Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 8: Convolutional Neural Networks (Part 5)

Logistics:

Assignment 2 …

- Was due last night
- Do not worry if you didn’t finish everything -> look forward
- For the next assignment **start early**

Logistics:

Assignment 3 …

- We will cover intro material today, most relevant content Thursday
- This is objectively THE hardest assignment in the course

Logistics:

Assignment 3 …

- We will cover intro material today, most relevant content Thursday
- This is objectively THE hardest assignment in the course

Hints:

- Read the full assignment and think about what we are asking you to do
- Write down pseudo-code and dimensions of variables on paper
- Test with one sentence for debugging (you should be able to over-fit)
- Plot or print training and validation losses (averaged every X batches)
- Temper your expectations (especially for Part 1)

Logistics:

Project (Survey & Self-defined) …

— Group formations by **October 18th** (there will be Google short form)

— Project pitches: **November 1 & 3**

— Project proposal document: **November 10** (?)

Logistics:

Paper Readings and Presentation Selection

Authors	Title	Venue	Link
J. Lu, V. Goswami, M. Rohrbach, D. Parikh, S. Lee	12-in-1: Multi-Task Vision and Language Representation Learning	CVPR 2020	https://arxiv.org/pdf/1912.02315.pdf
A. Jaegle, F. Gimeno, A. Brock	Perceiver: General Perception with Iterative Attention	ICML 2021	https://arxiv.org/pdf/2103.03206.pdf
A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A	PerceiverIO: A General Architccture for Structured Inputs and Outputs	ICLR 2022	https://arxiv.org/pdf/2107.14795.pdf
C. Lin, Y. Jiang, J. Cai, L. Qu, G. Haffari, Z. Yuan	Multimodal Transformer with Variable-length Memory for Vision-and-Language Navigation		https://arxiv.org/pdf/2111.05759.pdf
S. Chen, P.-L. Guhur, C. Schmid, I. Laptev	History Aware Multimodal Transformer for Vision-and-Language Navigation	NeurIPS 2021	https://arxiv.org/pdf/2110.13309.pdf
R. Bachmann, D. Mizrahi, A. Atanov, A. Zamir	MultiMAE: Multi-modal Multi-task Masked Autoencoders	ECCV 2022	https://arxiv.org/pdf/2204.01678.pdf
A. Yang, A. Miech, J. Sivic, I. Laptev, C. Schmid	TubeDETR: Spatio-Temporal Video Grounding with Transformers	CVPR 2022	https://arxiv.org/pdf/2203.16434.pdf
M. Li, R. Xu, S. Wang, L. Zhou, X. Lin, C. Zhu, M. Zeng, H. Ji, S.-F. Chang	CLIP-Event: Connecting Text and Images with Event Structures	CVPR 2022	https://arxiv.org/pdf/2201.05078.pdf
T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, C.Ross	Winoground: Probing Vision and Language Models for Visio Linguistic Compositionality	CVPR 2022	https://arxiv.org/pdf/2204.03162.pdf
J. Andreas, M. Rohrbach, T. Darrell, D. Klein	Neural module networks	CVPR 2016	https://arxiv.org/pdf/1511.02799.pdf
B. Zhao, B. Chang, Z. Jie, L. Sigal	Modular Generative Adversarial Networks	ECCV 2018	https://arxiv.org/pdf/1804.03343.pdf
R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko	Learning to reason: End-to-end module networks for visual question answering	ICCV 2017	https://openaccess.thecvf.com/content_ICCV_20
J.Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, R. Girshick	Inferring and Executing Programs for Visual Reasoning	ICCV 2017	https://arxiv.org/pdf/1705.03633.pdf
A. Das, S. Kottur, J. Moura, S. Lee, D. Batra	Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning	ICCV, 2017	https://arxiv.org/pdf/1703.06585.pdf
A. Maharana, D. Hannan, M. Bansal	StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation	ECCV 2022	https://arxiv.org/pdf/2209.06192.pdf
Y.-L. Sung, J. Cho, M.Bansal	VL-ADAPTER: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks	CVPR 2022	https://arxiv.org/pdf/2112.06825.pdf
R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, Y. Choi	MERLOT Reserve: Multimodal Neural Script Knowledge through Vision and Language and S	CVPR 2022	https://arxiv.org/pdf/2201.02639.pdf
J. Hessel, J. Hwang, J. S. Park , R. Zellers, C. Bhagavatula, A. Rohrbach, K. Saenko, Y. Choi	The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning	ECCV 2022	https://arxiv.org/pdf/2202.04800.pdf
Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler	Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and	ICCV 2015	https://arxiv.org/pdf/1506.06724.pdf
E. Perez, F. Strub, H. Vries, V. Dumoulin, A. Courville	FiLM: Visual Reasoning with a General Conditioning Layer	AAAI 2018	https://arxiv.org/pdf/1709.07871.pdf
	CLIP		
A. Douillard, A. Rame, G. Couairon, M. Cord	DyTox: Transformers for Continual Learning with DYnamic TOken eXpansion	CVPR 2022	https://openaccess.thecvf.com/content/CVPR202
N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. Feris, D. Harwath, J. Gla	Everything at Once – Multi-modal Fusion Transformer for Video Retrieval	CVPR 2022	https://openaccess.thecvf.com/content/CVPR202
J. Lu, C. Clark, R. Zellers, R. Mottaghi, A. Kembhavi	Unified-IO: A Unified Model for Vision, Language and Multi-modal Tasks		https://arxiv.org/pdf/2206.08916.pdf
	Video Diffusion Models		
TBD	PHENAKI: VARIABLE LENGTH VIDEO GENERATION FROM OPEN DOMAIN TEXTUAL DI	ICLR 2023	https://openreview.net/pdf?id=vOEXS39nOF
P. Seo, A. Lehrmann, B. Han, L. Sigal	Visual Reference Resolution using Attention Memory for Visual Dialog	NeurIPS 2017	
C. Xiong, S. Merity, R. Socher	Dynamic memory networks for visual and textual question answering	ICML 2016	
M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. Boyd-Graber, H. Daume, L. Davis	The Amazing Mysteries of the Gutter: Drawing Inferences between Panels in Comic Book Na	CVPR 2017	
E. Cubuk, B. Zoph, D. Mané, V. Vasudevan, Q. Le	AutoAugment: Learning Augmentation Policies from Data		

Computer **Vision Problems** (no language for now)

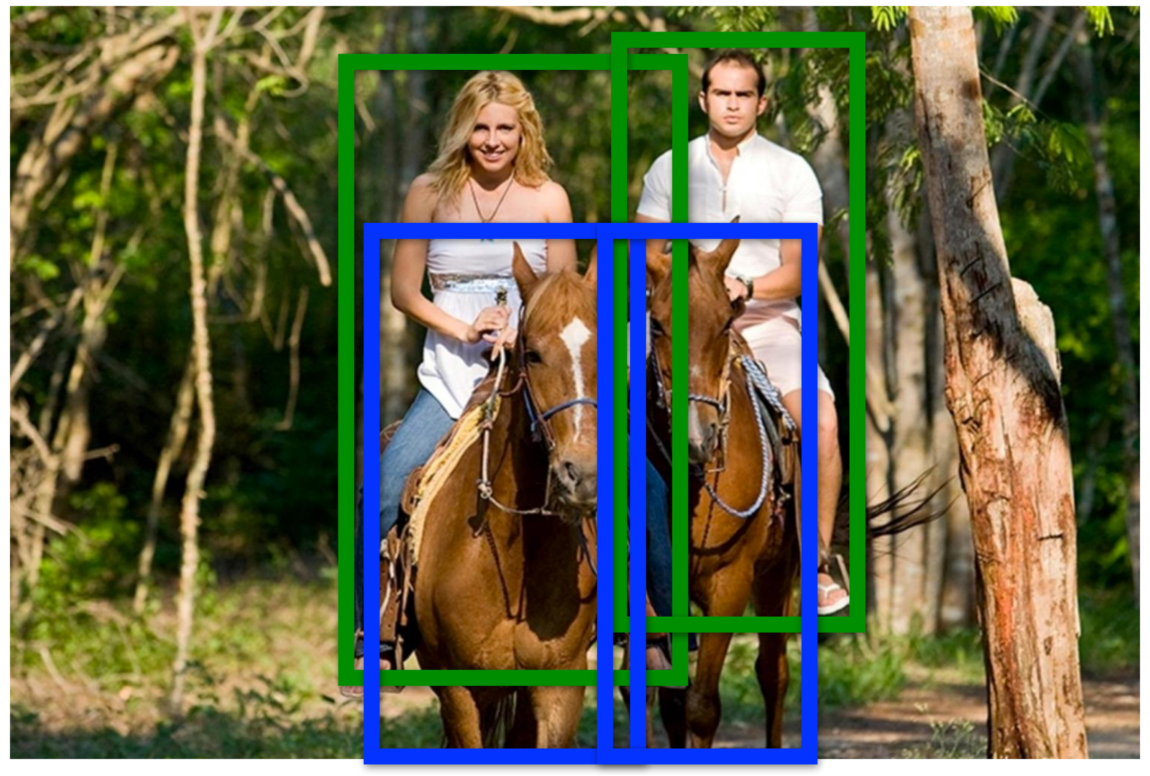
Categorization

Multi-**class**: Horse
Church
Toothbrush
Person

IMAGENET

Multi-**label**: **Horse**
Church
Toothbrush
Person

Detection



Horse (x, y, w, h)
Horse (x, y, w, h)
Person (x, y, w, h)
Person (x, y, w, h)

COCO Common Objects in Context

Segmentation

Horse
Person

COCO Common Objects in Context

Instance Segmentation

Horse1
Horse2
Person1
Person2

Computer **Vision Problems** (no language for now)

Categorization

Computer **Vision Problems** (no language for now)

Categorization

Multi-**class:** Horse
Church
Toothbrush
Person

IMAGENET

Computer **Vision Problems** (no language for now)

Categorization

Multi-**class**: Horse
Church
Toothbrush
Person

IMAGENET

Multi-**label**: **Horse**
Church
Toothbrush
Person

Computer **Vision Problems** (no language for now)

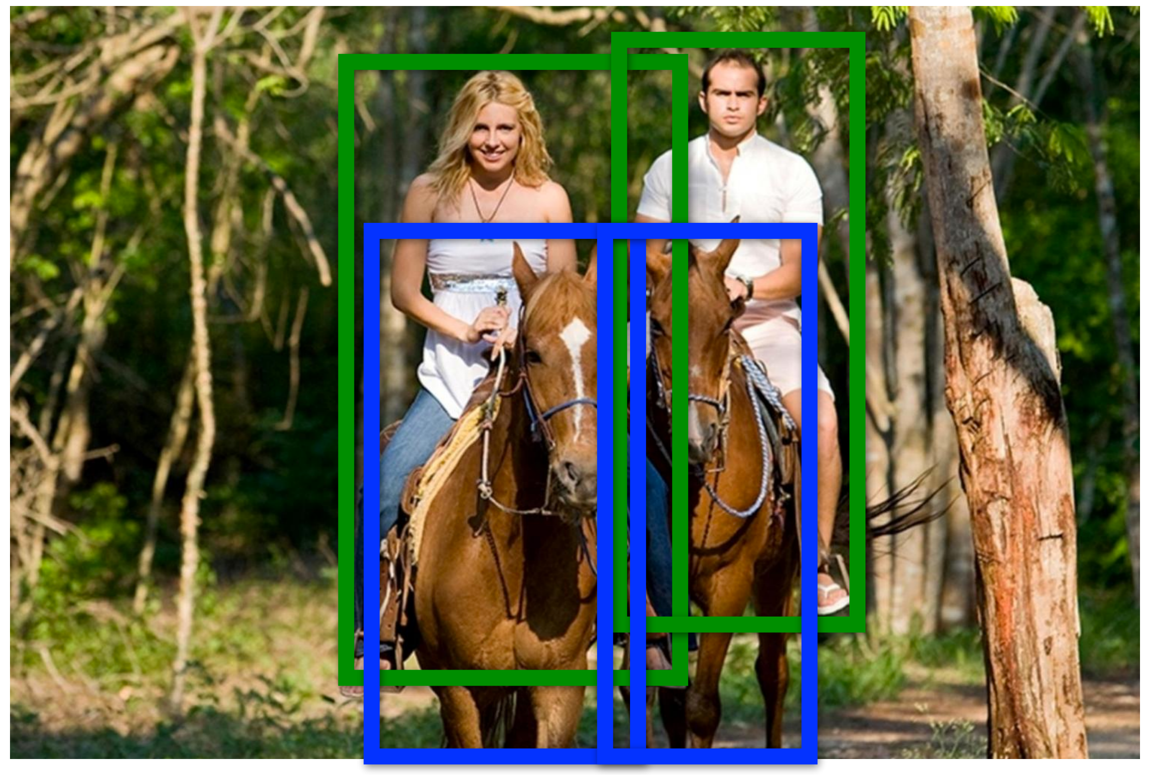
Segmentation

Horse
Person

COCO
Common Objects in Context

Computer **Vision Problems** (no language for now)

Detection



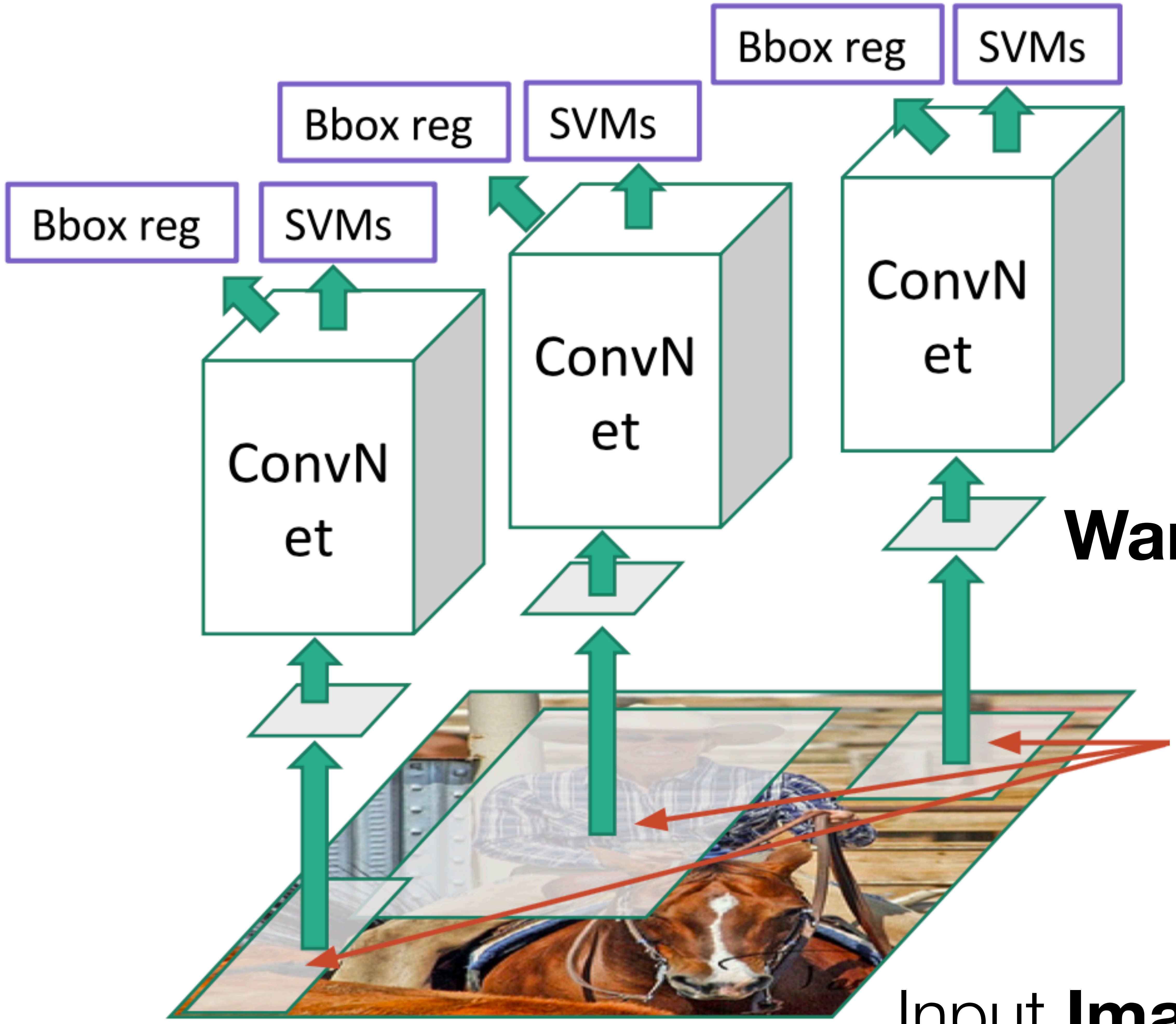
Horse (x, y, w, h)
Horse (x, y, w, h)
Person (x, y, w, h)
Person (x, y, w, h)

COCO
Common Objects in Context

R-CNN

Linear Regression for bounding box offsets

[Girshick et al, CVPR 2014]



Classify regions with SVM

Forward each region through a **CNN**

Warped image regions

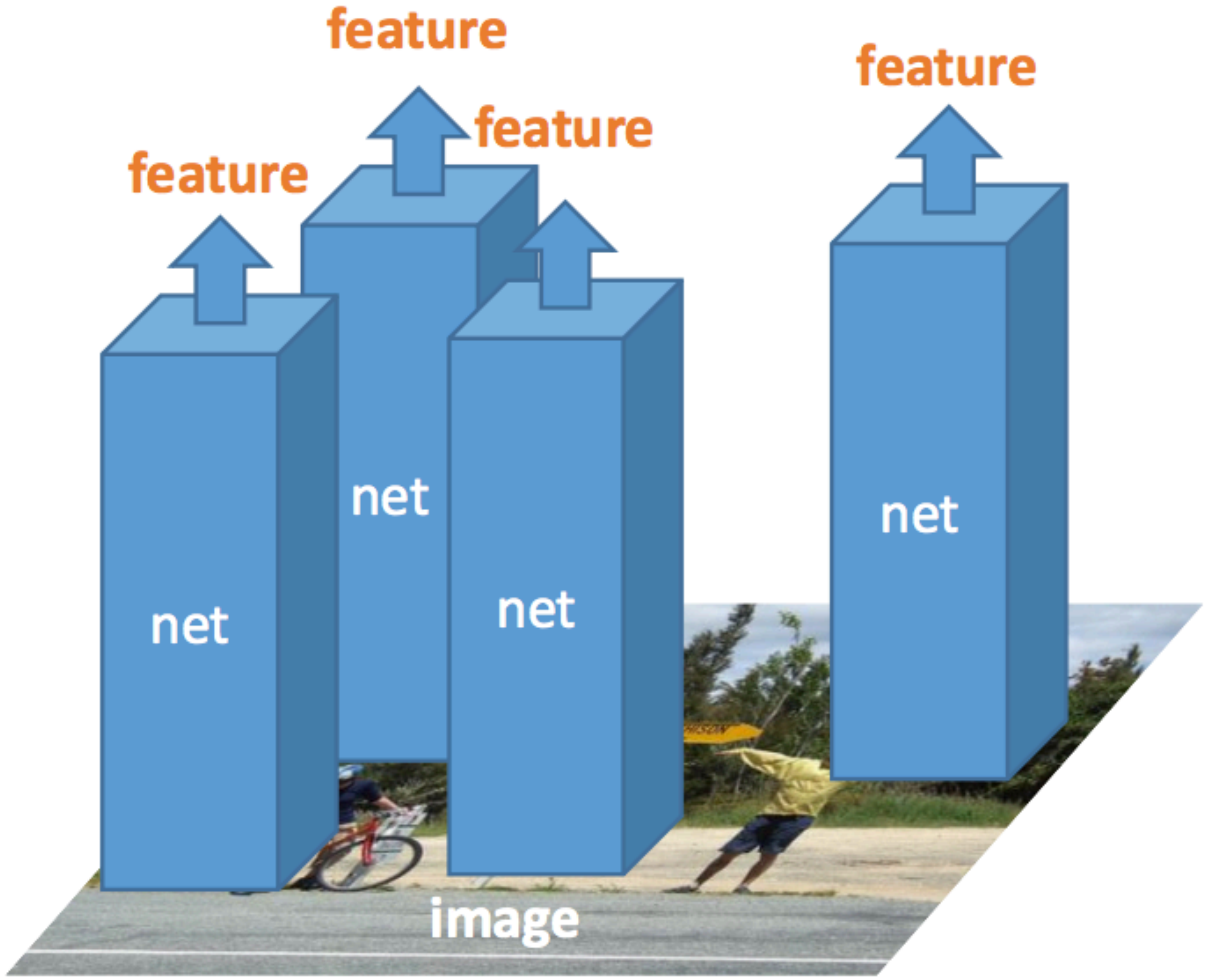
Regions of Interest from a proposal method (~2k)

