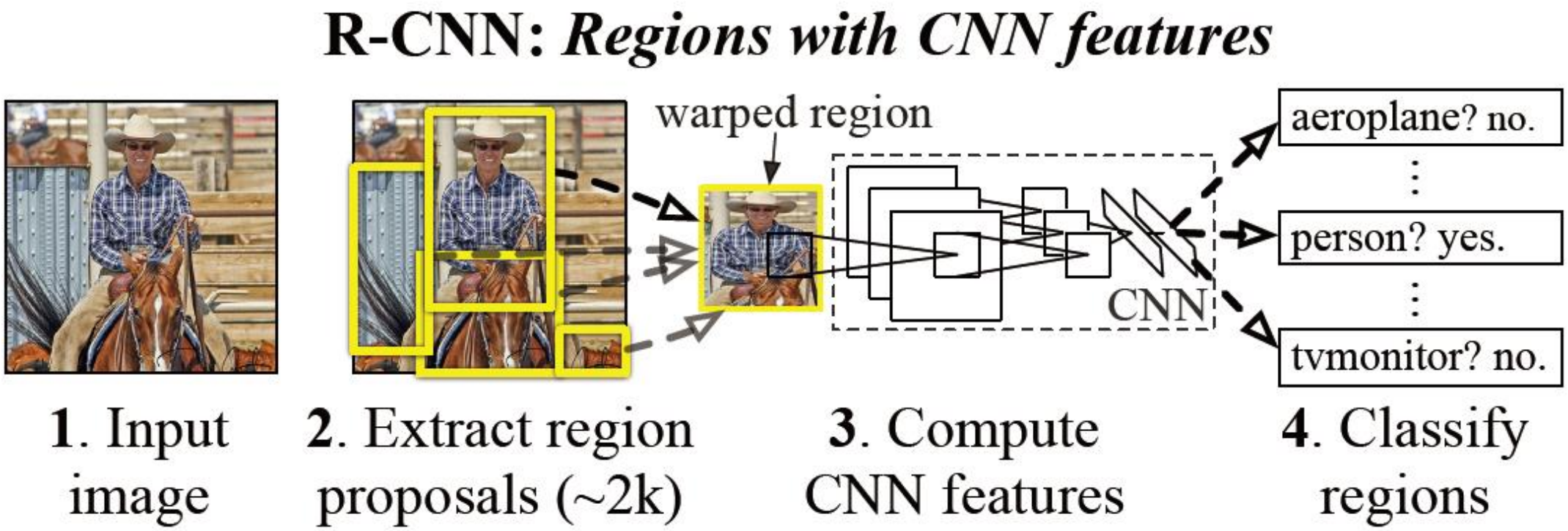
Fast R-CNN

- 66.6% mAP
- 10x than SPPnet

Girshick. Fast R-CNN
He et. al. Spatial pyramid pooling in deep convolutional networks for visual recognition
Girshick et. al. Rich feature hierarchies for accurate object detection and semantic segmentation

1. mAP based on PASCAL VOC 2007, results from Girshick

R-CNN



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

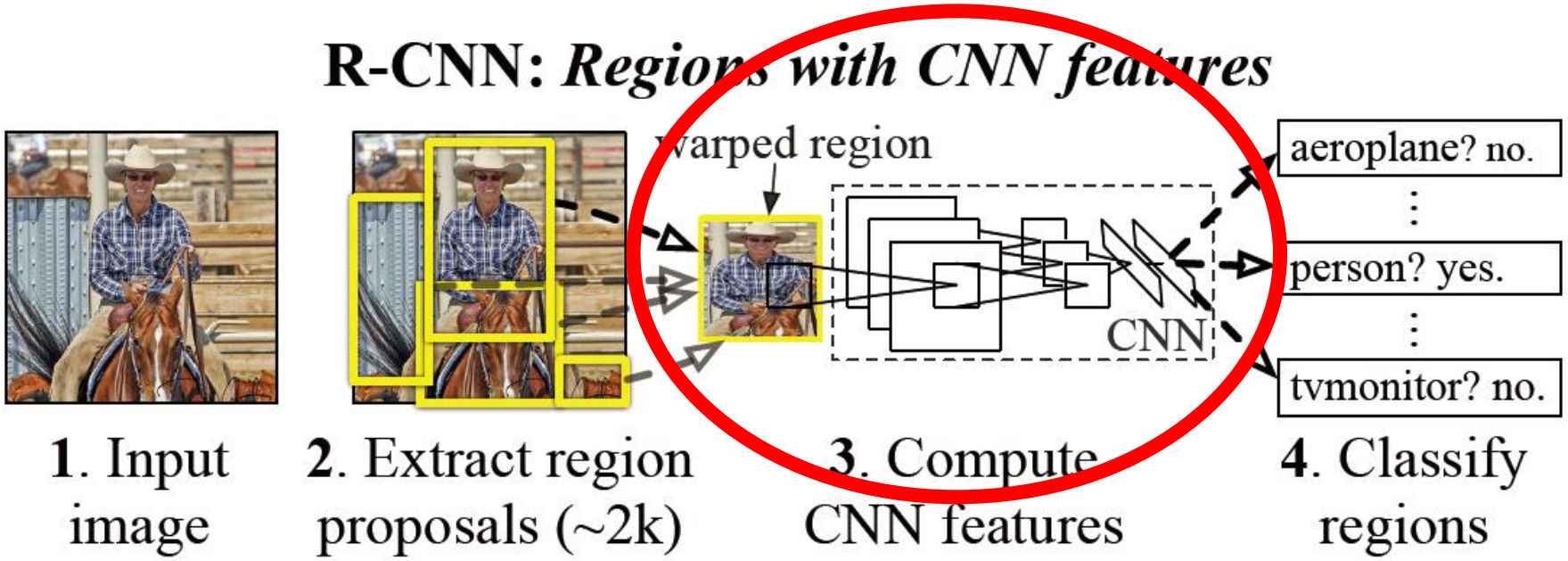
R-CNN Limitations

- Too slow
  - 13s/image on a GPU
  - 53s/image on a CPU
  - VGG-Net 7x slower

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

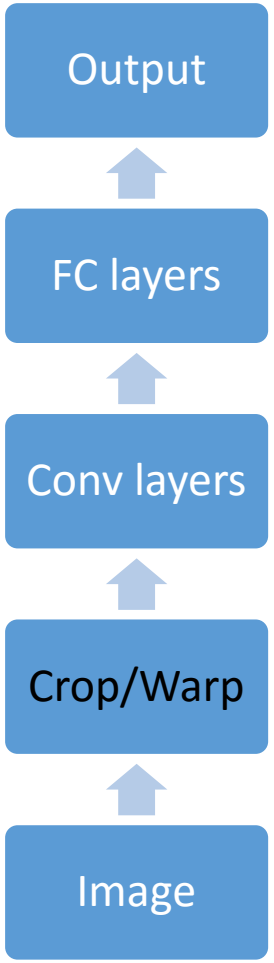
R-CNN Limitations

- Too slow
- Proposals need to be warped to a fixed size
  - Potential loss of accuracy



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

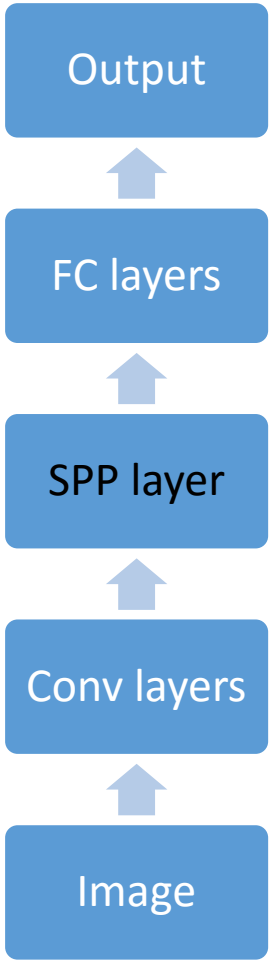
R-CNN Limitations



- Cropping may loss the object's information
- Warping may change the object's appearance

He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

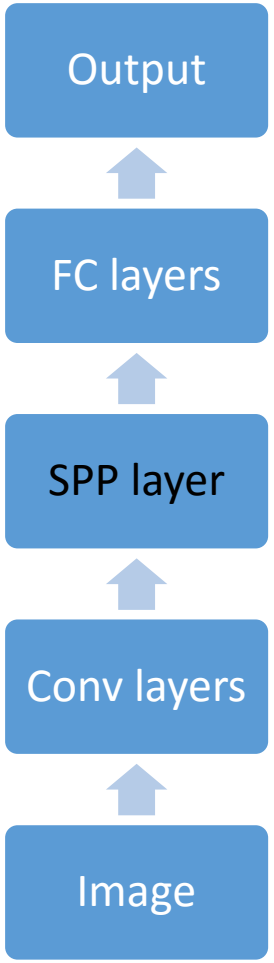
SPPnet: Motivation



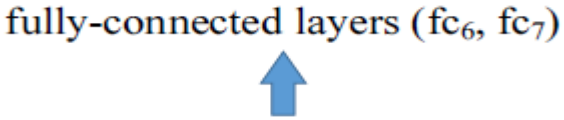
- SPP: Spatial Pyramid Pooling

He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

SPPnet: Motivation



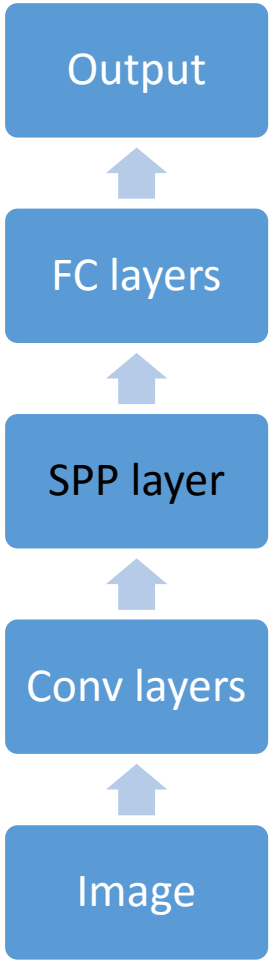
- SPP: Spatial Pyramid Pooling
  - Fixed size input