Input **Image**

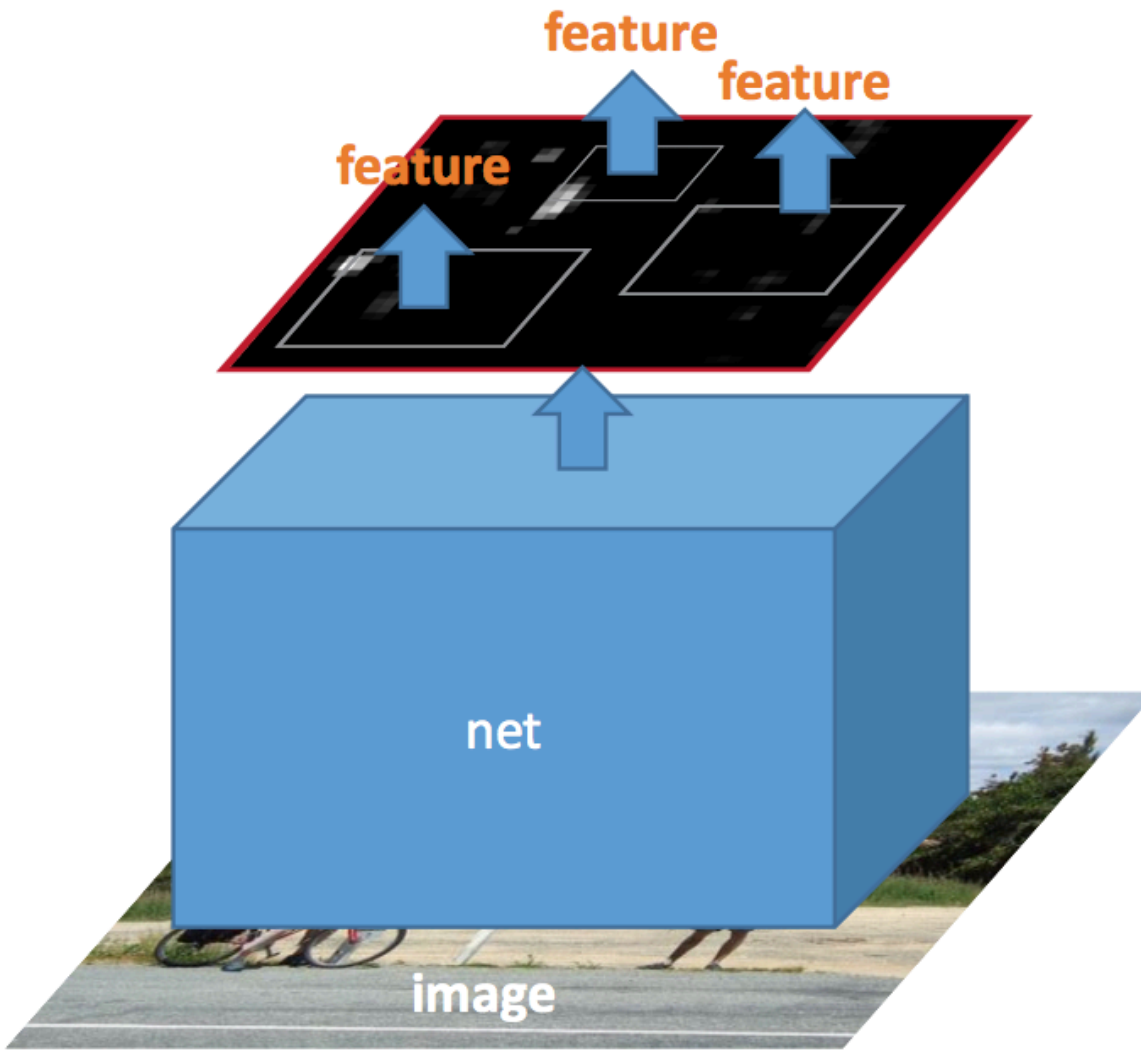
* image from Ross Girshick

R-CNN vs. SPP

[He et al, ECCV 2014]



R-CNN
2000 nets on image regions



SPP-net
1 net on full image

Fast **R-CNN**

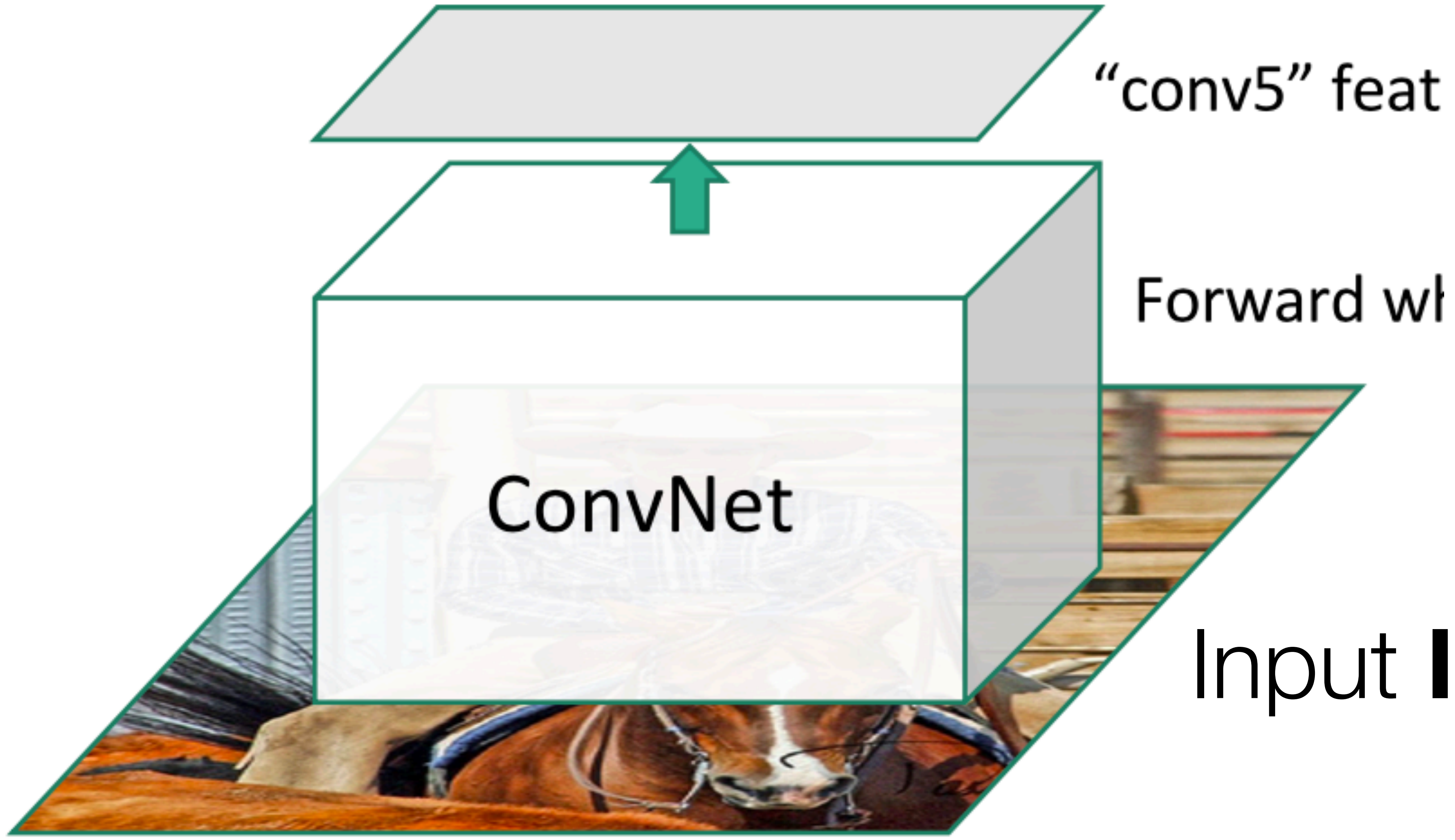
[Girshick et al, ICCV 2015]

Input **Image**

* image from Ross Girshick

Fast **R-CNN**

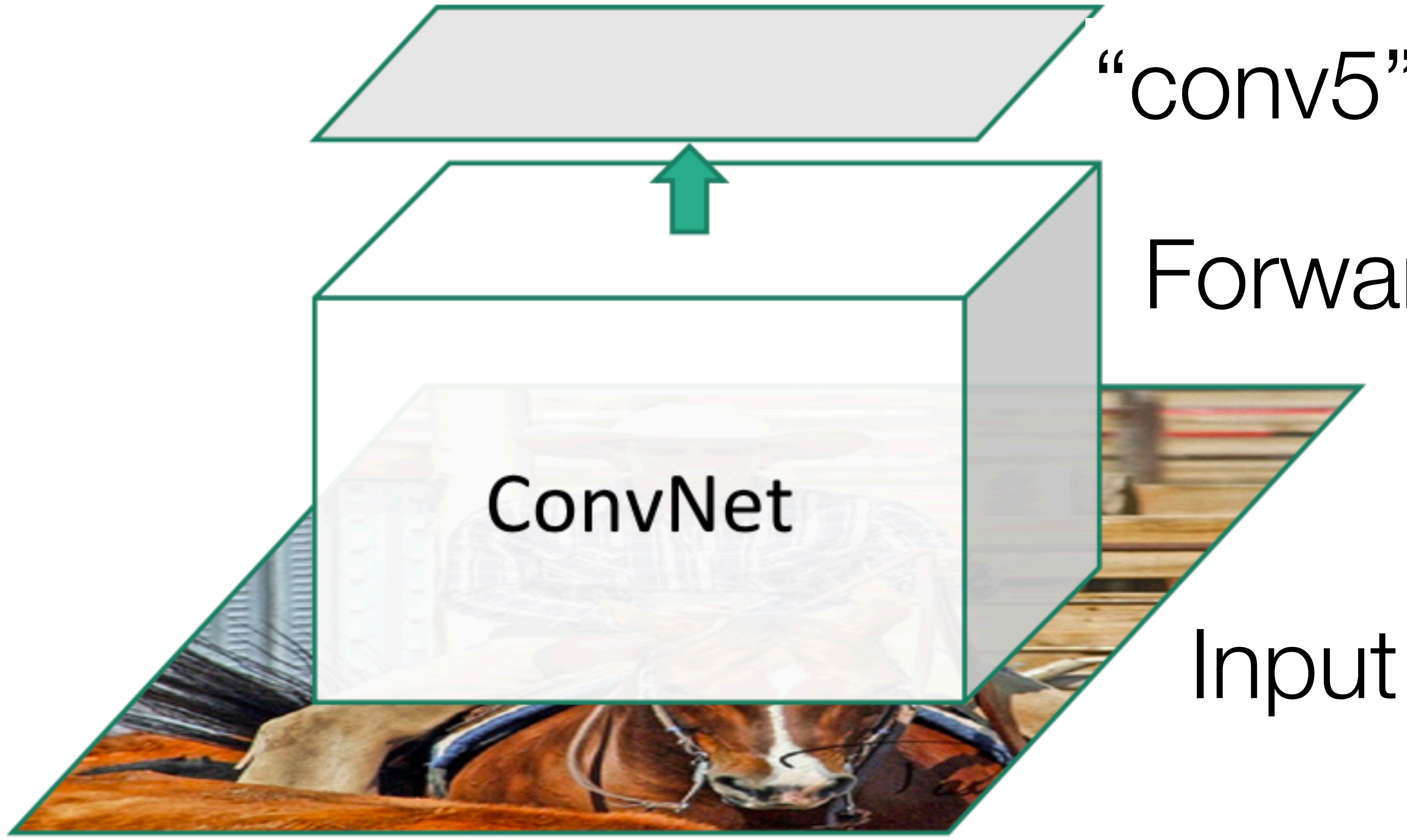
[Girshick et al, ICCV 2015]



* image from Ross Girshick

Fast **R-CNN**

[Girshick et al, ICCV 2015]



"conv5" feature map

Forward prop the **whole image** through CNN

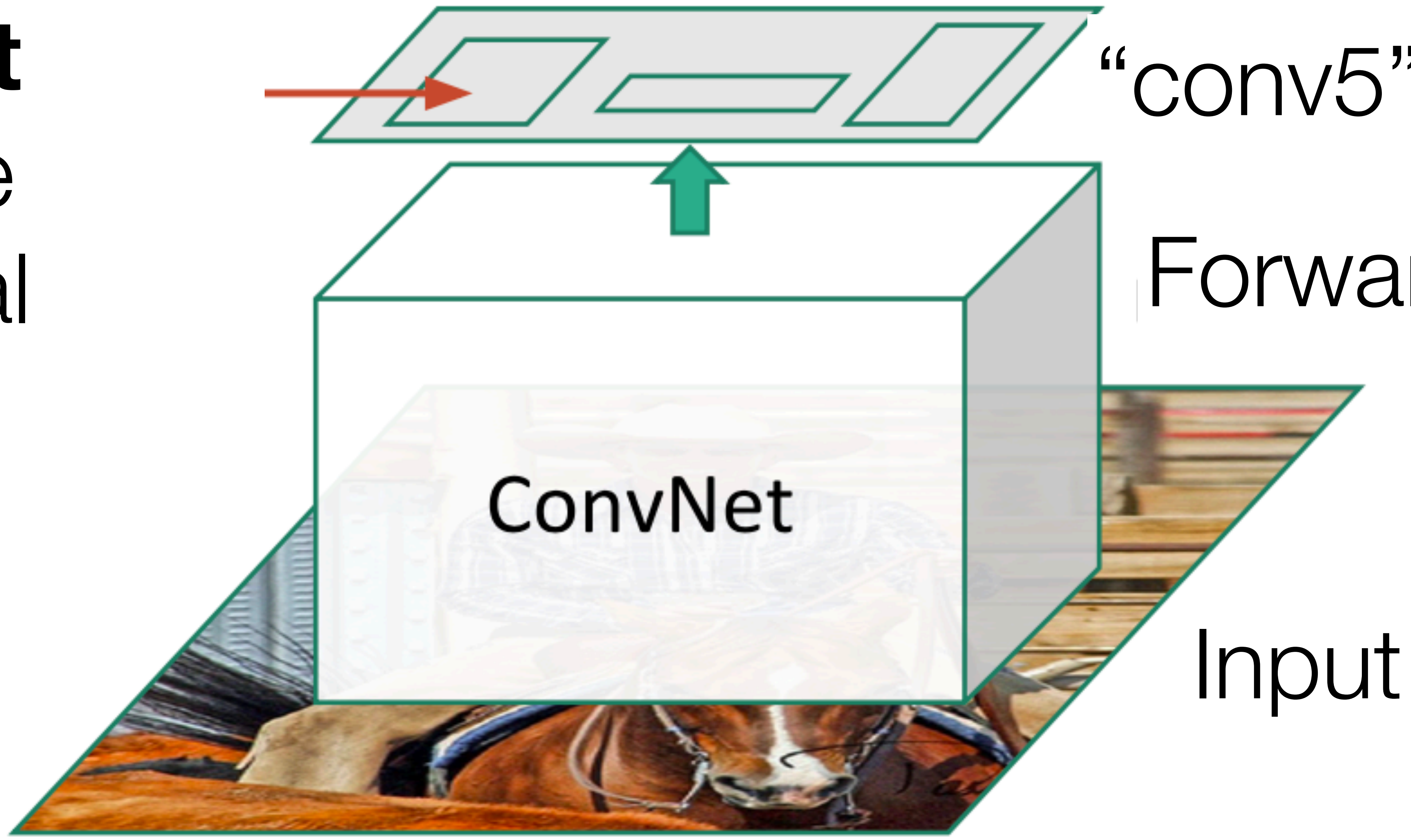
Input **Image**

* image from Ross Girshick

Fast **R-CNN**

[Girshick et al, ICCV 2015]

Regions of Interest from the proposal method



“conv5” feature map

Forward prop the **whole image** through CNN

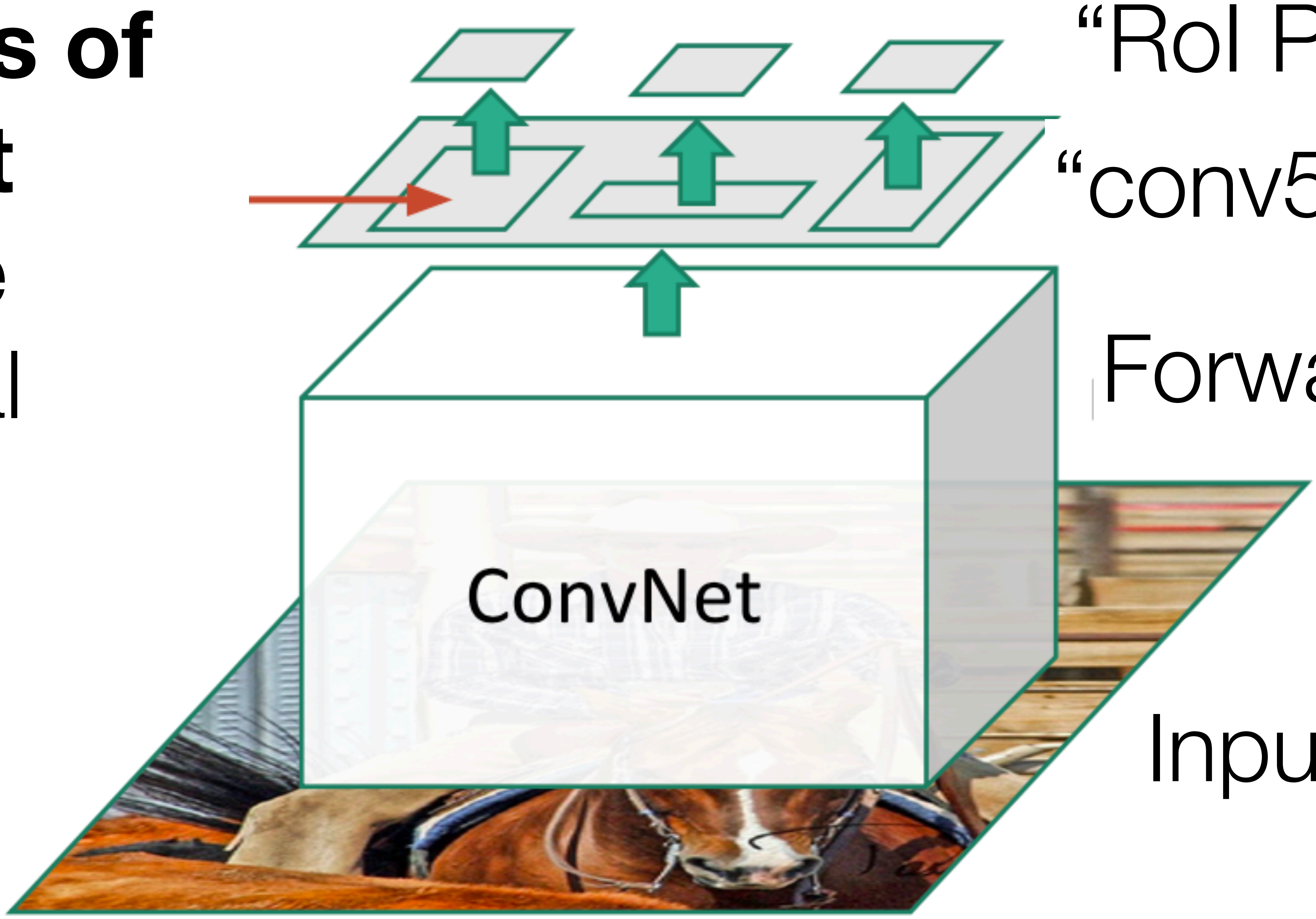
Input **Image**

* image from Ross Girshick

Fast **R-CNN**

[Girshick et al, ICCV 2015]