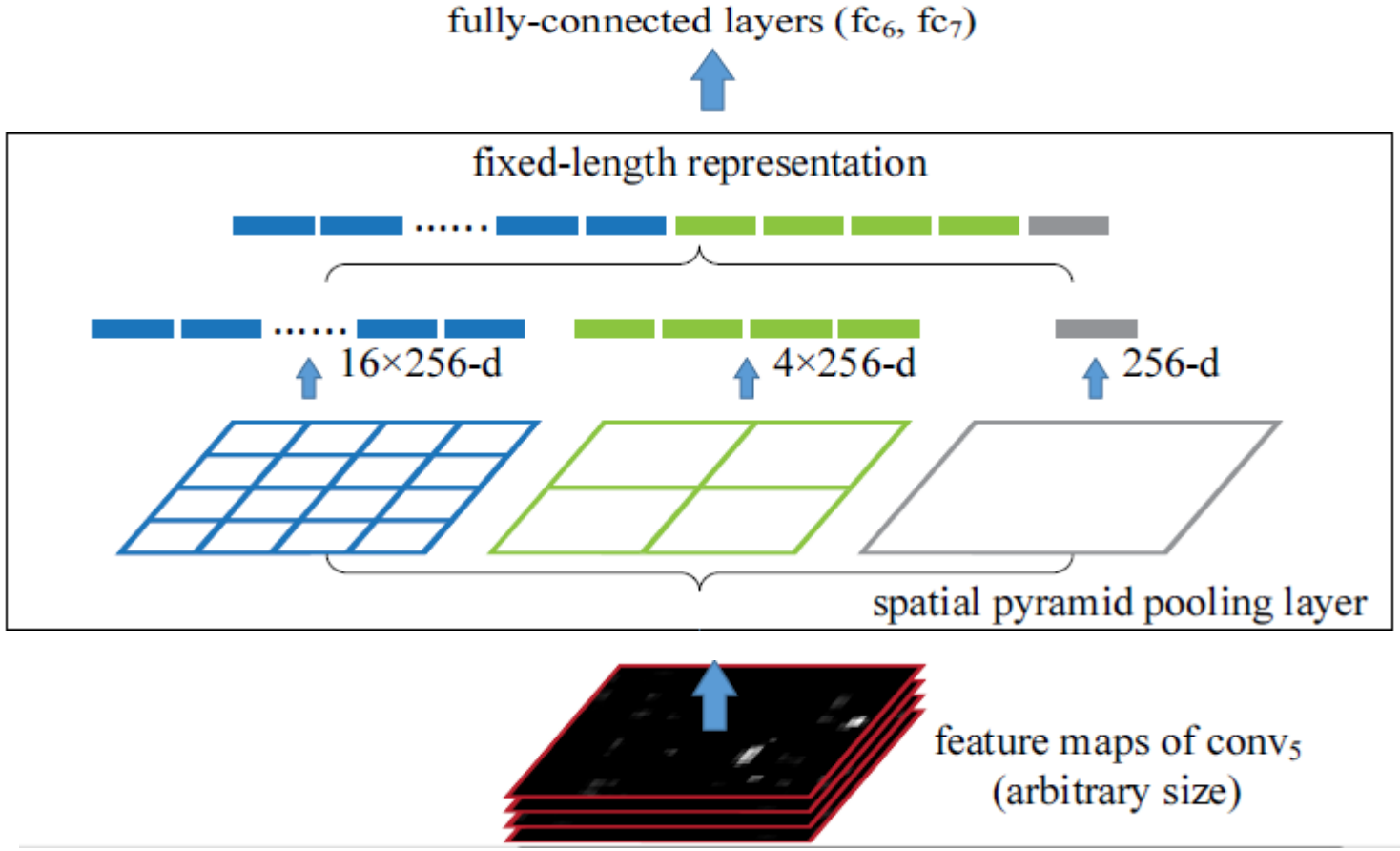
  - Arbitrary size output

He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

SPPnet: Motivation

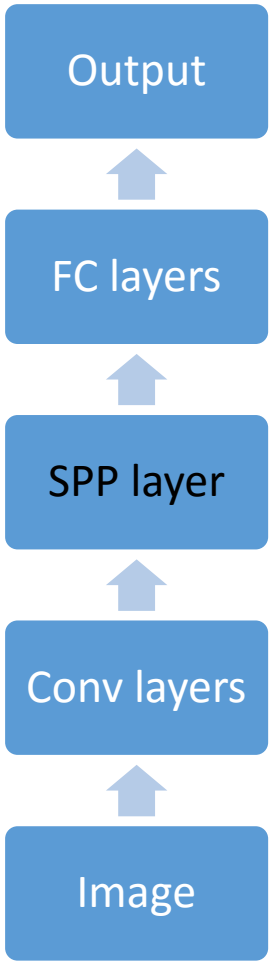


- SPP: Spatial Pyramid Pooling

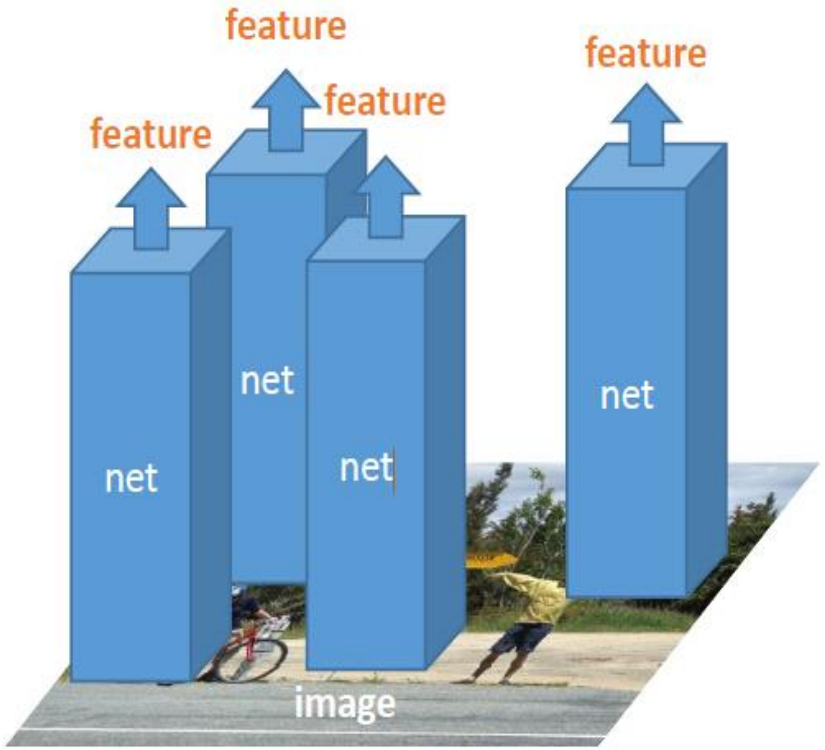


He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

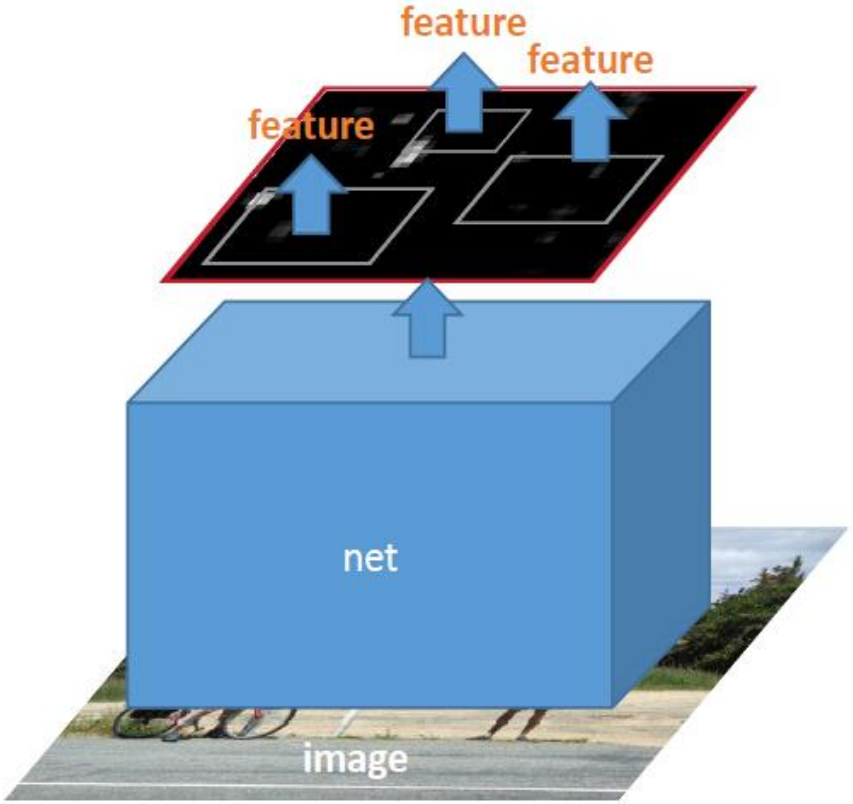
SPPnet: Motivation



R-CNN: 2000 nets on 1 image

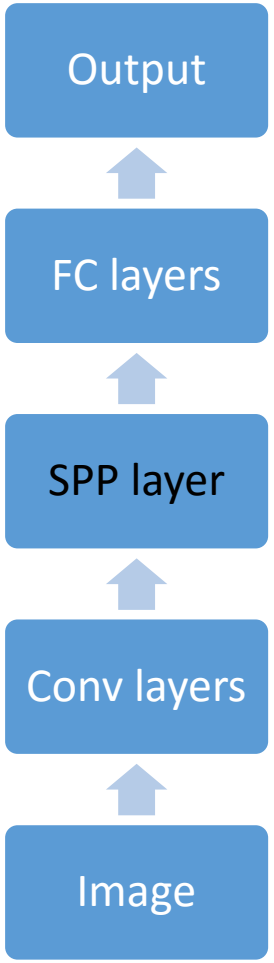


SPPnet: 1 net on 1 image

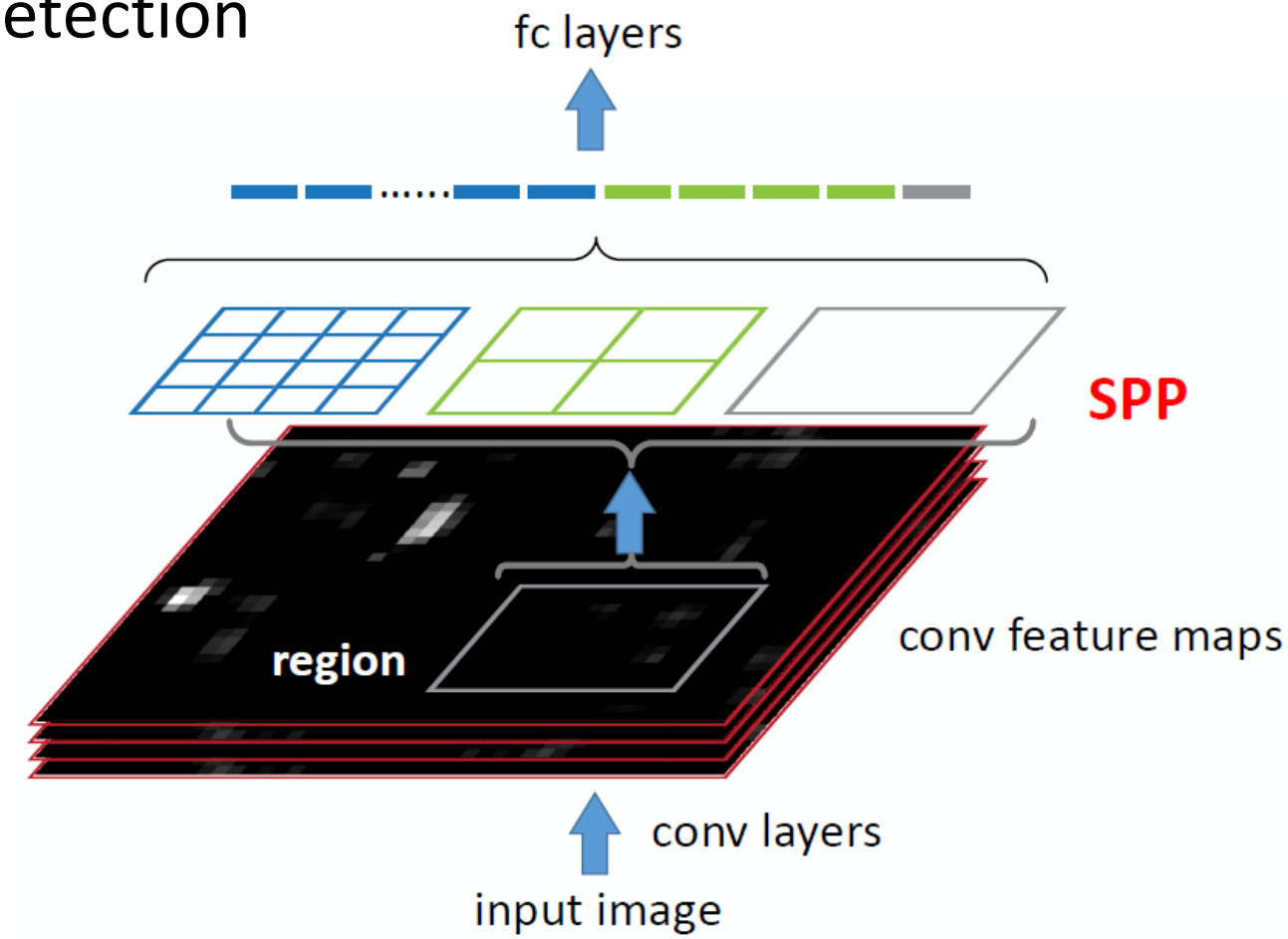