Regions of Interest from the proposal method



"RoI Pooling" layer

"conv5" feature map

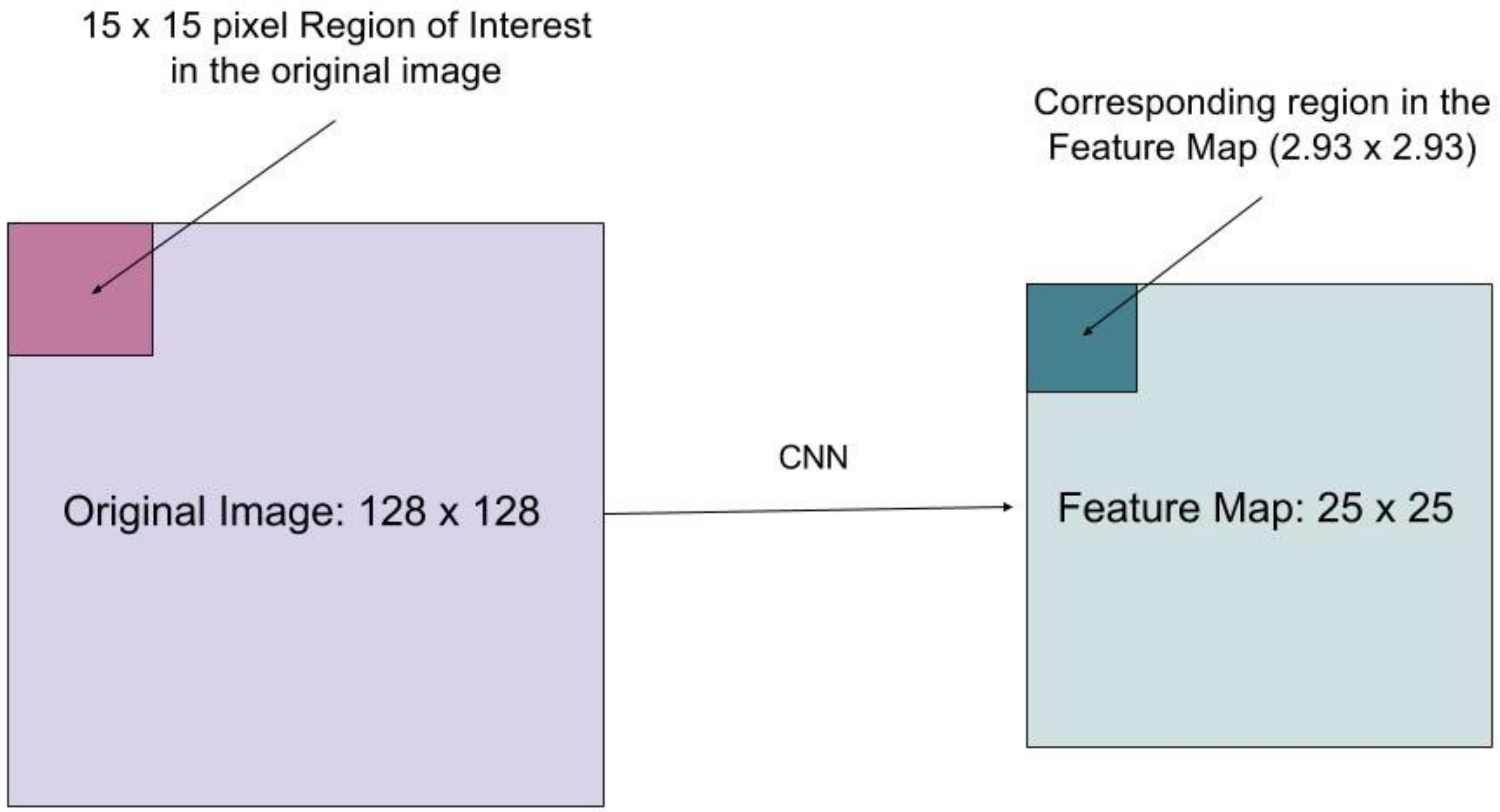
Forward prop the **whole image** through CNN

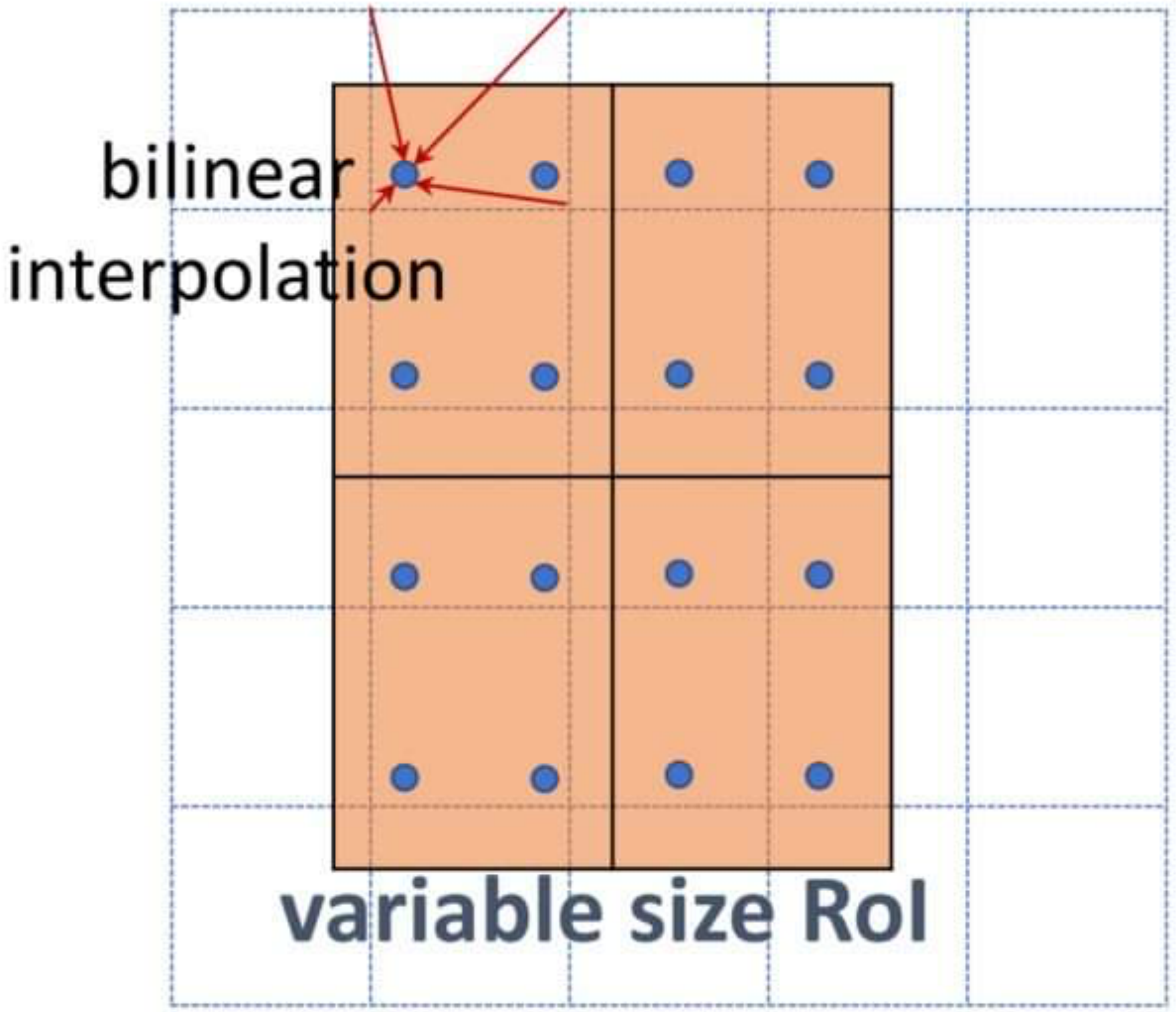
Input **Image**

Girshick, "Fast R-C
Figure copyright R

* image from Ross Girshick

RoI Align


15 x 15 pixel Region of Interest
in the original image
Corresponding region in the
Feature Map (2.93 x 2.93)
Original Image: 128 x 128
CNN
Feature Map: 25 x 25


bilinear
interpolation
variable size RoI

Fast **R-CNN**

[Girshick et al, ICCV 2015]

Object classification

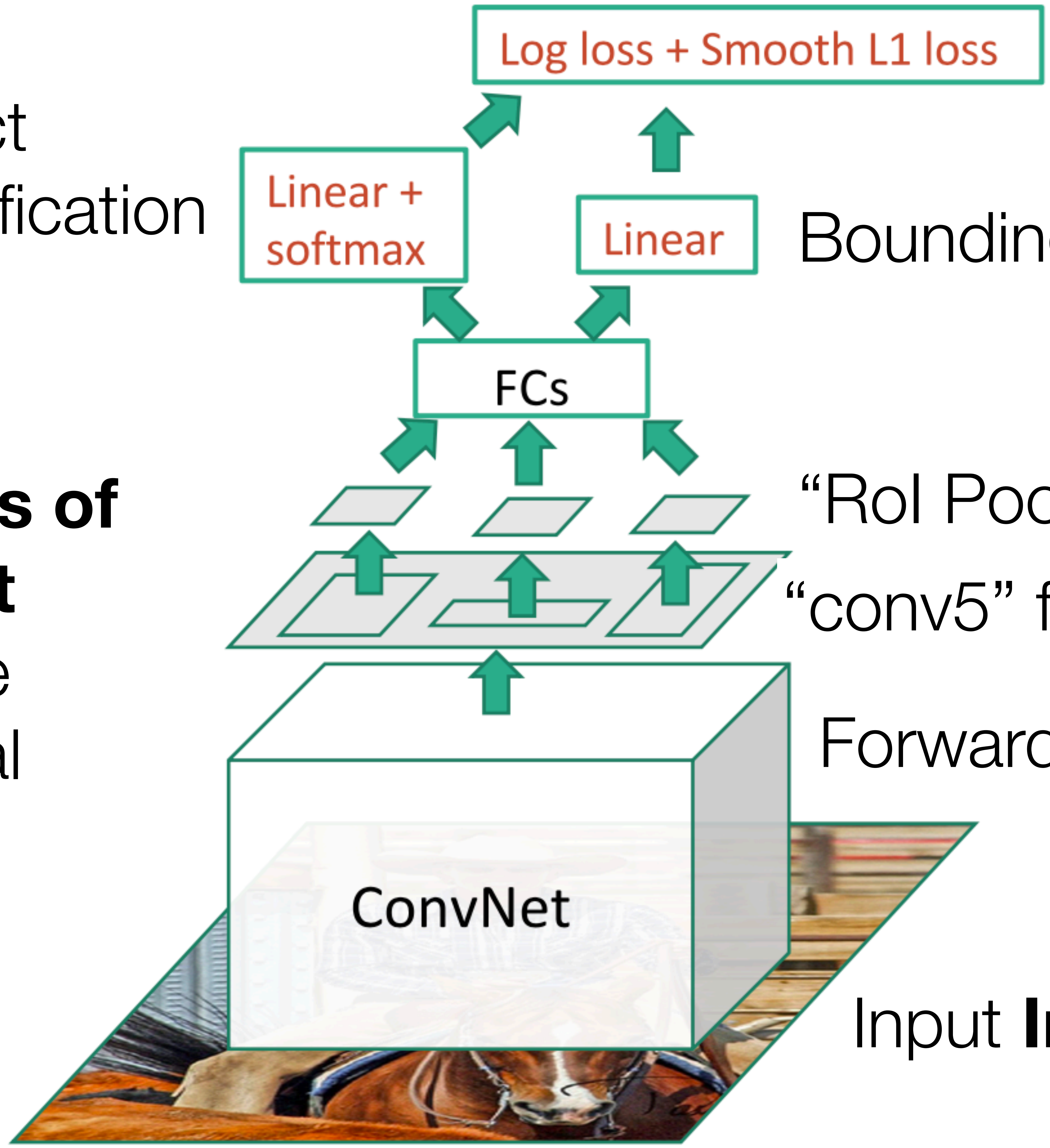
Multi-task loss

Bounding box regression

“RoI Pooling” layer

“conv5” feature map

Forward prop the **whole image** through CNN



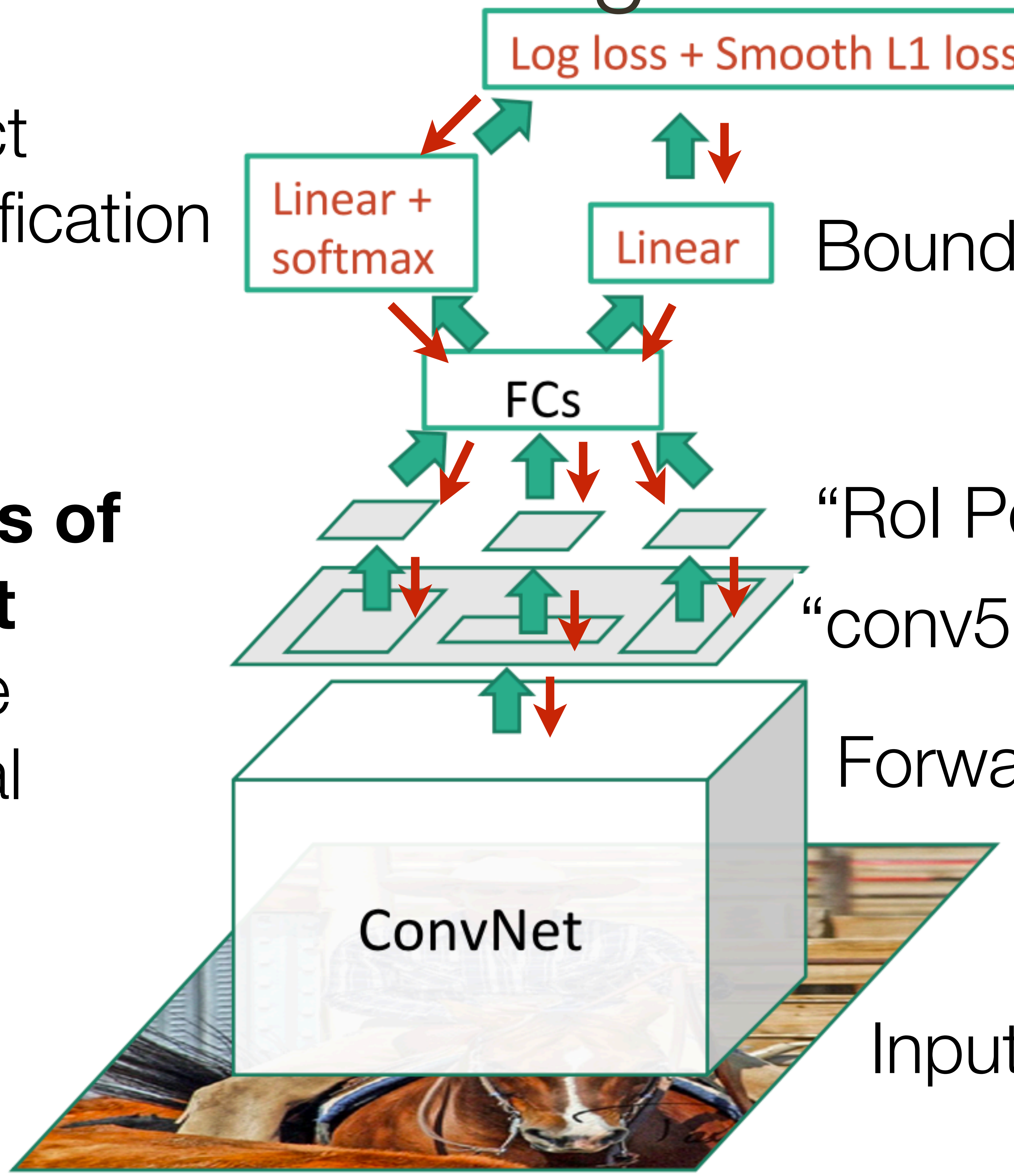
Regions of Interest from the proposal method