He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

SPPnet: Motivation



- SPP Detection



He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

SPPnet: Result

	R-CNN	SPPnet 1-scale	SPPnet 5-scale
mAP	58.5	58.0	**59.2**
GPU time / img	9s	**0.14s**	0.38s
Speed-up[2]	1x	64x	24x

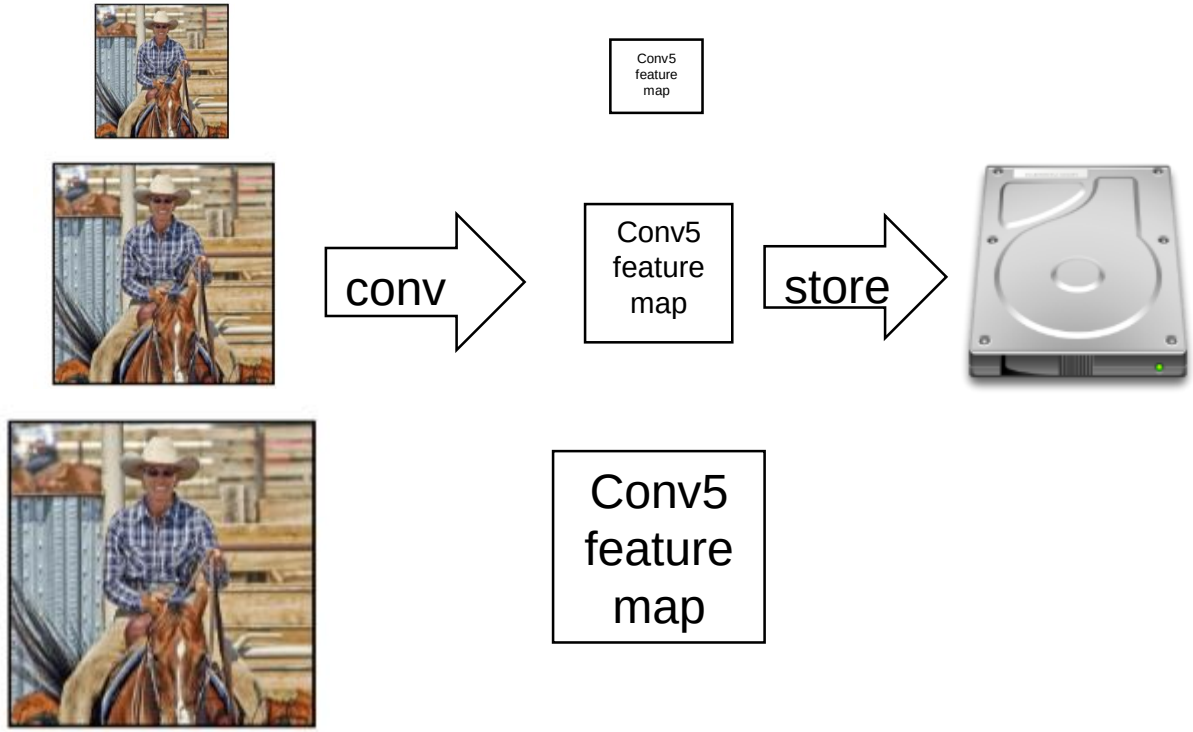
VOC 2007[1]

He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

1. mAP based on PASCAL VOC 2007, results from *He et al.*
2. Speed stands for testing speed

SPPnet: Limitations

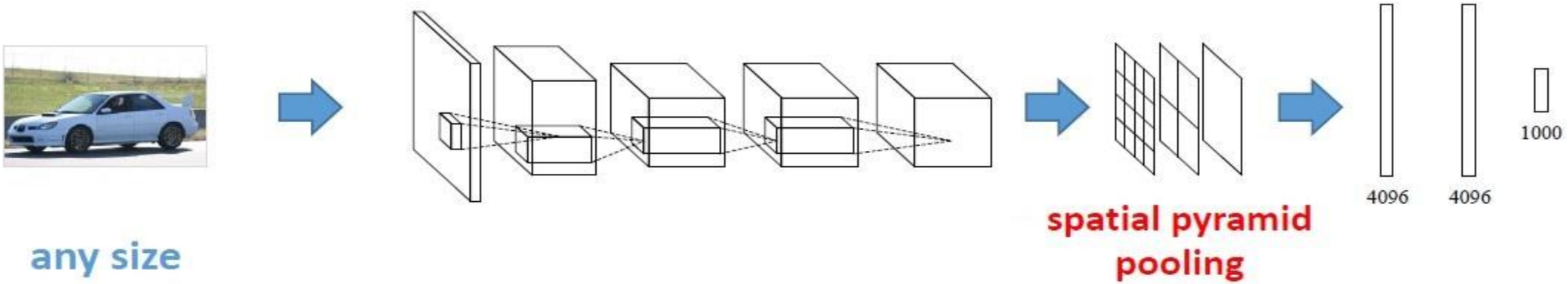
- High memory consumption



He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

SPPnet: Limitations

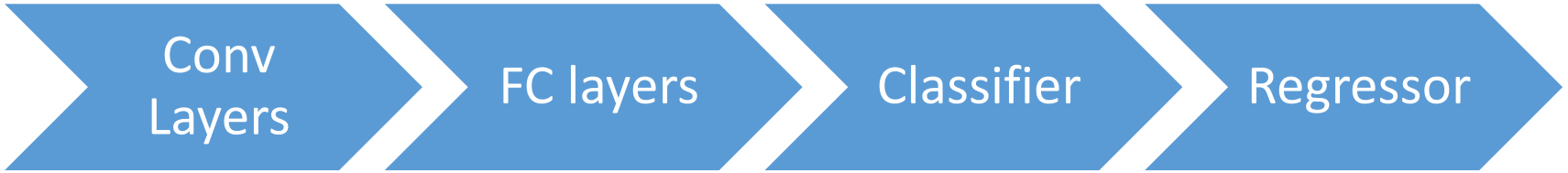
- High memory consumption
- Training is inefficient
  - Inefficient back-propagation through SPP



He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

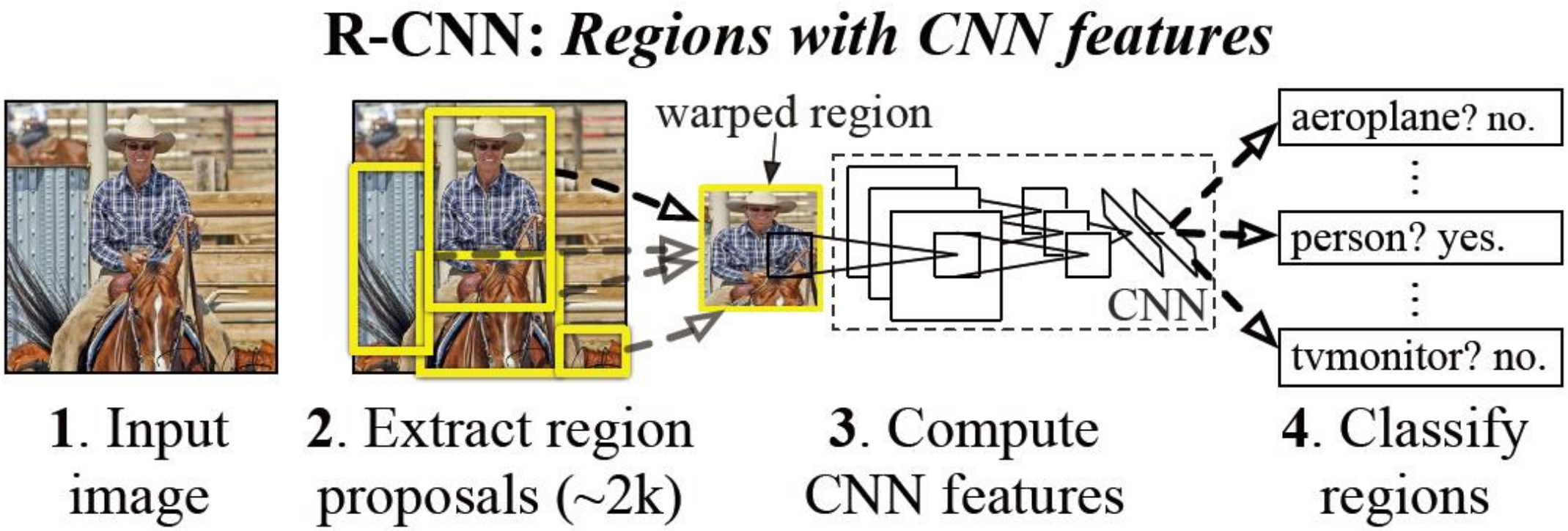
SPPnet: Limitations

- High memory consumption
- Training is inefficient
  - Inefficient back-propagation through SPP
- Multi-stage pipeline



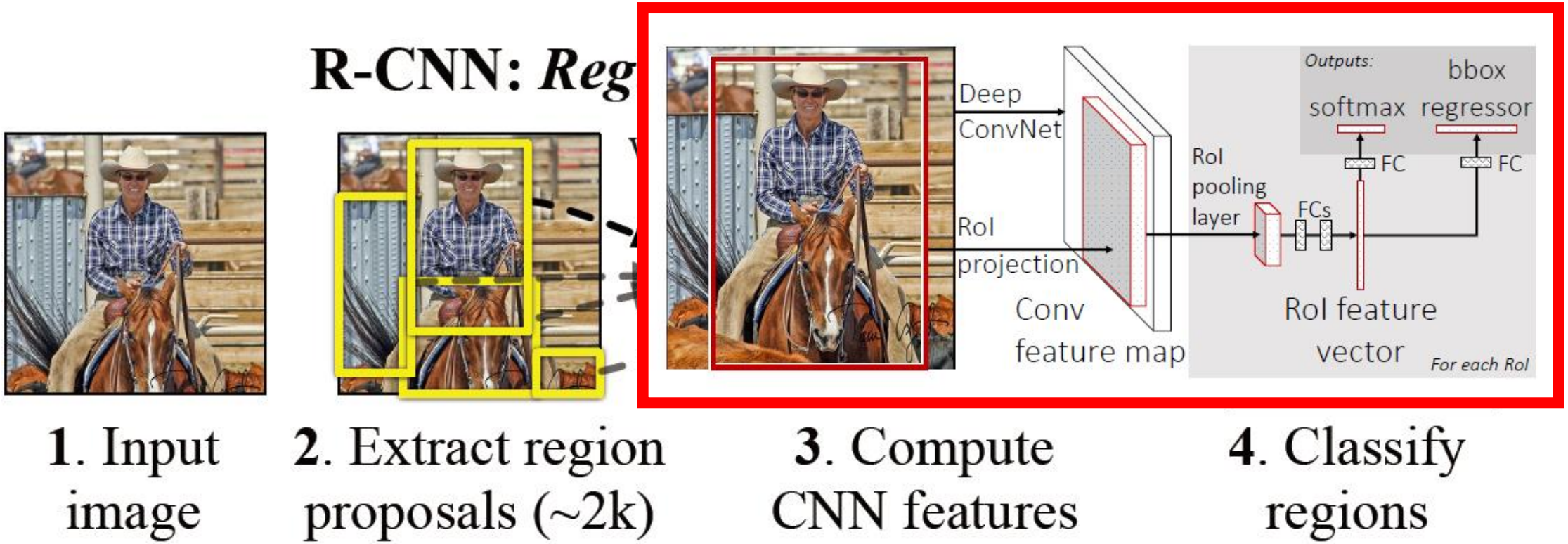
He et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Fast R-CNN: What's New



Girshick et. al. Rich feature hierarchies for accurate object detection and semantic segmentation
Girshick. Fast R-CNN

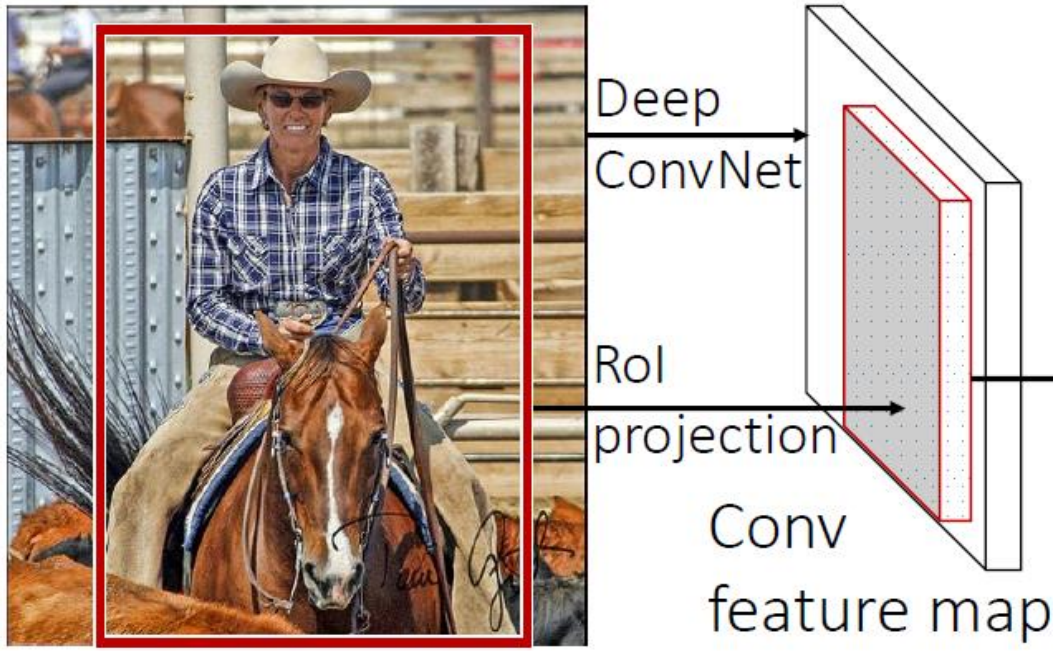
Fast R-CNN: What's New



Girshick et. al. Rich feature hierarchies for accurate object detection and semantic segmentation
Girshick. Fast R-CNN