Input **Image**

* image from Ross Girshick

Fast **R-CNN:** Training

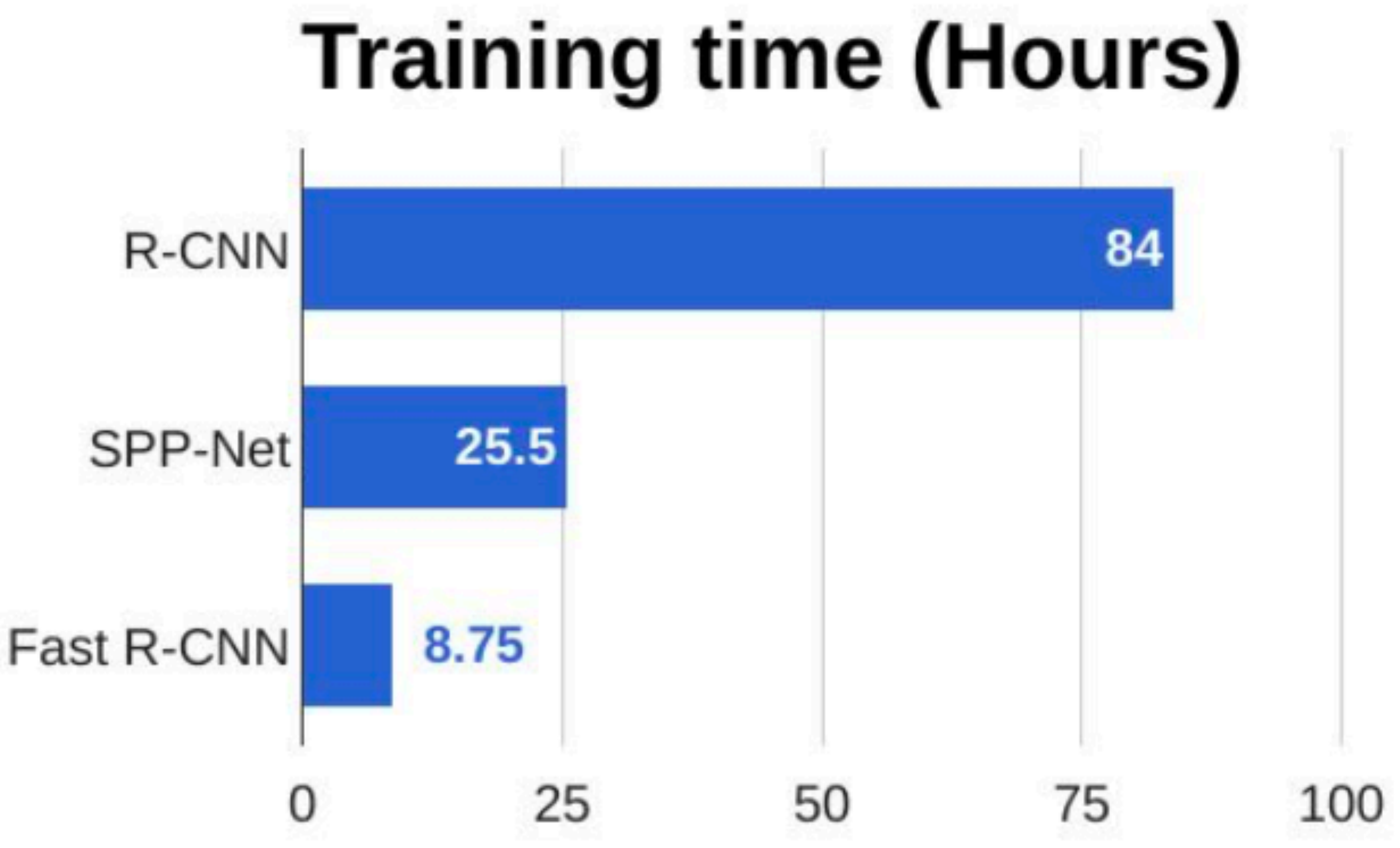
[Girshick et al, ICCV 2015]

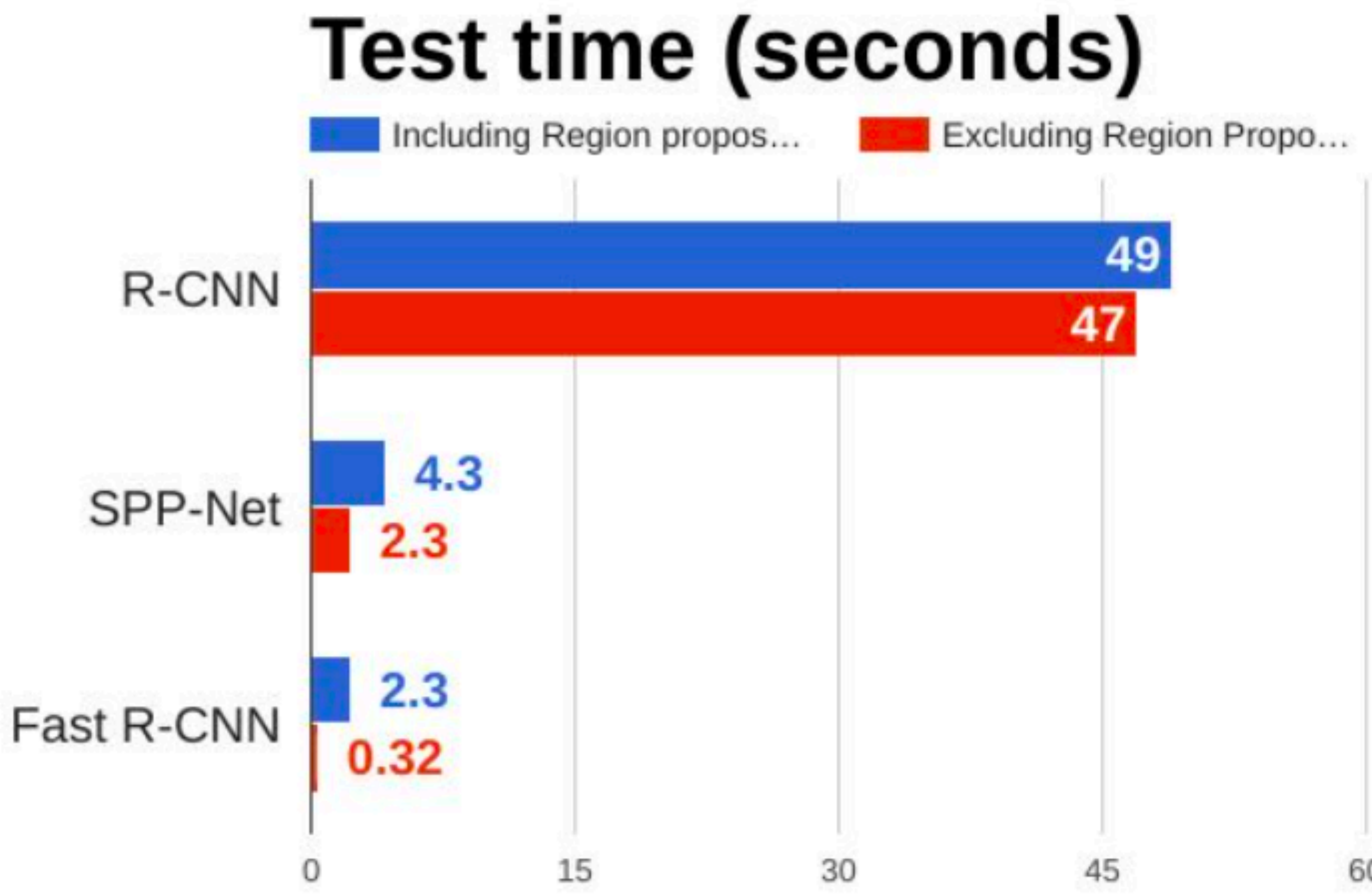


* image from Ross Girshick

R-CNN vs. SPP vs. Fast R-CNN

[Girshick et al, CVPR 2014]
[Girshick et al, ICCV 2015]
[He et al, ECCV 2014]

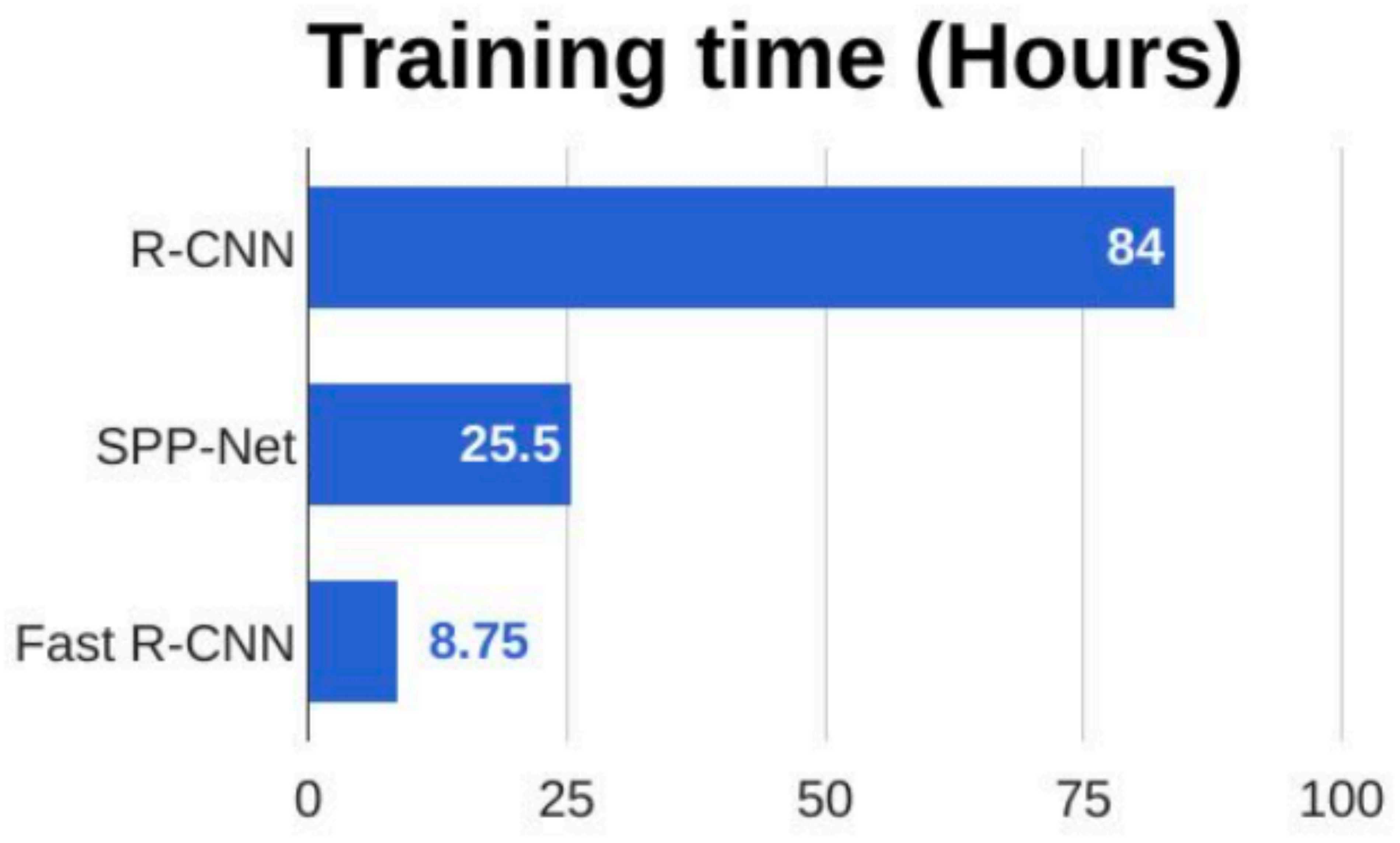


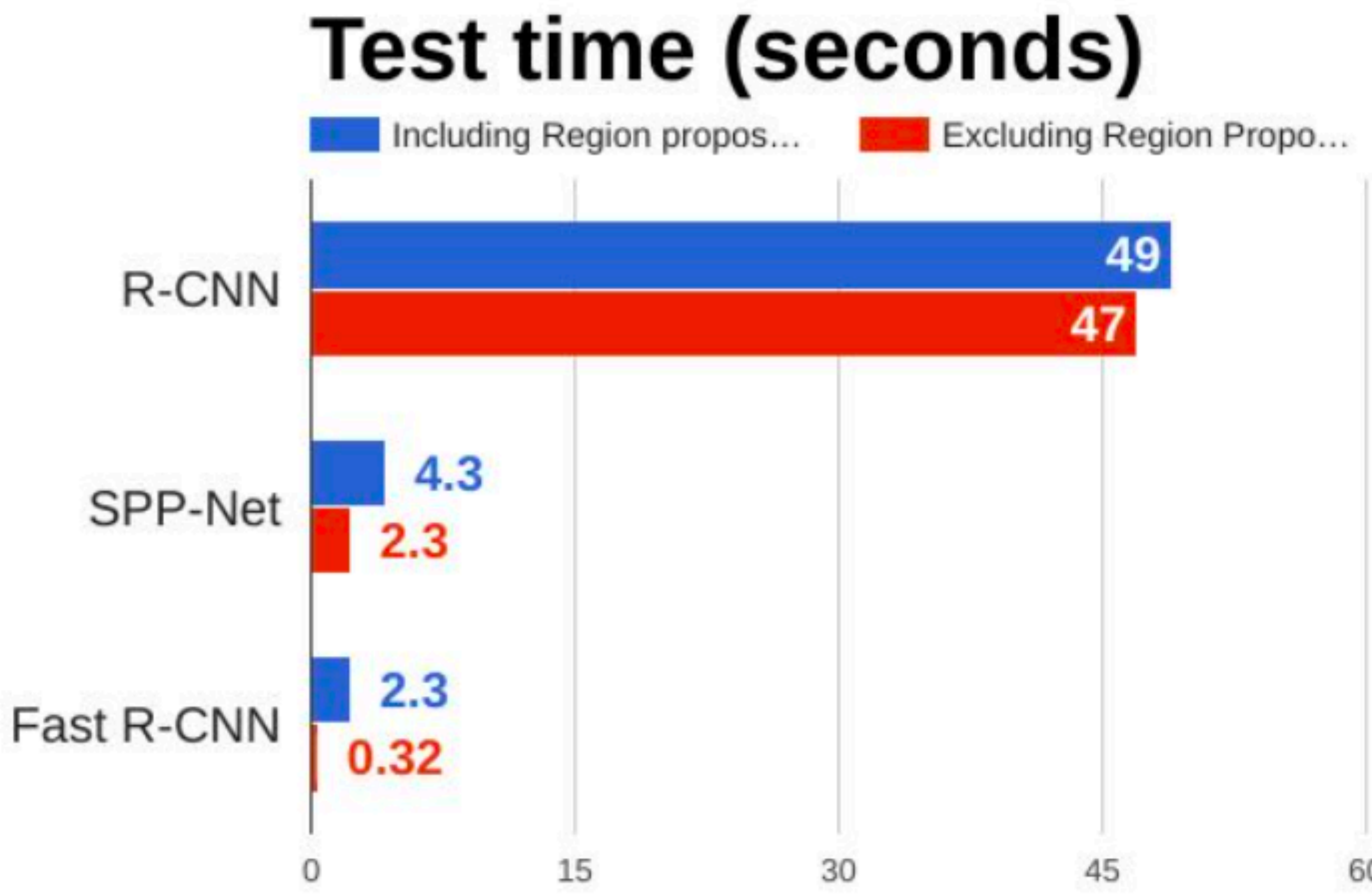


* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

R-CNN vs. **SPP** vs. **Fast R-CNN**

[Girshick et al, CVPR 2014]
[Girshick et al, ICCV 2015]
[He et al, ECCV 2014]





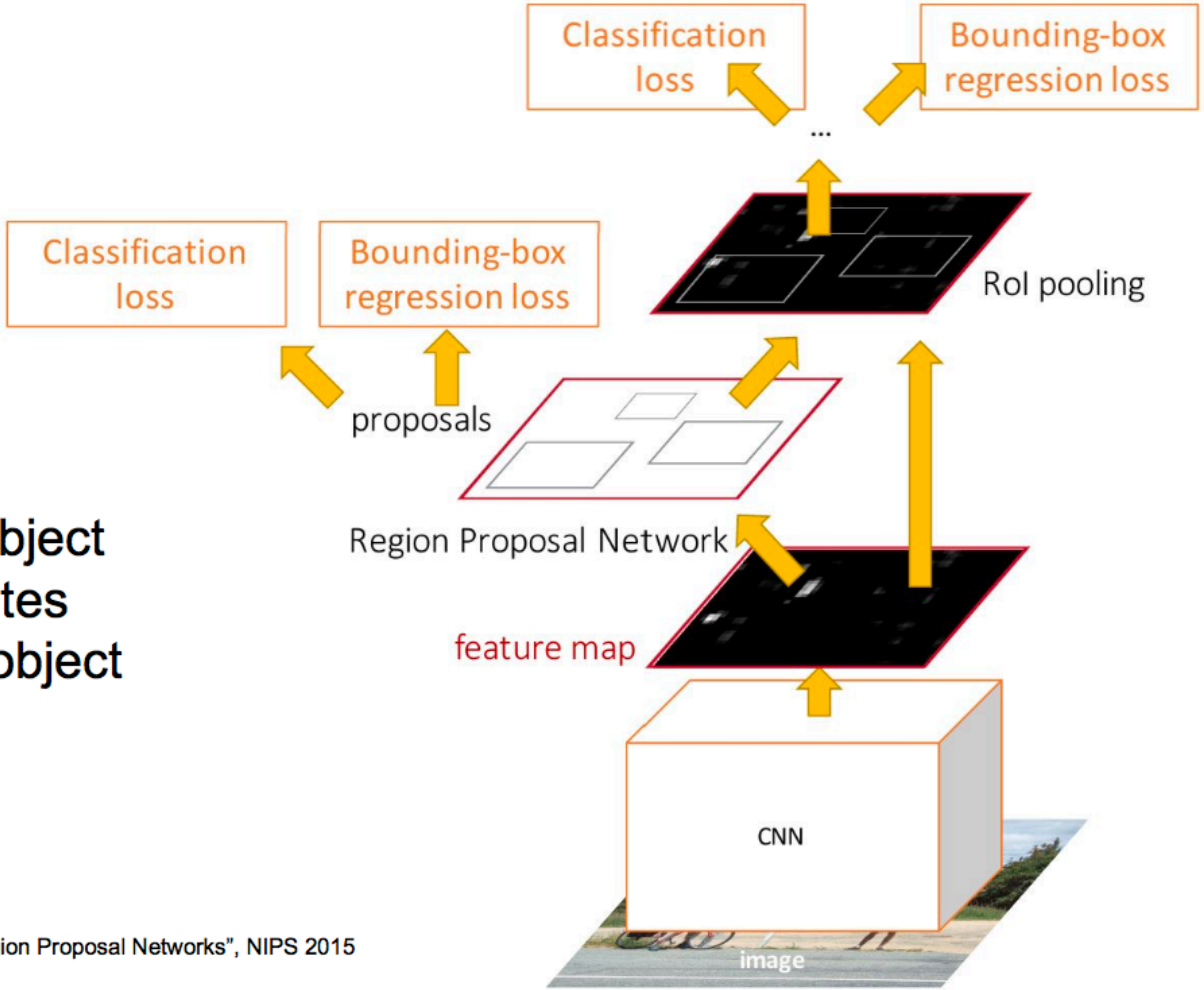
Observation: Performance dominated by the region proposals at this point!

Faster **R-CNN**

Make CNN do proposals!