Fast R-CNN: What's New

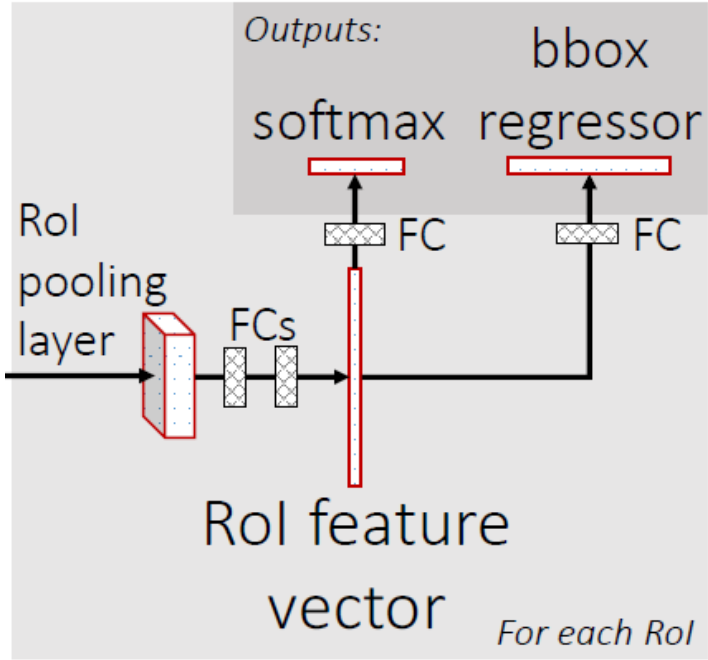
- RoI Pooling Layer
  - ≈ one scale SPP layer



Girshick. Fast R-CNN

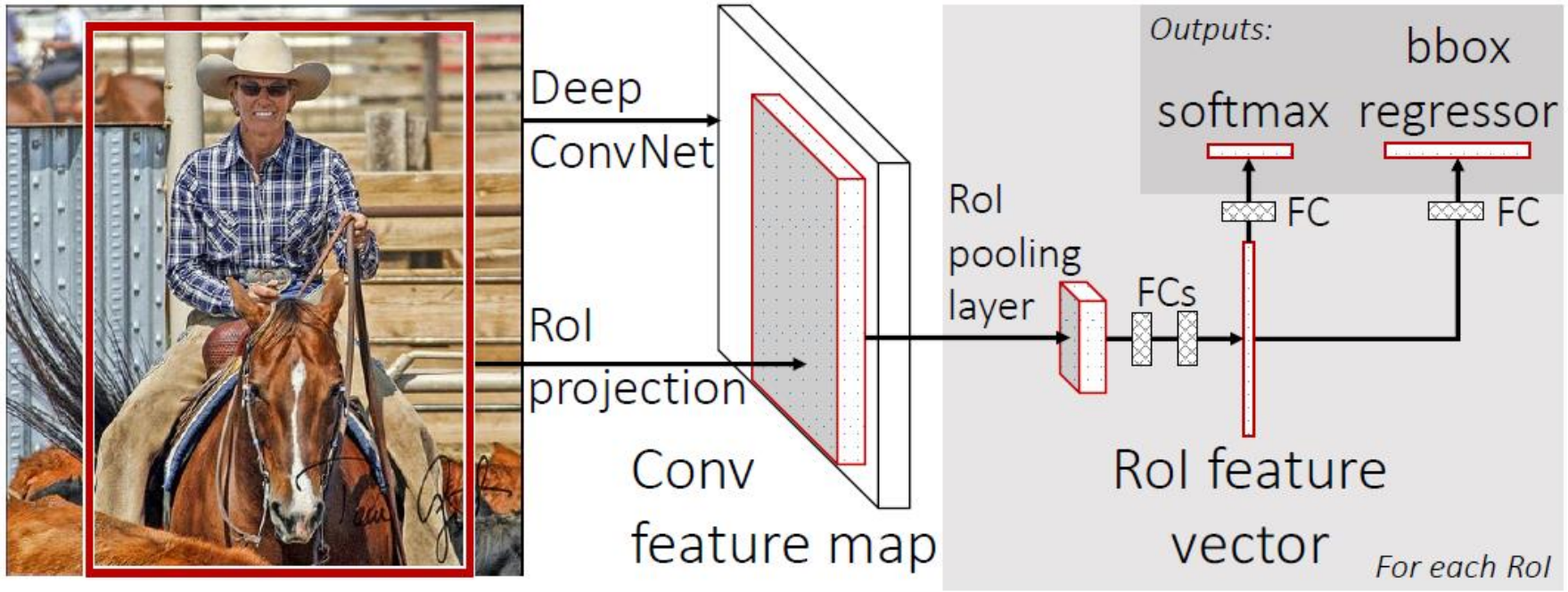
Fast R-CNN: What's New

- Sibling Classifier & Regressor Layers



Girshick. Fast R-CNN

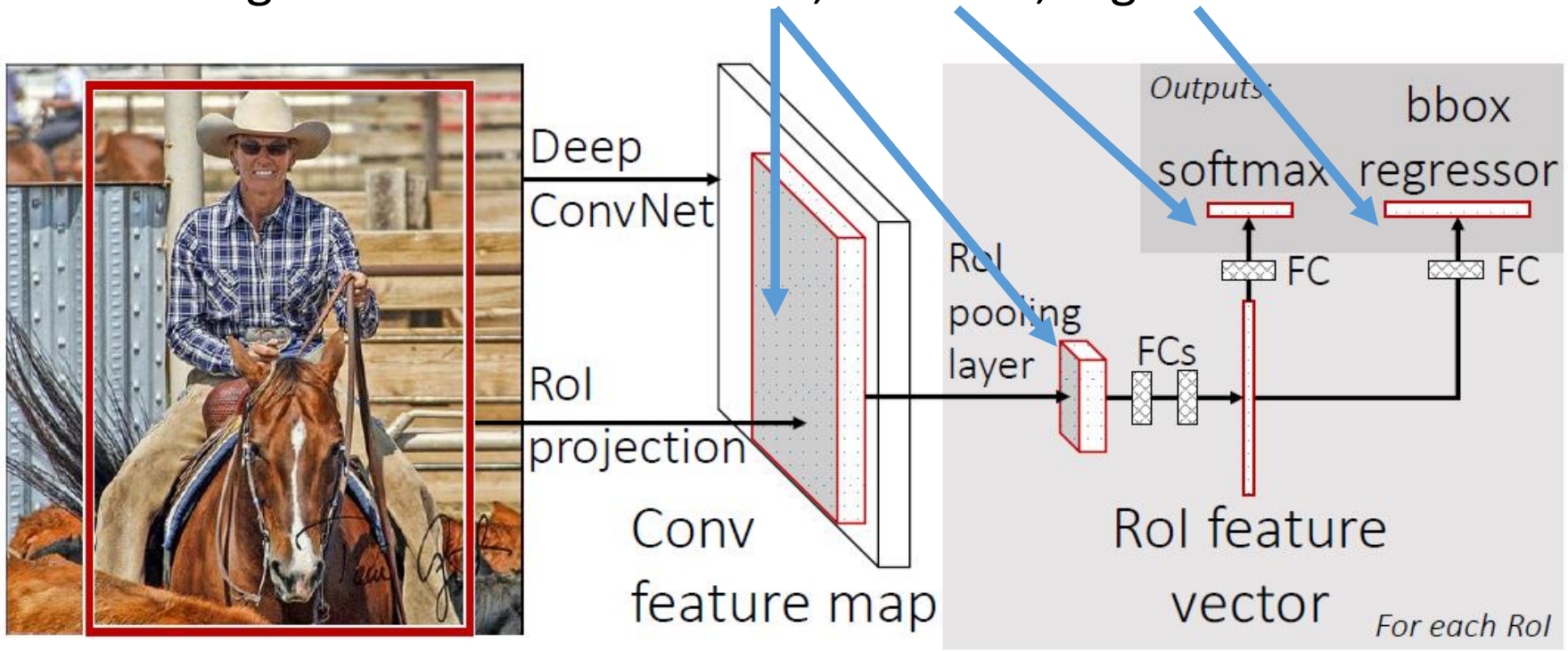
Fast R-CNN: What's New



Girshick. Fast R-CNN

Fast R-CNN: What's New

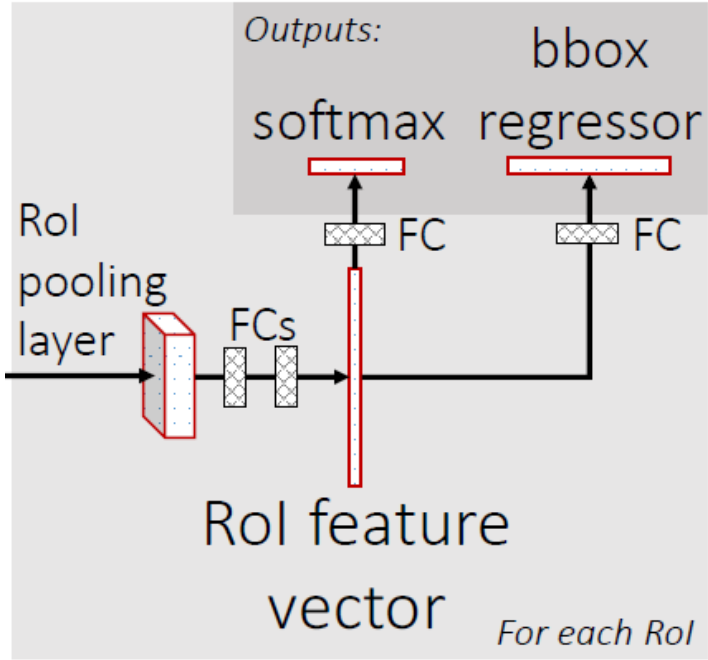
- Joint Training – one stage framework
  - Joint them together: feature extractor, classifier, regressor