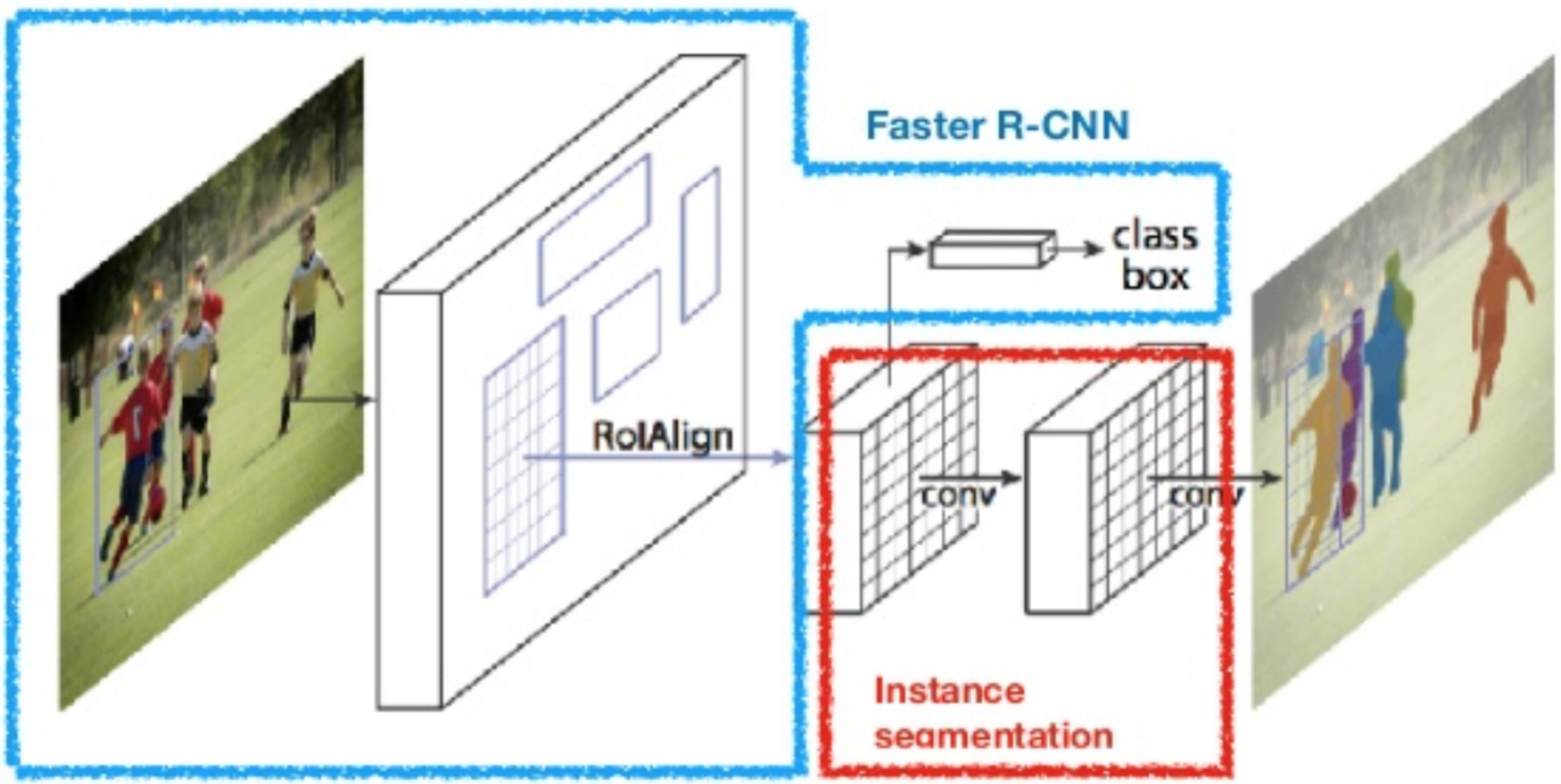
Insert **Region Proposal Network (RPN)** to predict proposals from features

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

Mask R-CNN

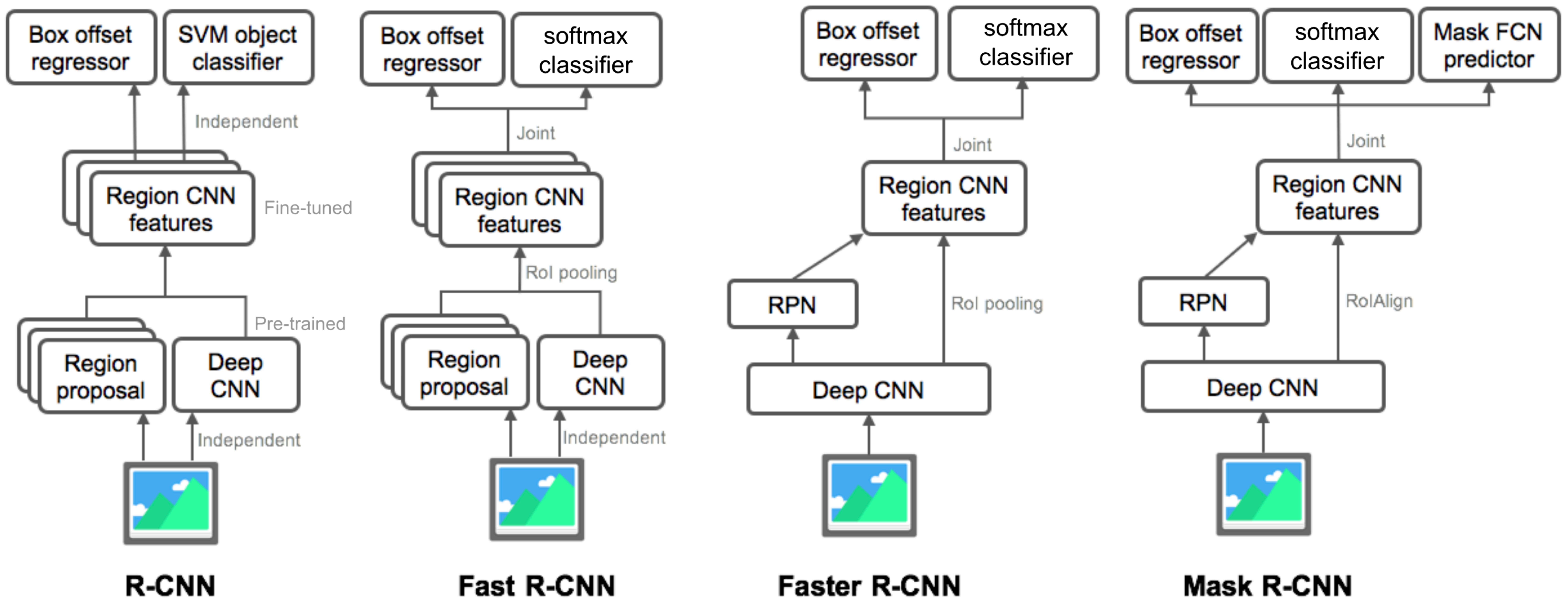


[He et al, 2017]

Mask R-CNN

[He et al, 2017]

Summary of R-CNN Family of Models

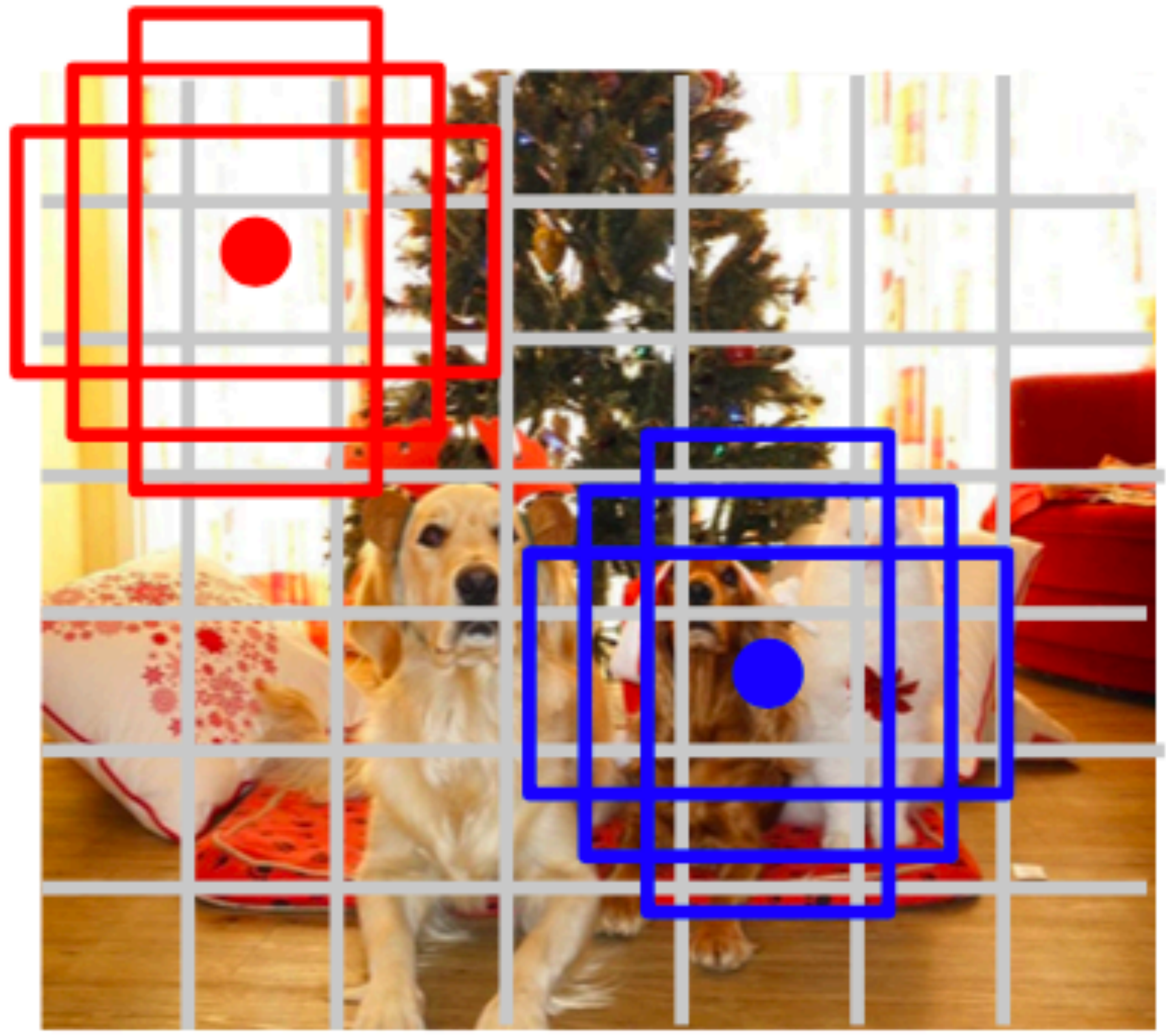


https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html

YOLO: You Only Look Once

[Redmon et al, CVPR 2016]

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

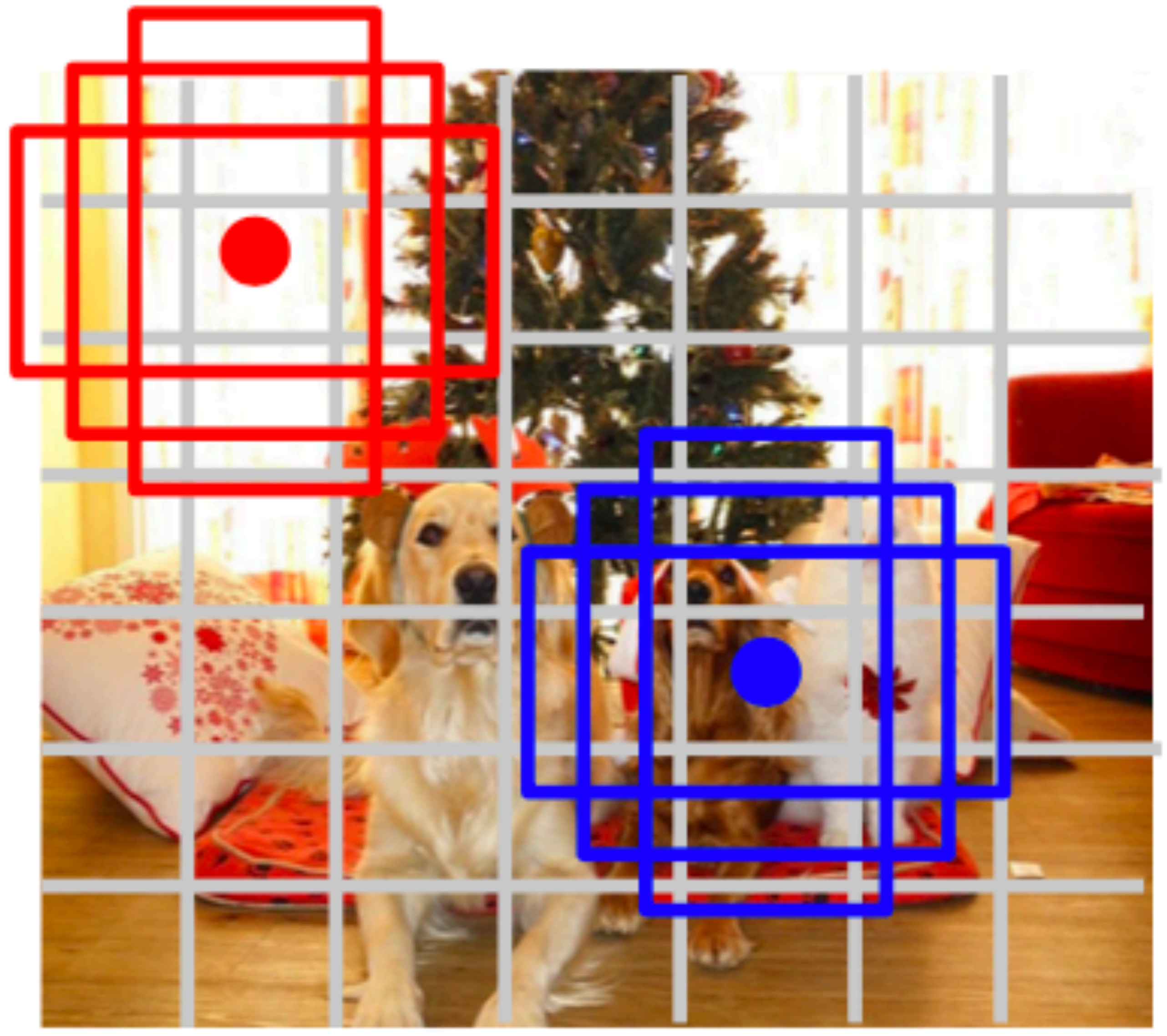
Within each grid cell:
- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

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

YOLO: You Only Look Once

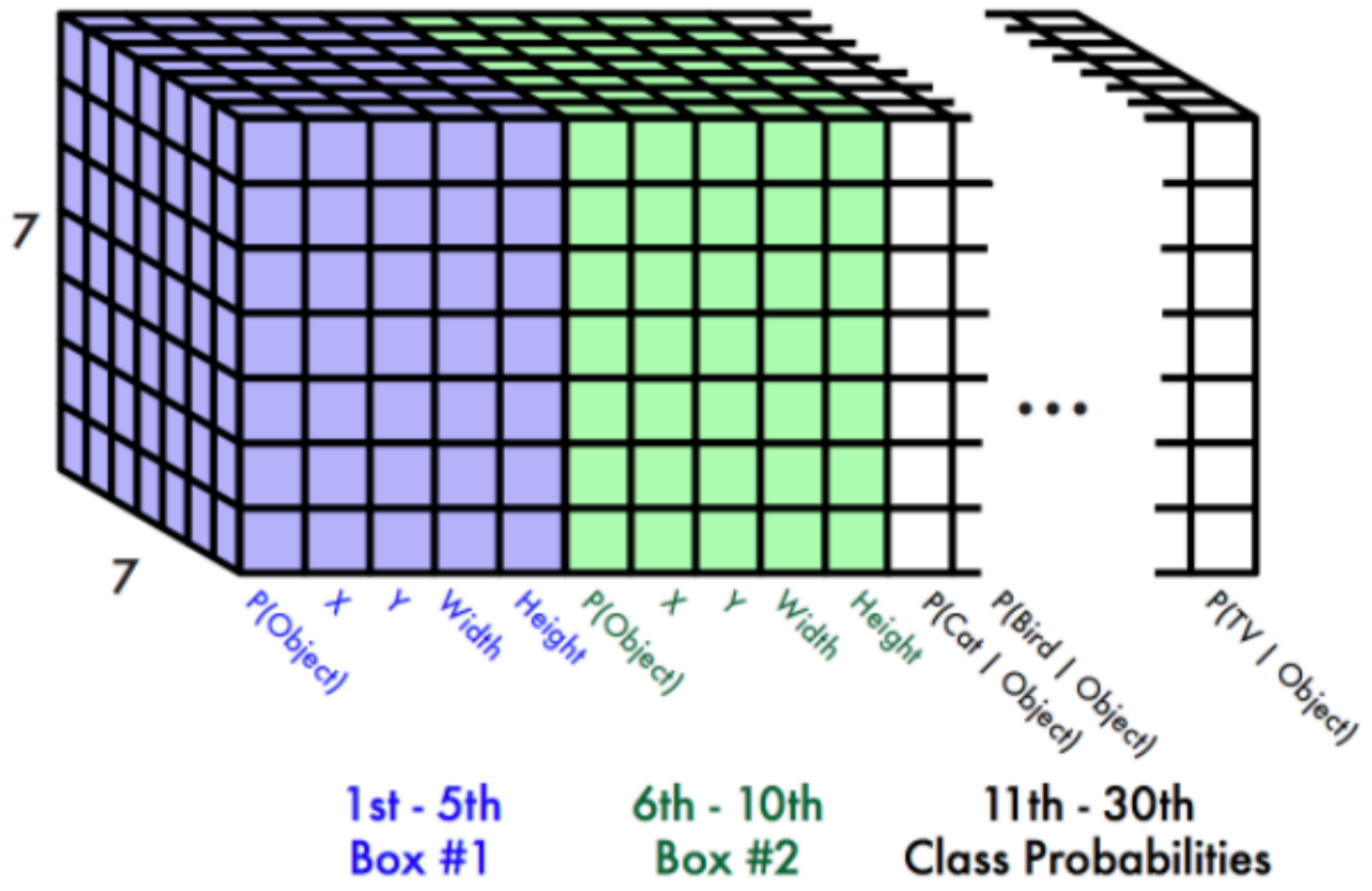
[Redmon et al, CVPR 2016]

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



YOLO v2

http://pjreddie.com/yolo

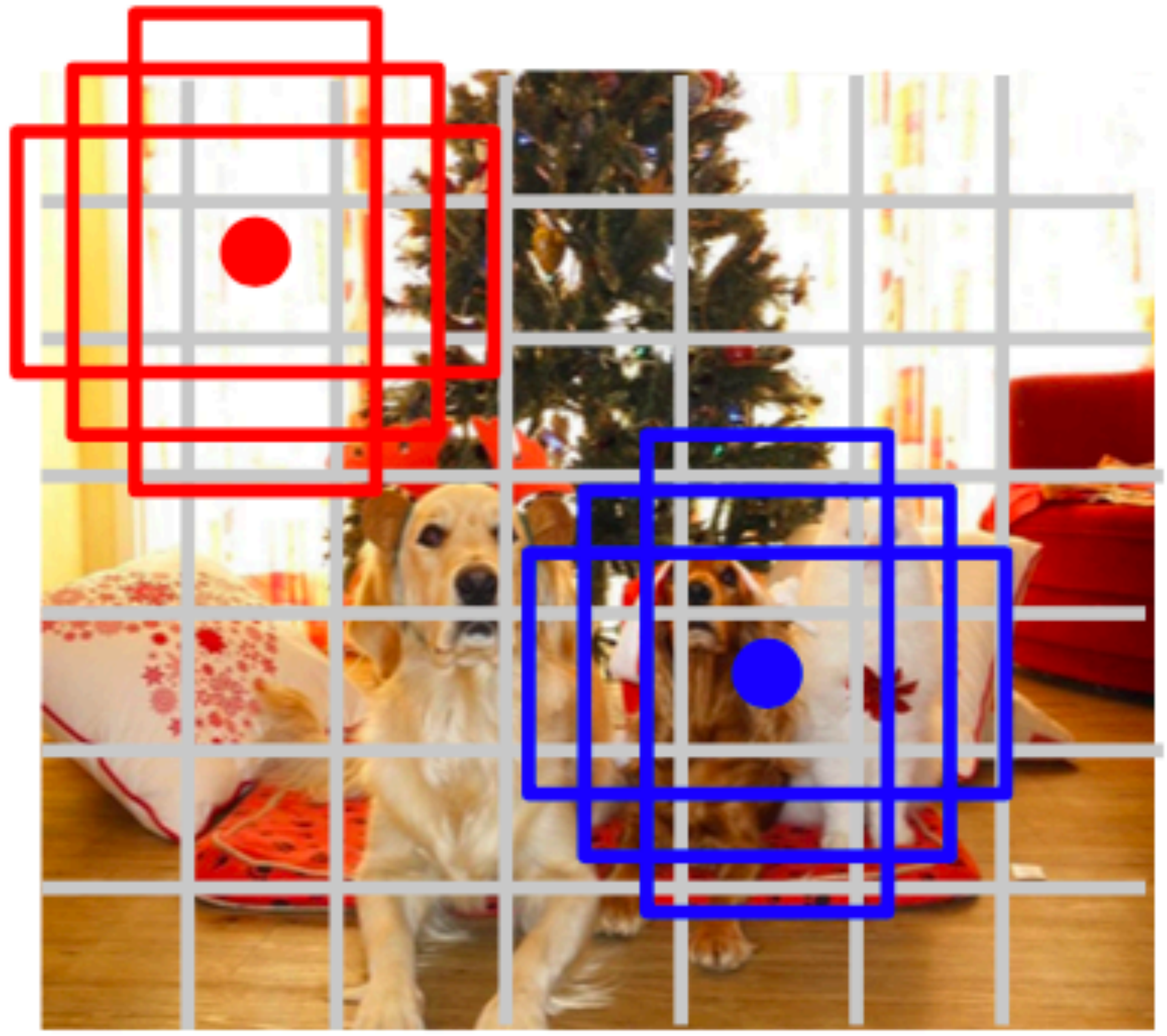
YOLO v2

http://pjreddie.com/yolo

YOLO: You Only Look Once

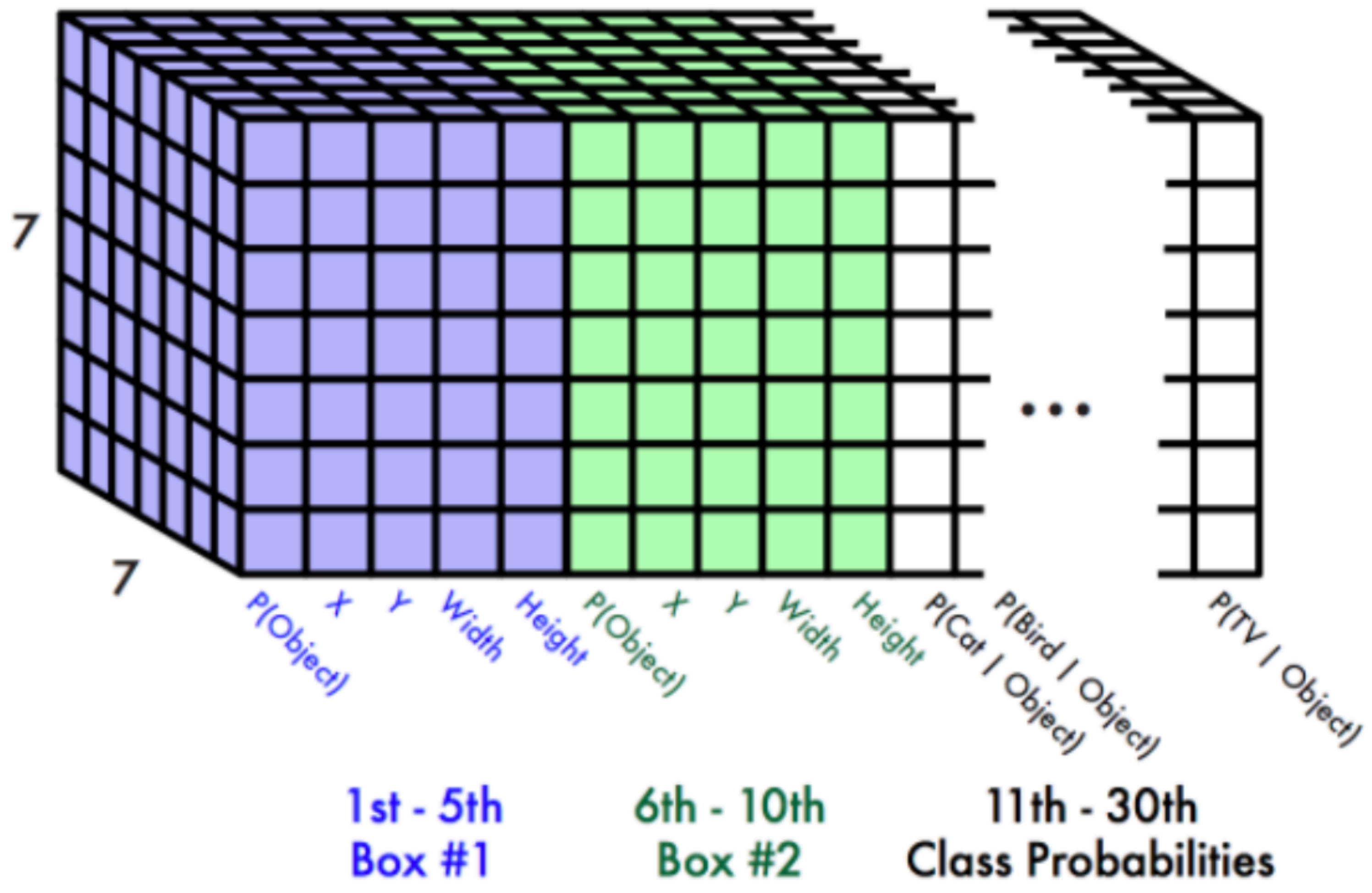
[Redmon et al, CVPR 2016]

Input image
3 x H x W

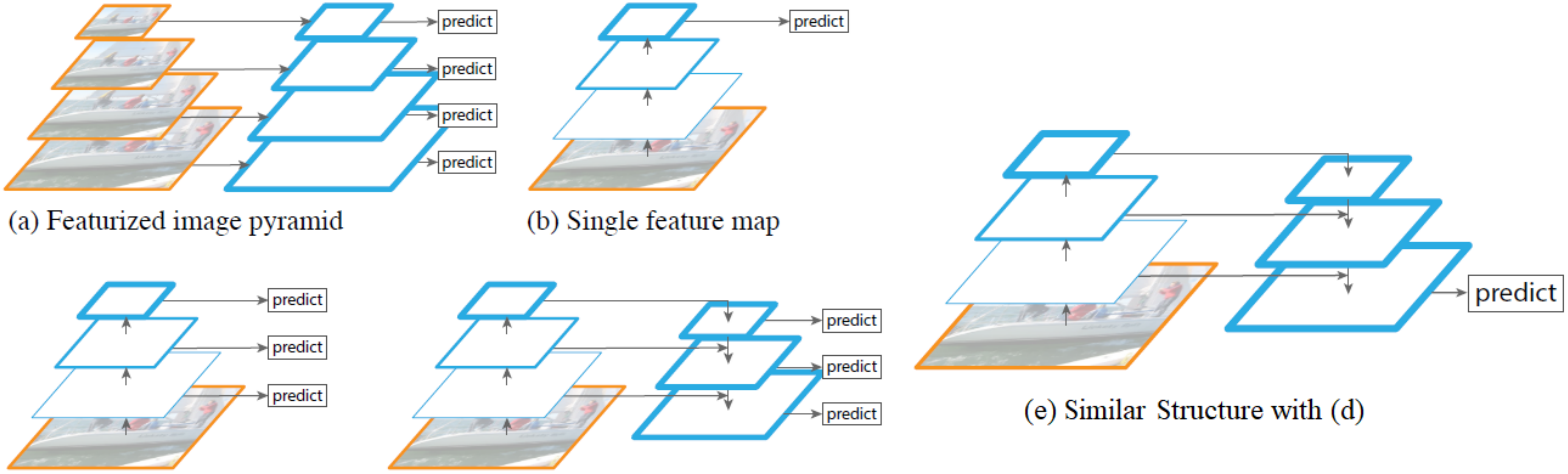


Divide image into grid
7 x 7

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



Feature **Pyramid** Networks



(a) Featurized image pyramid

(b) Single feature map

(c) Pyramidal feature hierarchy

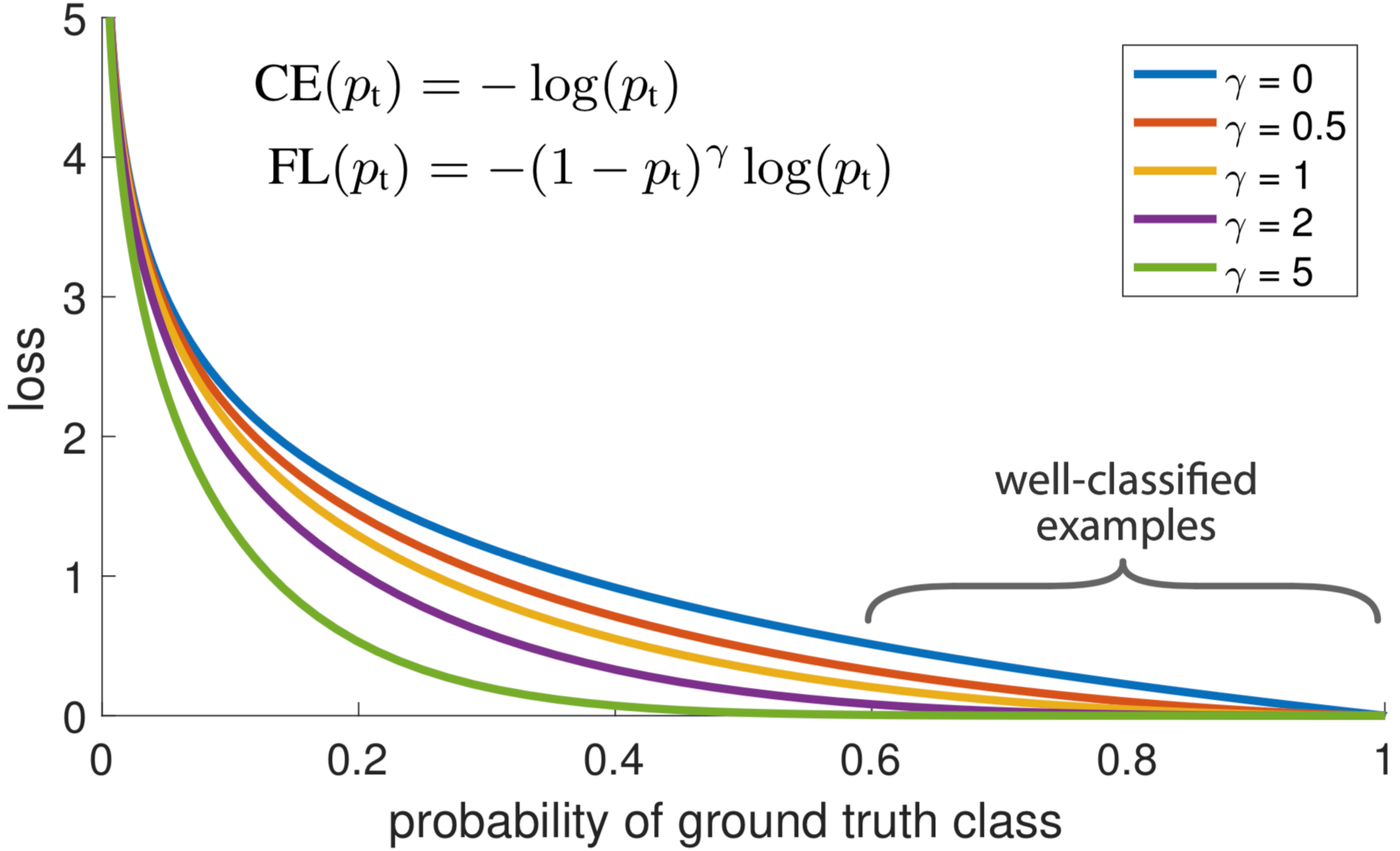
(d) Feature Pyramid Network

(e) Similar Structure with (d)

[Lin et al, CVPR 2017]

Focal Loss

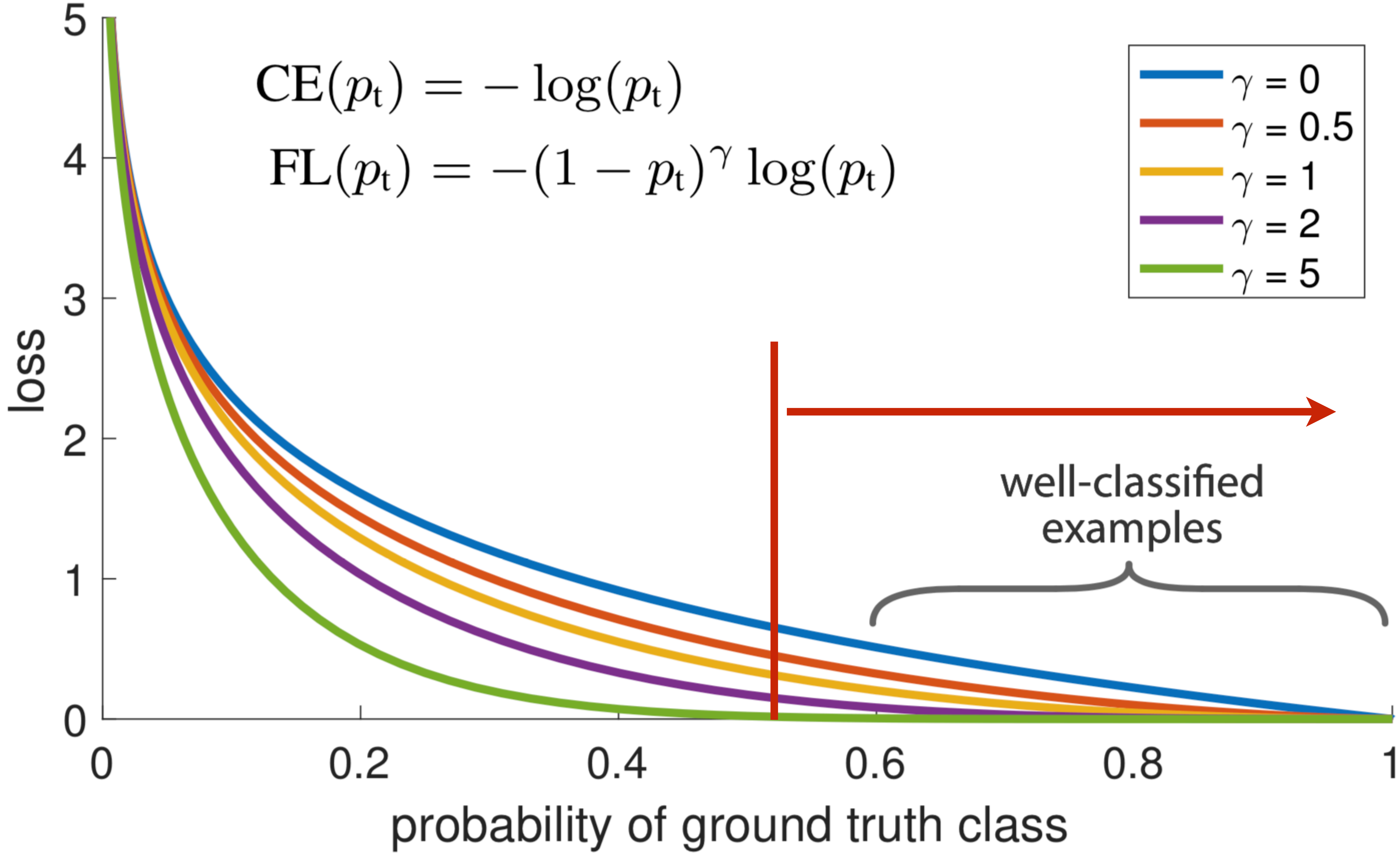
$$p_t = \begin{cases} p & \text{if } \; y = 1 \\ 1 - p & \text{otherwise} \end{cases}$$



[Lin et al, ICCV 2017]

Focal Loss

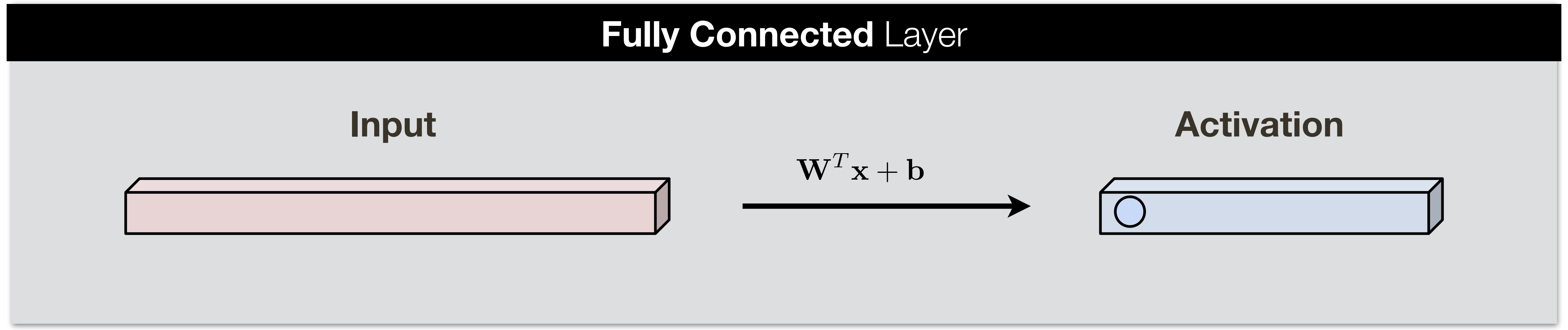
$$p_t = \begin{cases} p & \text{if } \; y = 1 \\ 1 - p & \text{otherwise} \end{cases}$$



[Lin et al, ICCV 2017]