Girshick. Fast R-CNN
Liliang Zhang, Detection: From R-CNN to Fast R-CNN

Fast R-CNN: Advantages

- One fine-tuning stage
  - Optimizes the classifier and bbox regressor
    - Loss function:
    - $L = L_{classifier} + \lambda L_{regressor}$



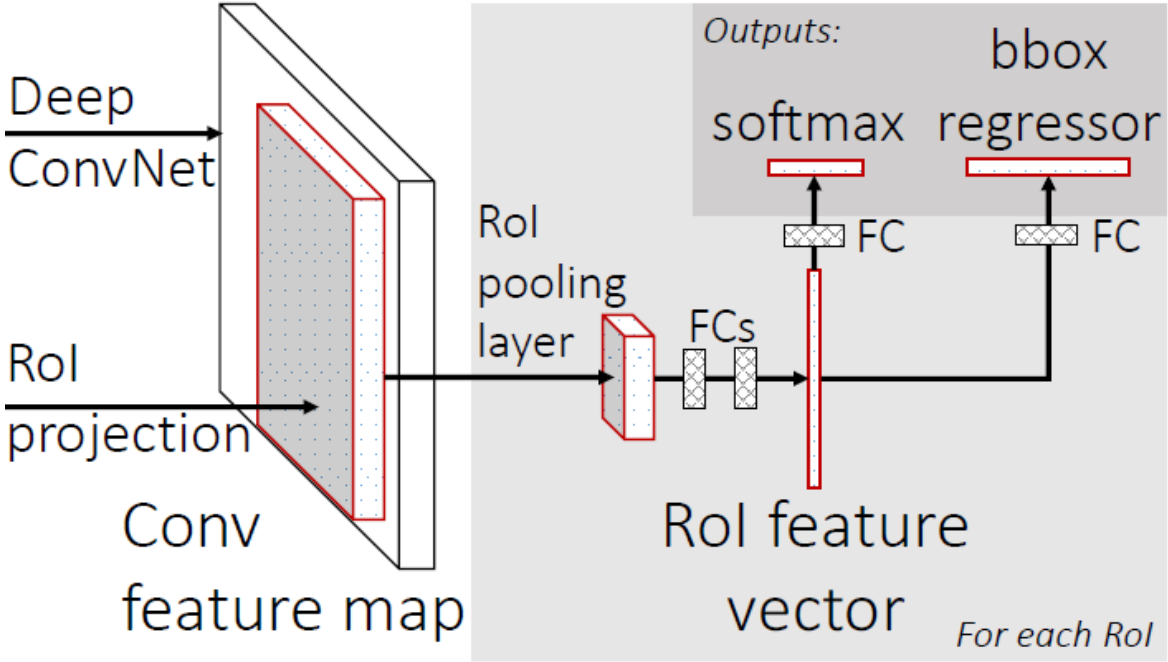
Girshick. Fast R-CNN

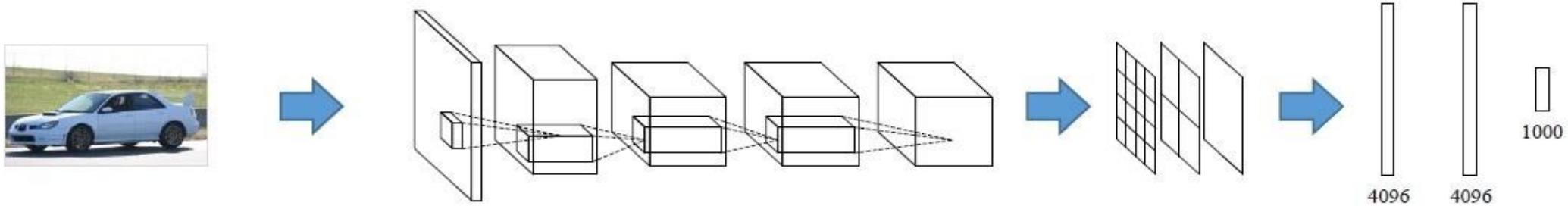
Fast R-CNN: Advantages

- One fine-tuning stage
- Fast training
  - Using mini-batch stochastic gradient descent
    - E.g. 2 images * 64 RoIs each
    - 64x faster than 128 images * 1 RoI each (strategies of R-CNN, SPPnet)

Fast R-CNN: Advantages

- One fine-tuning stage
- Fast training
- Efficient back-propagation

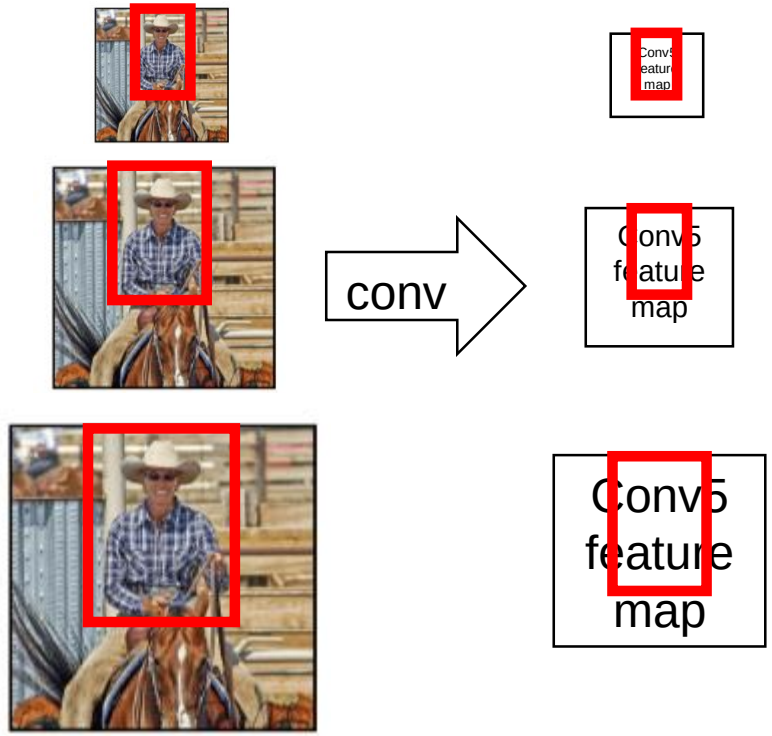




Girshick. Fast R-CNN
Liliang Zhang, Detection: From R-CNN to Fast R-CNN

Fast R-CNN: Advantages

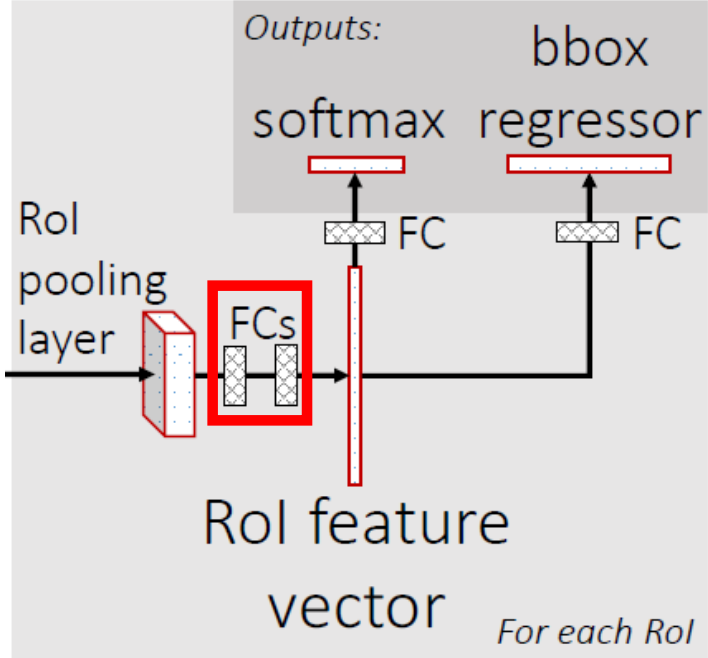
- One fine-tuning stage
- Fast training
- Efficient back-propagation
- Scale invariance
  - In practice, single scale is good enough
  - Single scale: faster x10 than SPP-Net



Girshick. Fast R-CNN
Liliang Zhang, Detection: From R-CNN to Fast R-CNN

Fast R-CNN: Advantages

- One fine-tuning stage
- Fast training
- Efficient back-propagation
- Scale invariance
- Fast detecting
  - Truncated SVD[1] $W \approx U\Sigma_t V^T$
  - Single FC layer -> two FC layers
  - Reduce parameters[2] from uv to $t(u + v)$