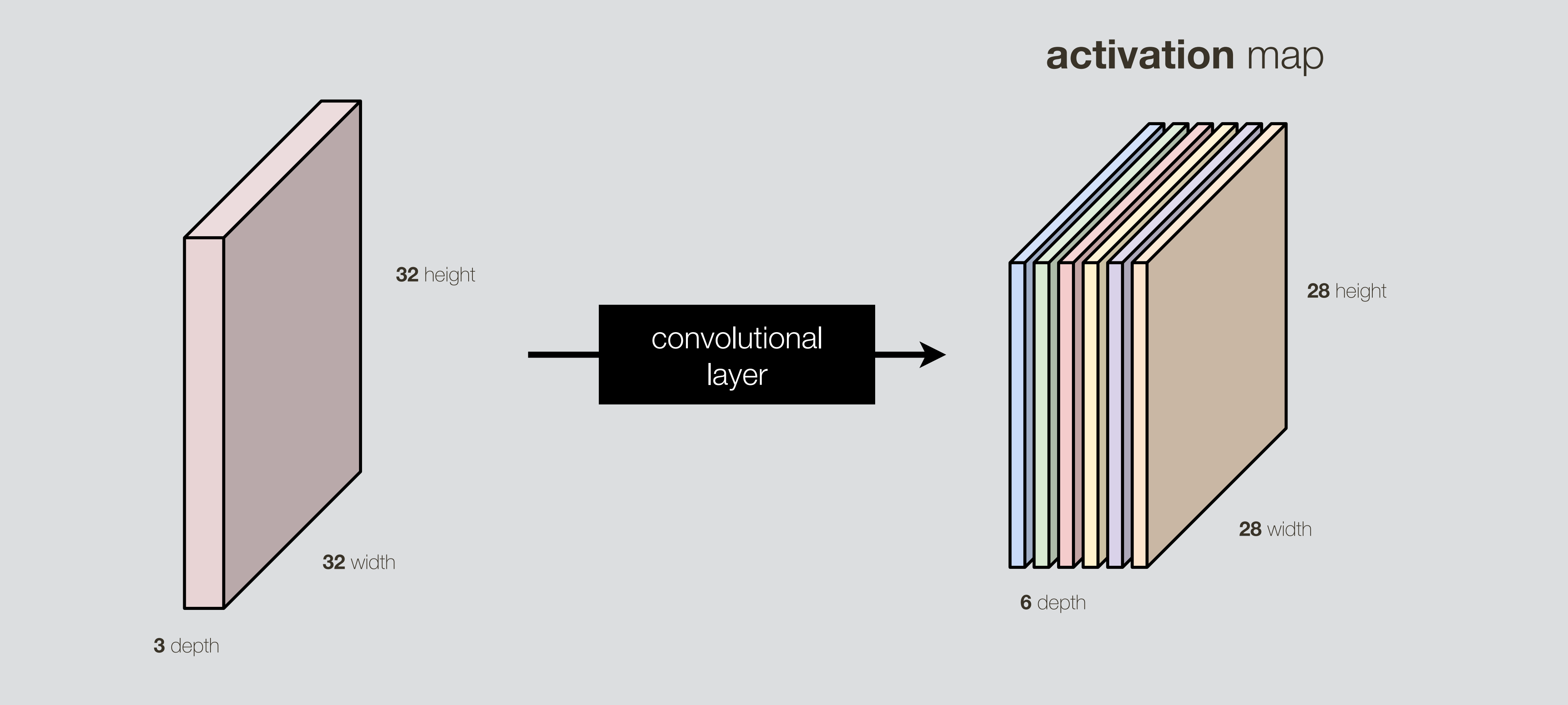
Review of **CNNs**

Review of **CNNs**

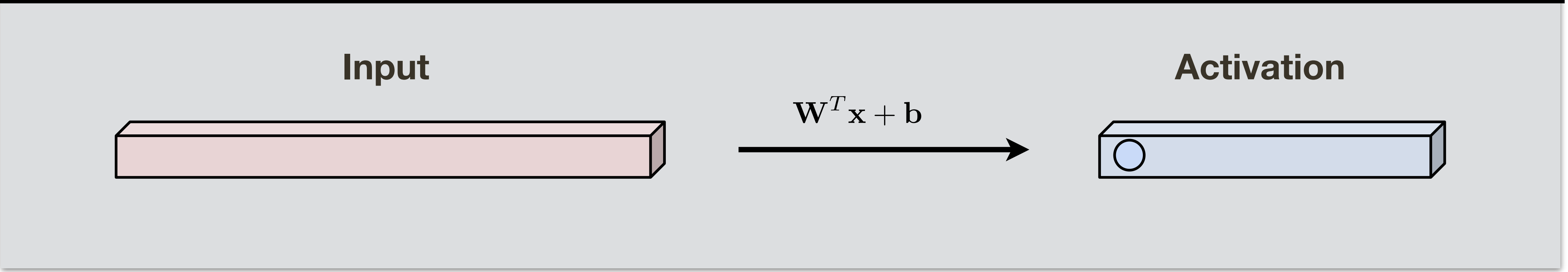


Review of **CNNs**

Convolutional Layer

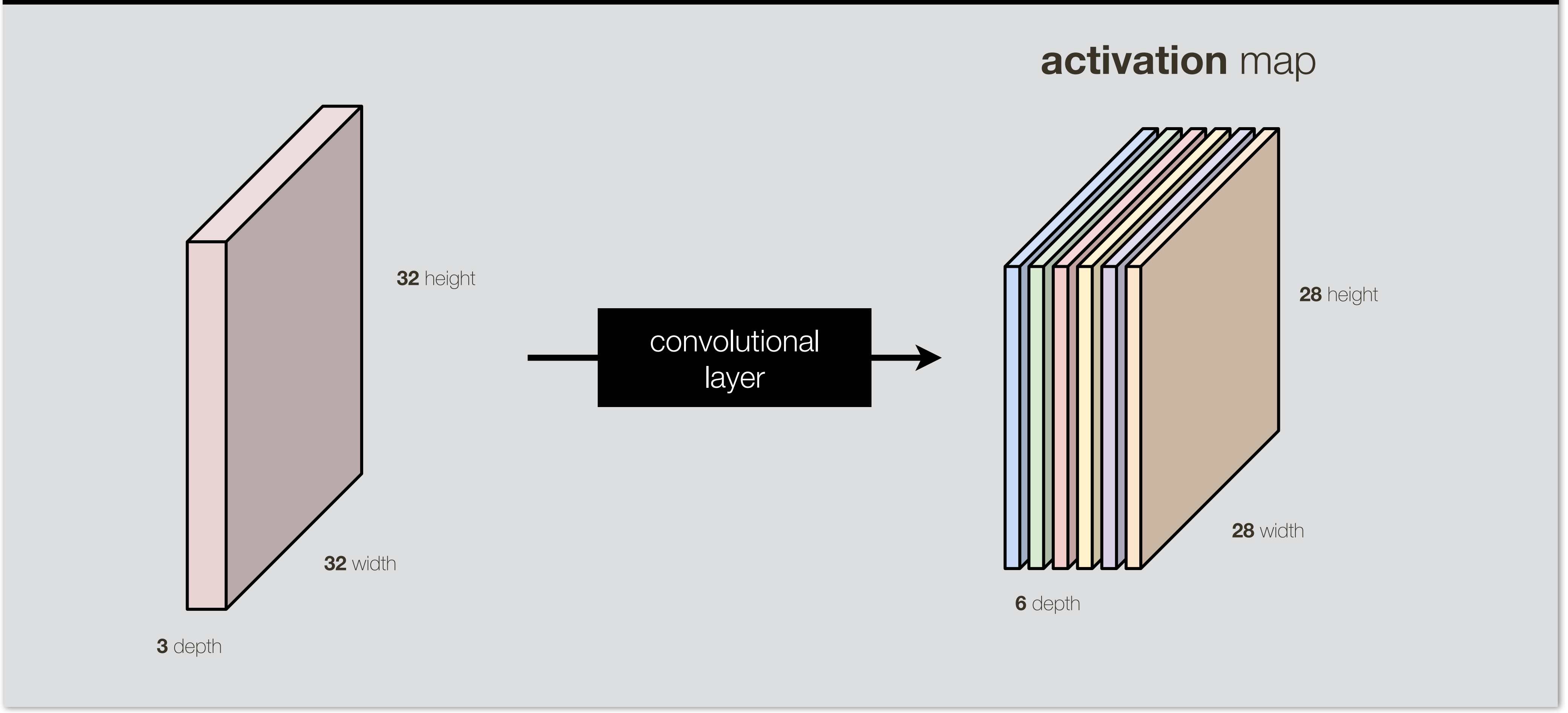


Fully Connected Layer

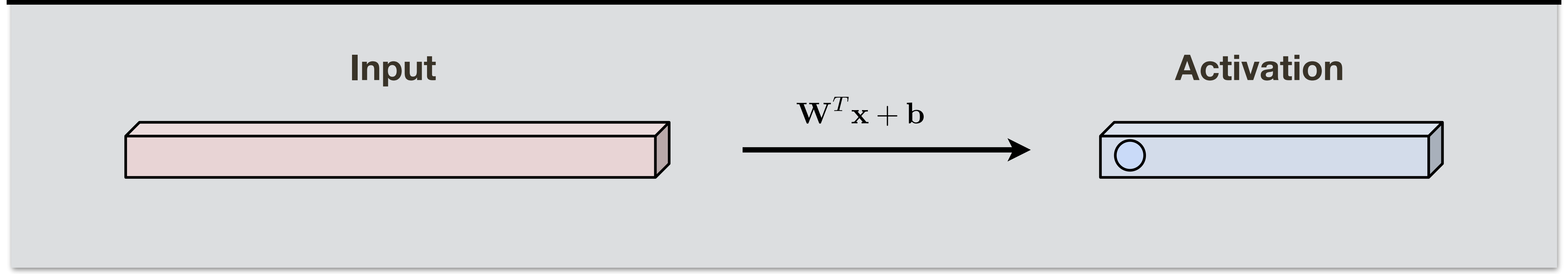


Review of **CNNs**

Convolutional Layer

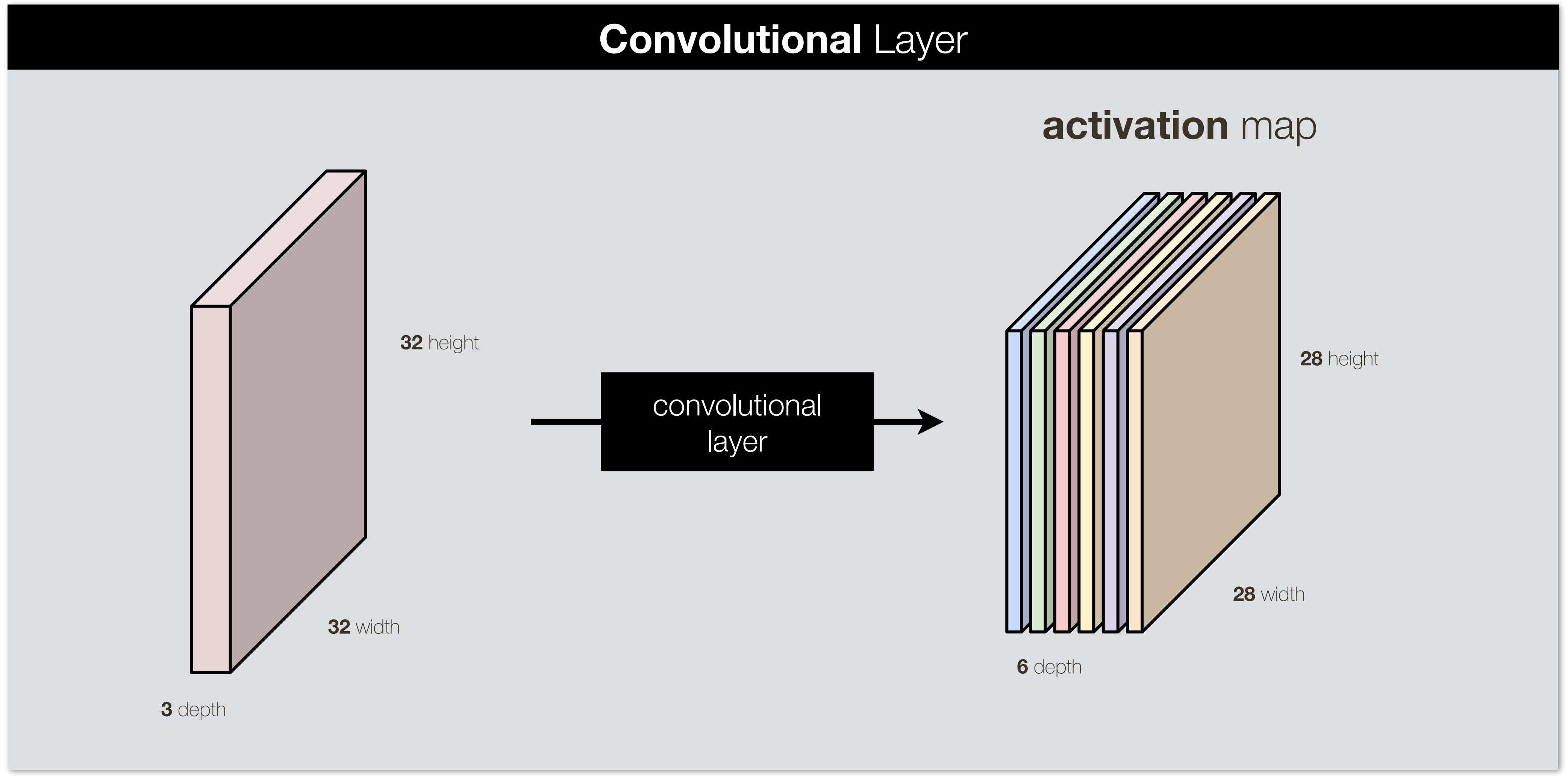


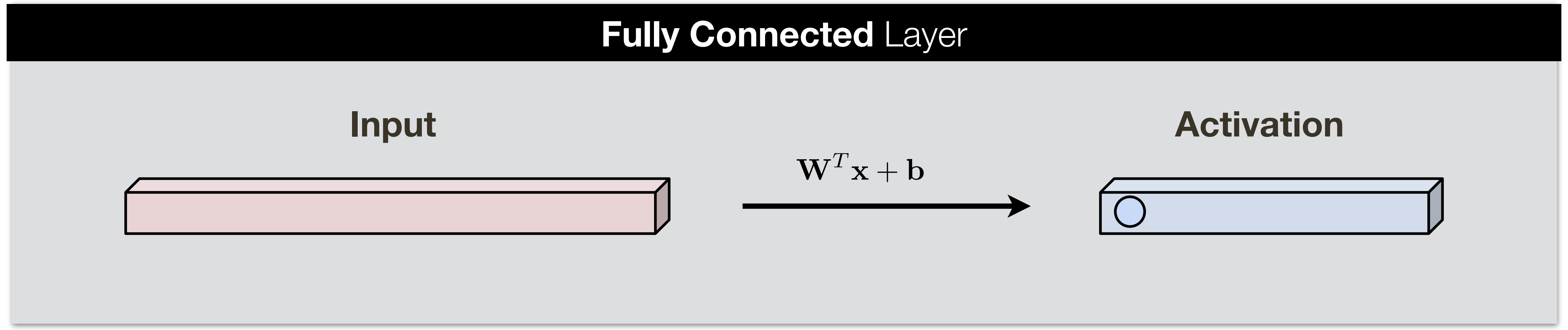
Fully Connected Layer

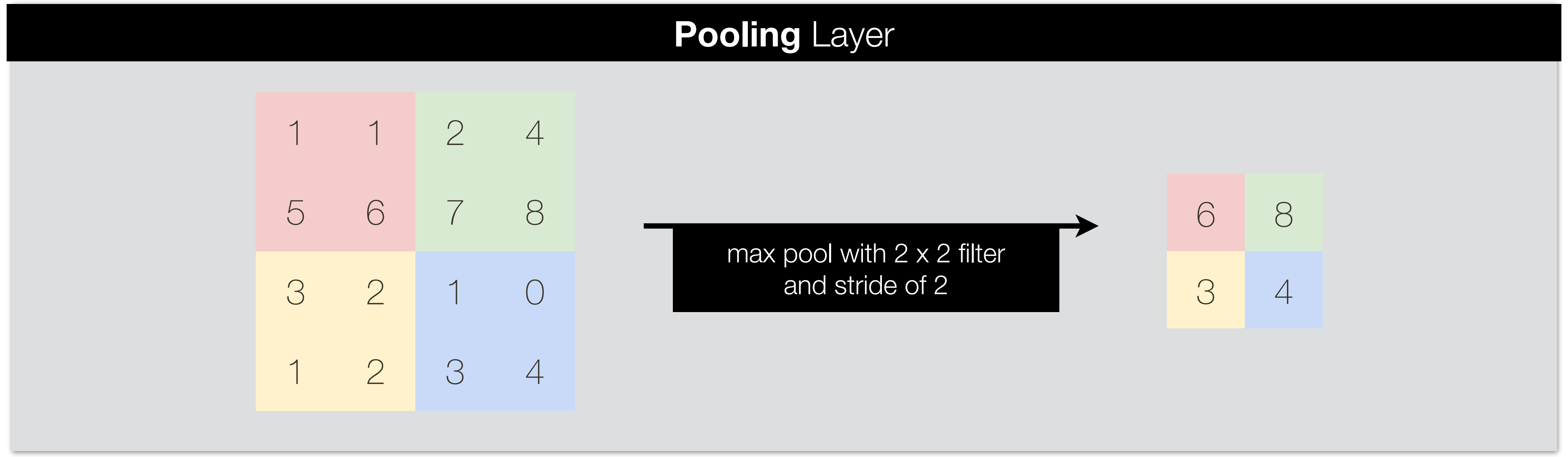


Pooling Layer

Review of **CNNs**





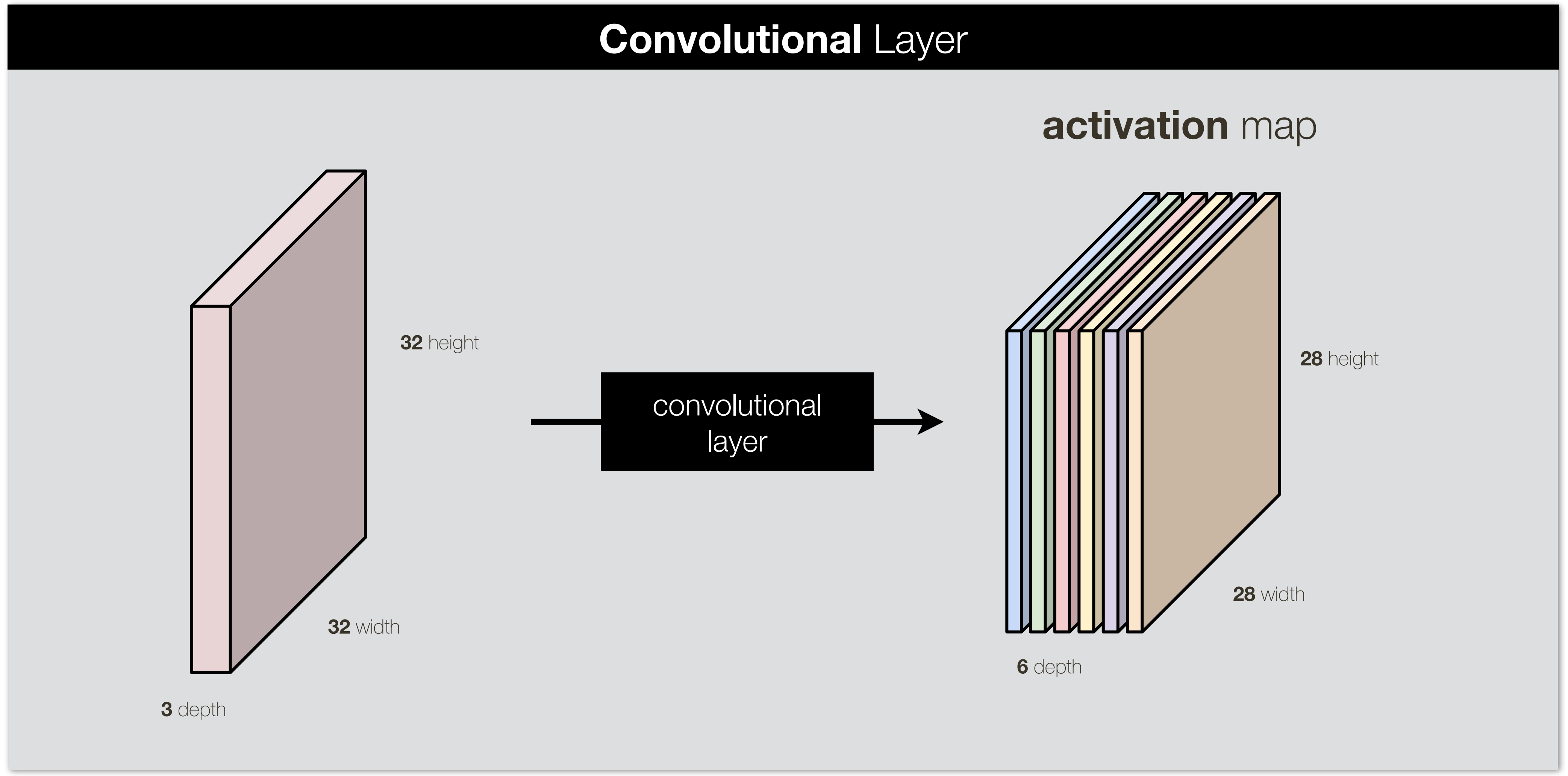


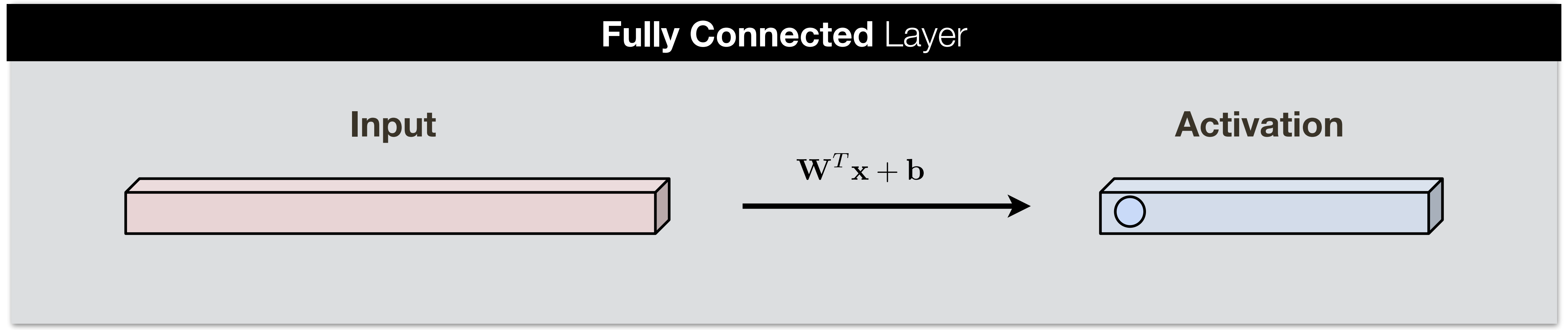
Effective Techniques for **Training**

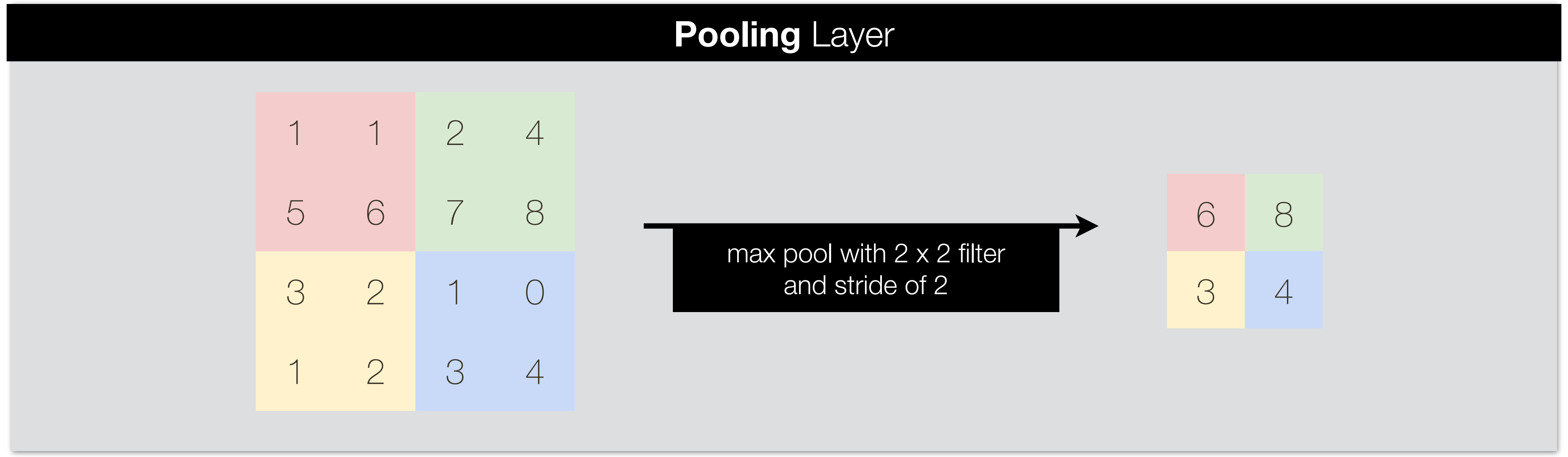
— **Regularization:** L1, L2, data augmentation

— **Transfer Learning:** fine-tuning networks

Review of **CNNs**







Effective Techniques for **Training**

— **Regularization:** L1, L2, data augmentation

— **Transfer Learning:** fine-tuning networks

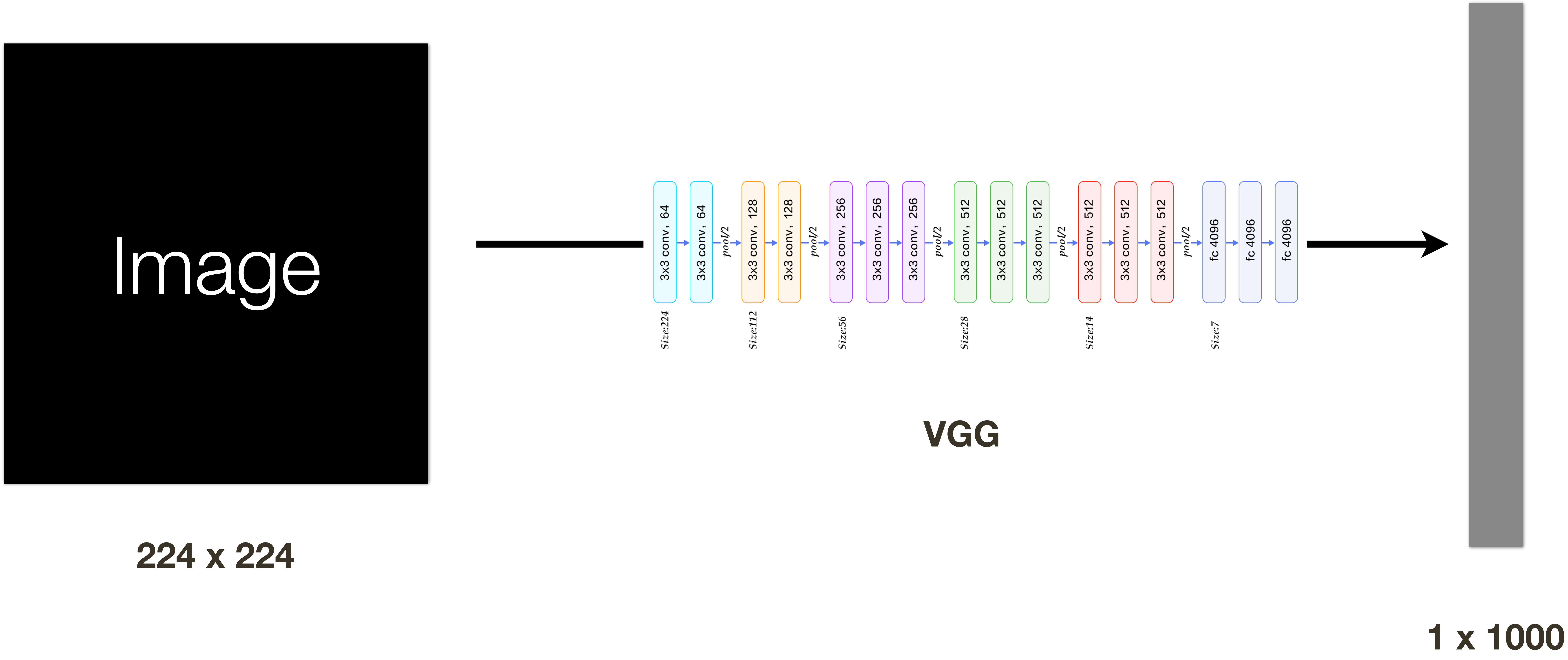
Vision **Applications** of CNNs

— **Classification:** AlexNet, VGG, GoogleLeNet, ResNet

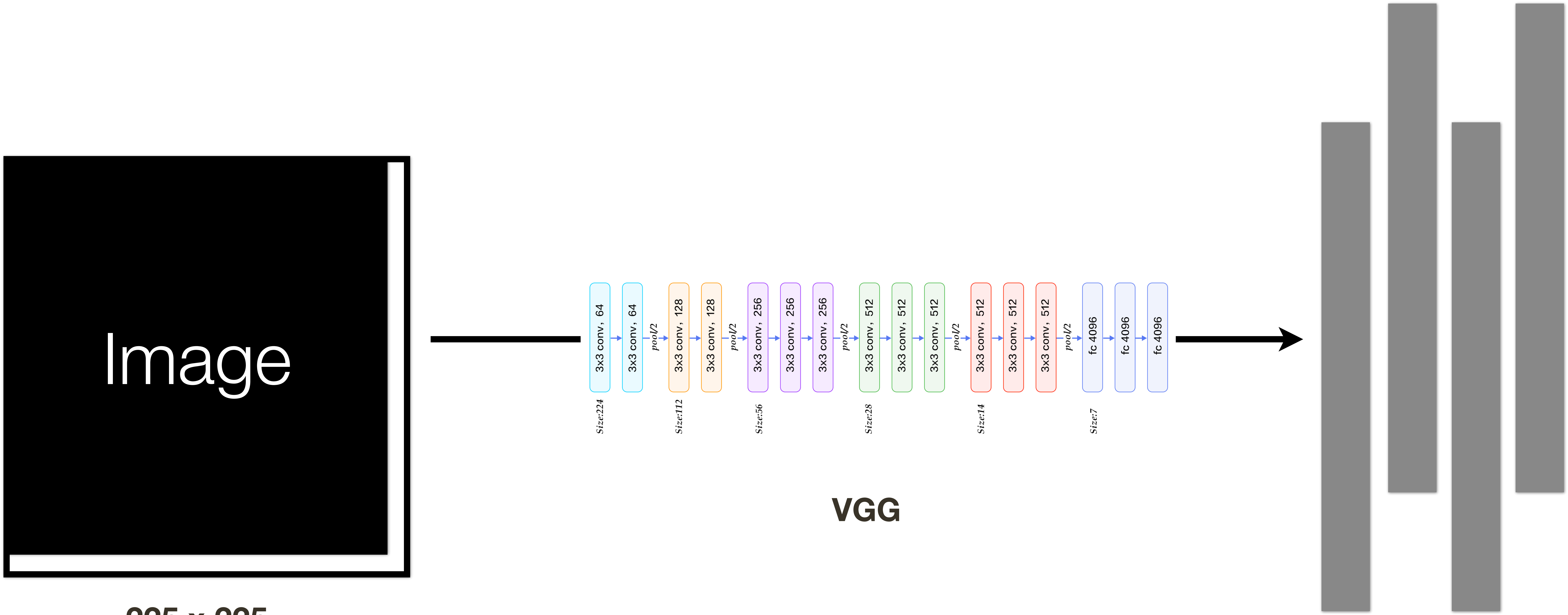
— **Segmentation:** Fully convolutional CNNs

— **Detection:** R-CNN, Fast R-CNN, Faster R-CNN, YOLO

Any **CNN** Could be **Fully Convolutional**



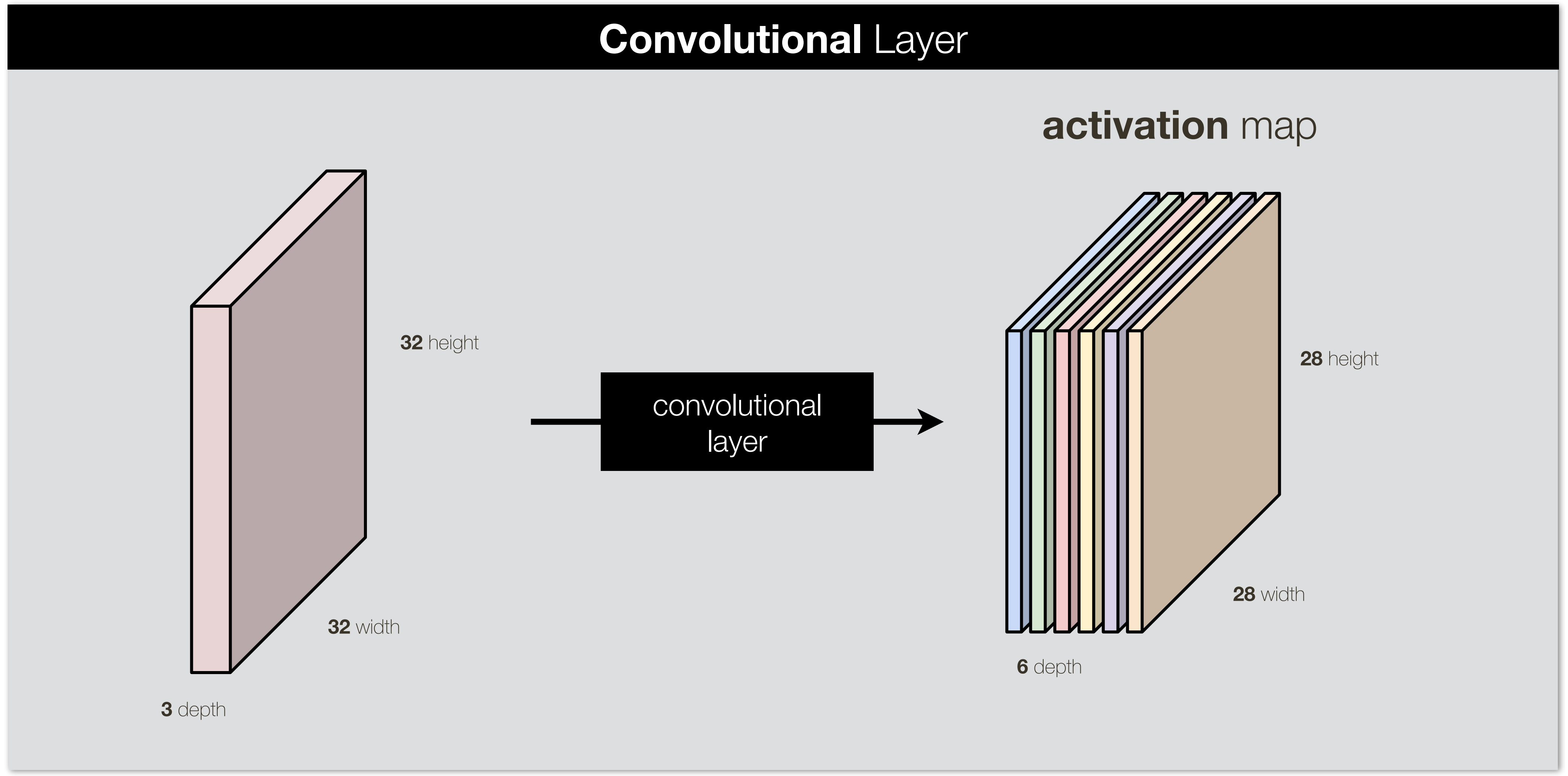
Any **CNN** Could be **Fully Convolutional**

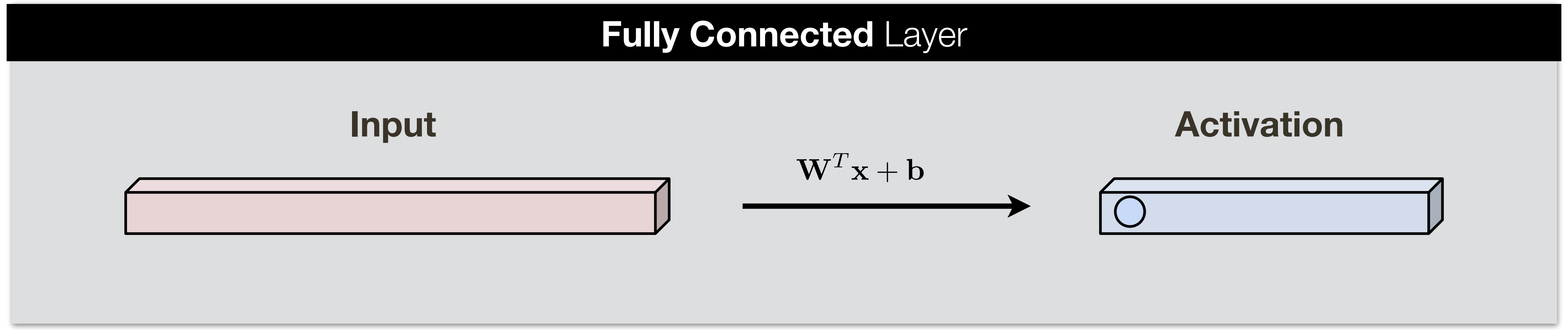


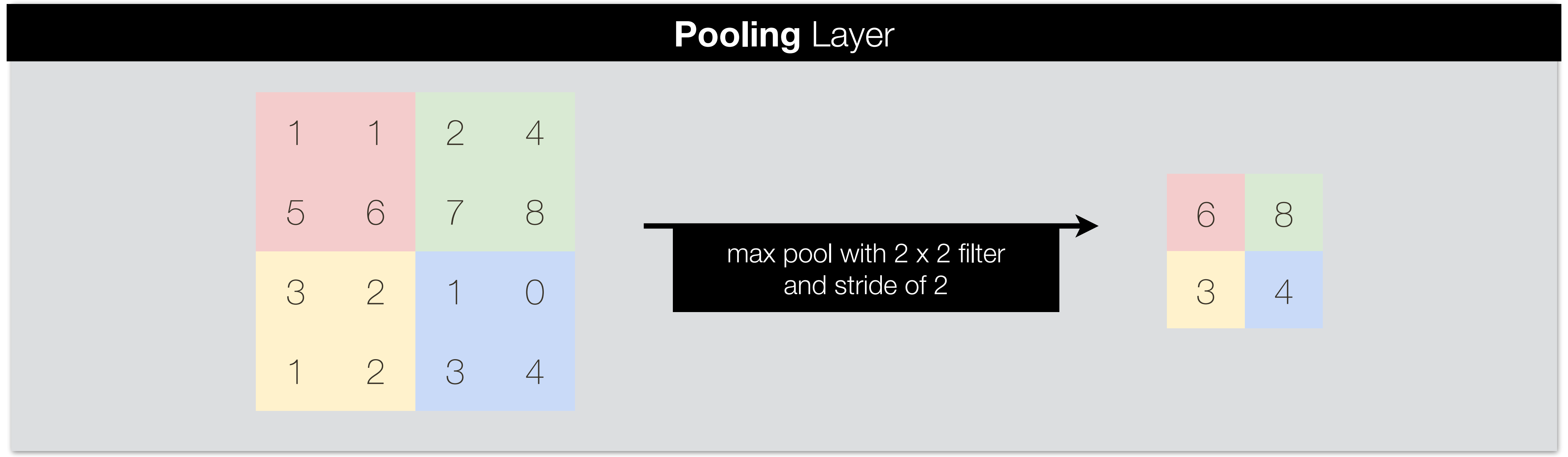
225 x 225

2 x 2 x 1000

Review of **CNNs**







Effective Techniques for **Training**

— **Regularization:** L1, L2, data augmentation

— **Transfer Learning:** fine-tuning networks

Vision **Applications** of CNNs

— **Classification:** AlexNet, VGG, GoogleLeNet, ResNet

— **Segmentation:** Fully convolutional CNNs

— **Detection:** R-CNN, Fast R-CNN, Faster R-CNN, YOLO