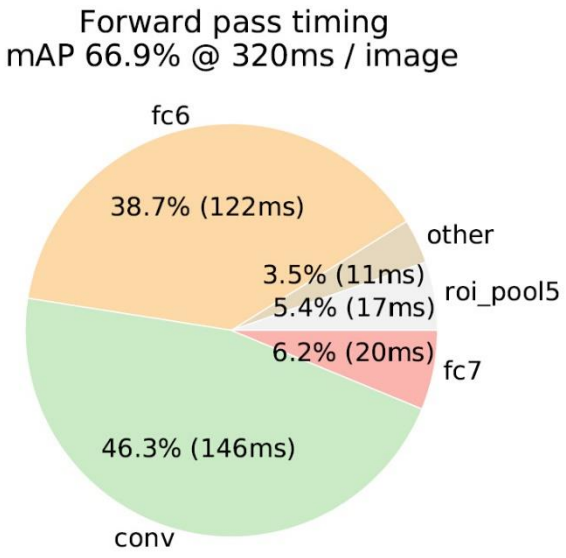


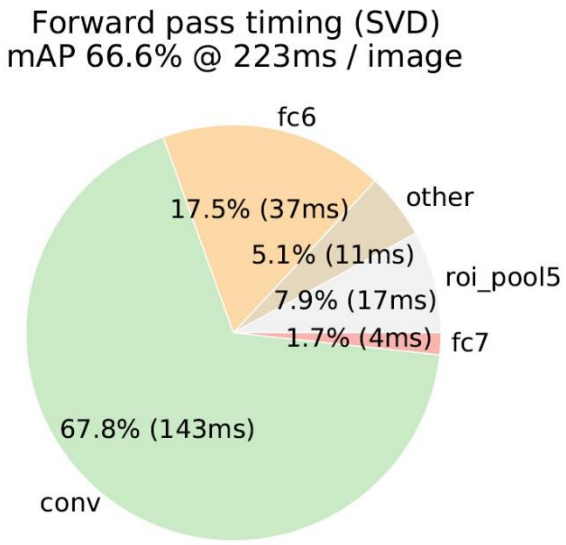
1. *U: size(u * t); Σ: size(t * t); V: size(v * t);*
2. *In practice, t << min(u, v)*

Girshick. Fast R-CNN
Liliang Zhang, Detection: From R-CNN to Fast R-CNN

Fast R-CNN: Advantages

- One fine-tuning stage
- Fast training
- Efficient back-propagation
- Scale invariance
- Fast detecting
  - Truncated SVD[1] $W \approx U\Sigma_t V^T$
  - Single FC layer -> two FC layers
  - Reduce parameters[2] from *uv* to *t(u + v)*

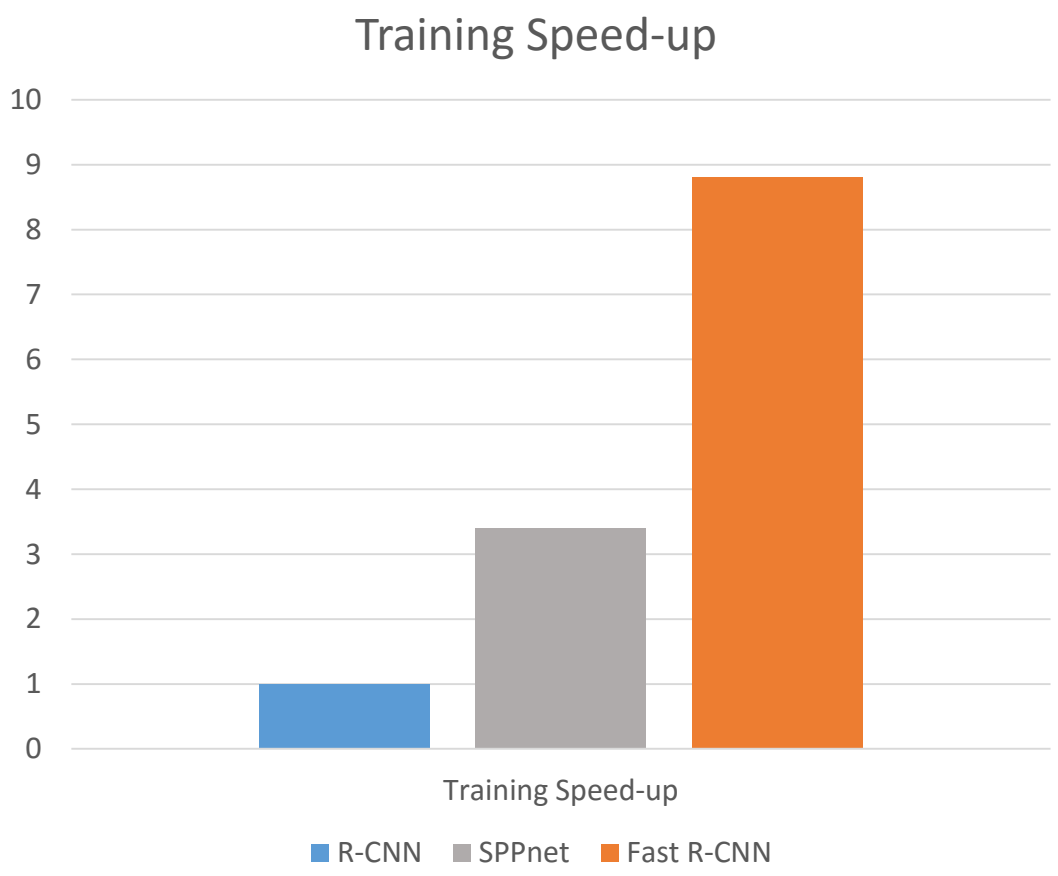




1. *U: size(u * t); Σ: size(t * t); V: size(v * t);*
2. *In practice, t << min(u, v)*

Girshick. Fast R-CNN
Liliang Zhang, Detection: From R-CNN to Fast R-CNN

Fast R-CNN: Result[1]

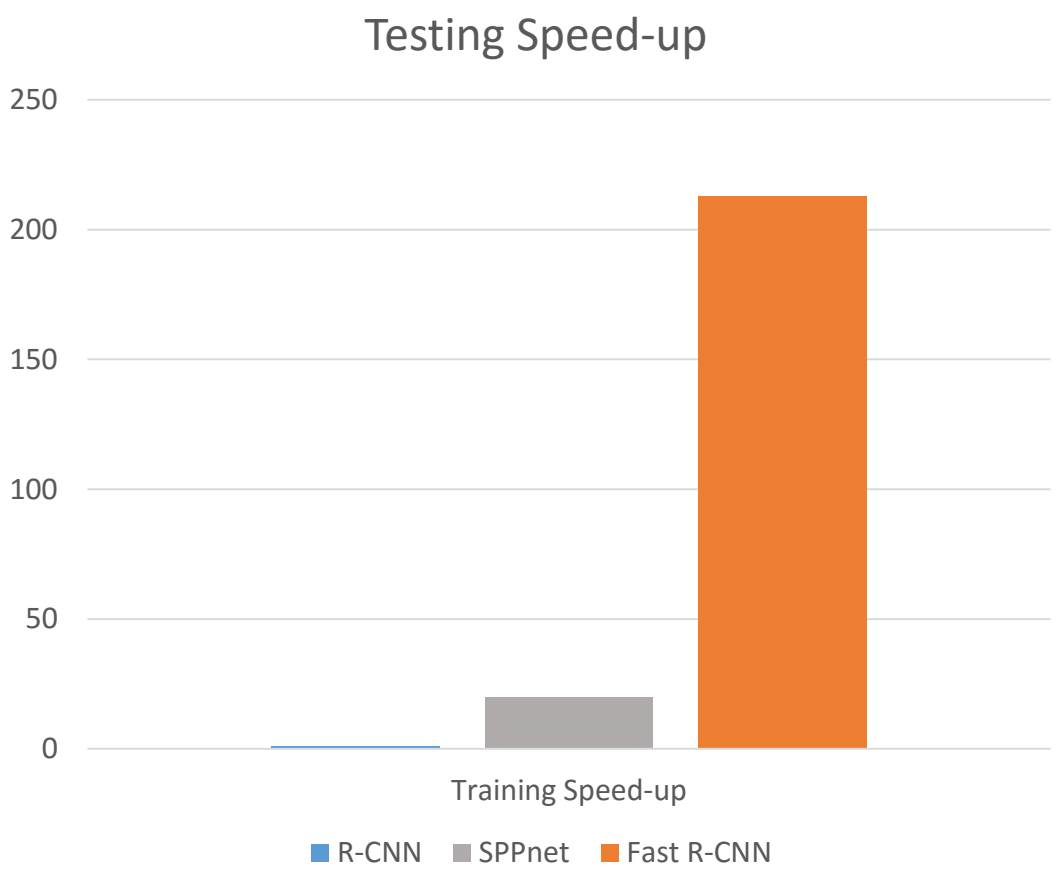


Results using VGG16

	R-CNN	SPPnet	Fast R-CNN
Train Time (h)	84	25	**9.5**
Train Speedup	1x	3.4x	**8.8x**
Test Rate (s/im)	47.0	2.3	0.22
Test Speedup	1x	20x	213x
VOC07 mAP	66.0%	63.1%	66.6%

1. Results from Girshick

Fast R-CNN: Result[1]



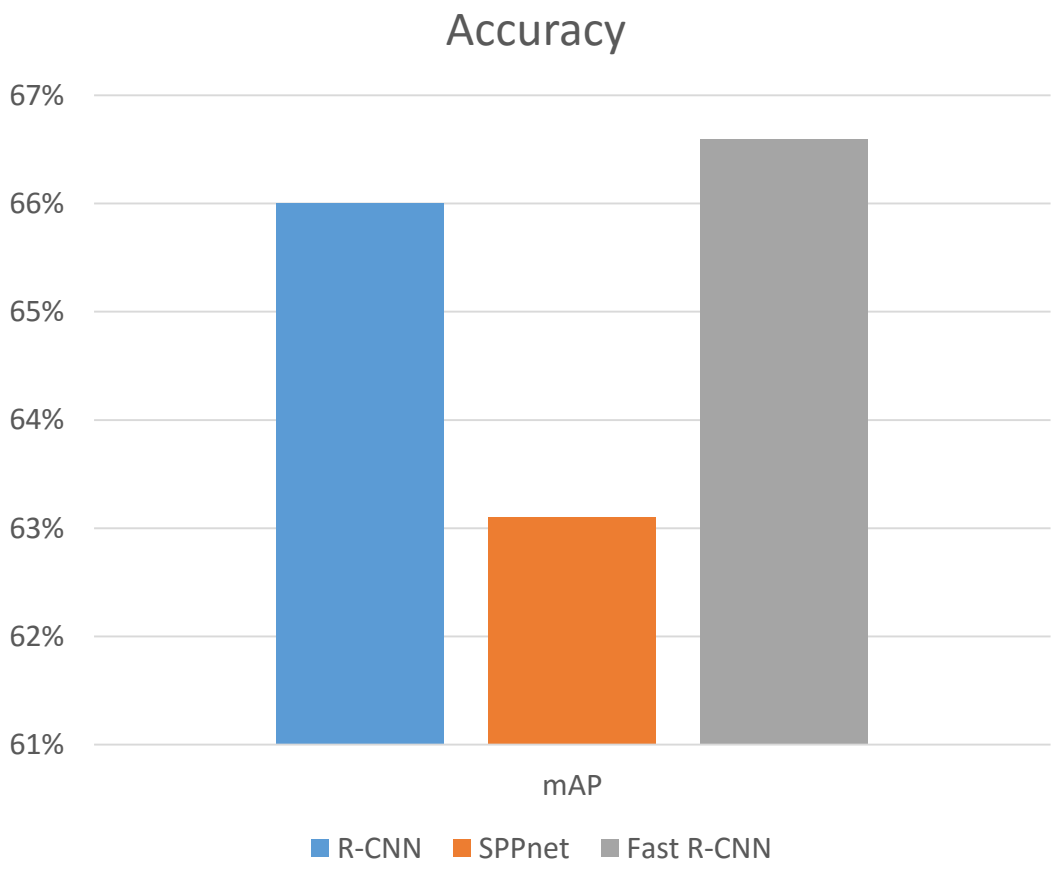
Results using VGG16

	R-CNN	SPPnet	Fast R-CNN
Train Time (h)	84	25	9.5
Train Speedup	1x	3.4x	8.8x
Test Rate (s/im)	47.0	2.3	**0.22**
Test Speedup	1x	20x	**213x**
VOC07 mAP	66.0%	63.1%	66.6%

Girshick. Fast R-CNN

1. Results from Girshick

Fast R-CNN: Result[1]



Results using VGG16

	R-CNN	SPPnet	Fast R-CNN
Train Time (h)	84	25	9.5
Train Speedup	1x	3.4x	8.8x
Test Rate (s/im)	47.0	2.3	0.22
Test Speedup	1x	20x	213x
VOC07 mAP	66.0%	63.1%	**66.6%**

Girshick. Fast R-CNN

1. Results from Girshick

Fast R-CNN: Discussions

- Which layers to fine-tune?
  - Fine-tuning $conv_1$: over-runs GPU memory.
  - Fine-tuning $conv_2$: mAP +0.3%, 1.3x slower.
  - Fine-tuning $conv_3$ and up.

Fast R-CNN: Discussions

- Which layers to fine-tune? $conv_3$ and up.
- Does multi-task training help?
  - Multi-task: potential to improve results
  - Tasks influence each other through a shared representation
  - mAP +0.8% to 1.1%

Fast R-CNN: Discussions

- Which layers to fine-tune? $conv_3$ and up.
- Does multi-task training help? Yes, mAP +0.8% to 1.1%.
- Scale invariance: single scale or multi scale?
  - Single scale: faster
  - Multi scale: accurate
  - Best trade-off: single scale

Fast R-CNN: Discussions

- Which layers to fine-tune? $conv_3$ and up.
- Does multi-task training help? Yes, mAP +0.8% to 1.1%.
- Scale invariance: single scale or multi scale? Single.
- Do we need more training data?
  - VOC07: 66.9% -> 70% with augment of VOC12 dataset.

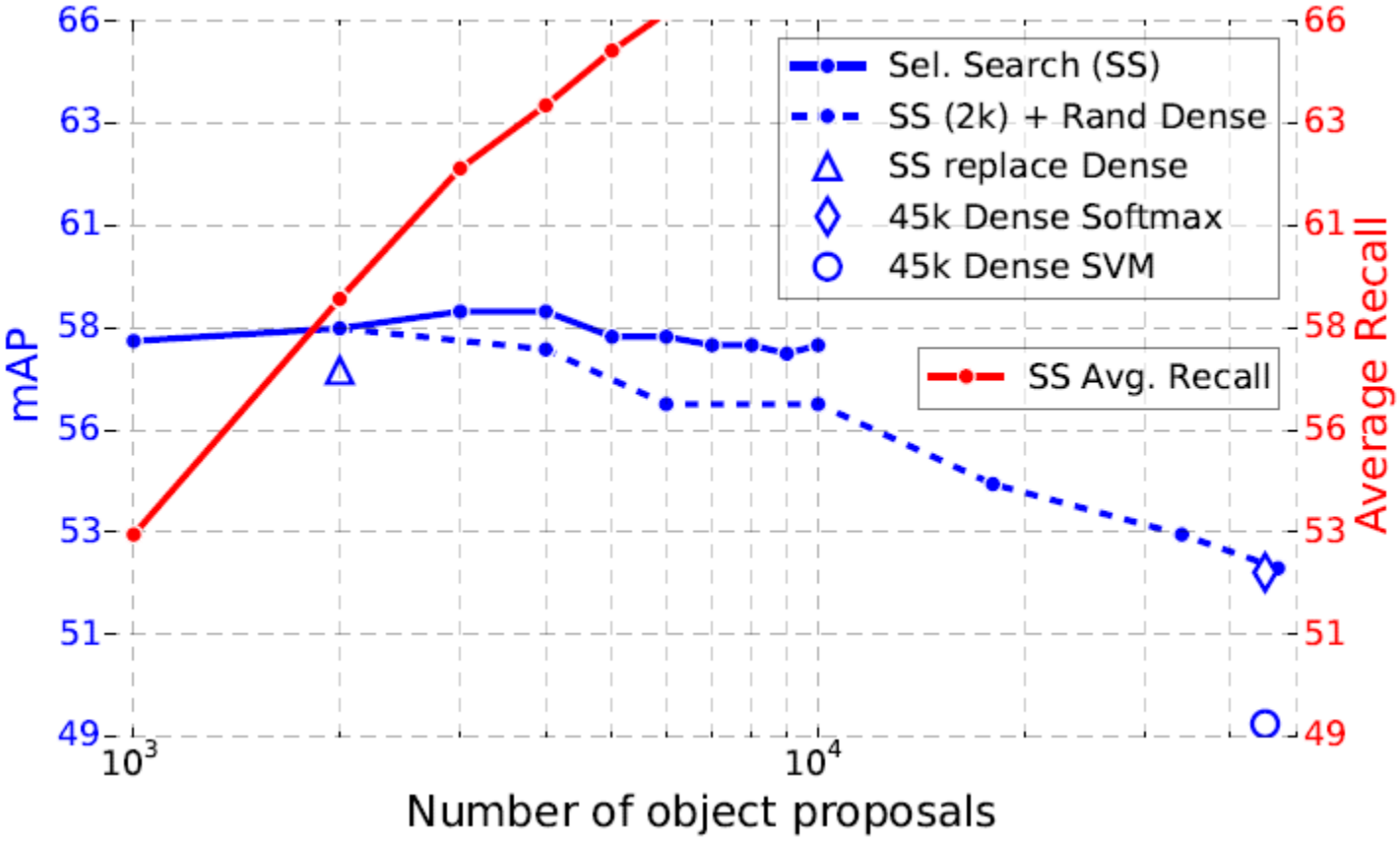
Fast R-CNN: Discussions

- Which layers to fine-tune? $conv_3$ and up.
- Does multi-task training help? Yes, mAP +0.8% to 1.1%.
- Scale invariance: single scale or multi scale? Single.
- Do we need more training data? Yes.
- Do SVMs outperform softmax?
  - SVM: Yes/No; Softmax: 1 vs. all.
  - Softmax: mAP +0.1% to 0.8%, competition between classes.

Fast R-CNN: Discussions

- Which layers to fine-tune? $conv_3$ and up.
- Does multi-task training help? Yes, mAP +0.8% to 1.1%.
- Scale invariance: single scale or multi scale? Single.
- Do we need more training data? Yes.
- Do SVMs outperform softmax? Softmax, mAP +0.1% to 0.8%.
- Are more proposals always better?

Fast R-CNN: Discussions



Girshick. Fast R-CNN

Fast R-CNN: Discussions

- Which layers to fine-tune? $conv_3$ and up.
- Does multi-task training help? Yes, mAP +0.8% to 1.1%.
- Scale invariance: single scale or multi scale? Single.
- Do we need more training data? Yes.
- Do SVMs outperform softmax? Softmax, mAP +0.1% to 0.8%.
- Are more proposals always better? No.

Questions?

Thanks!