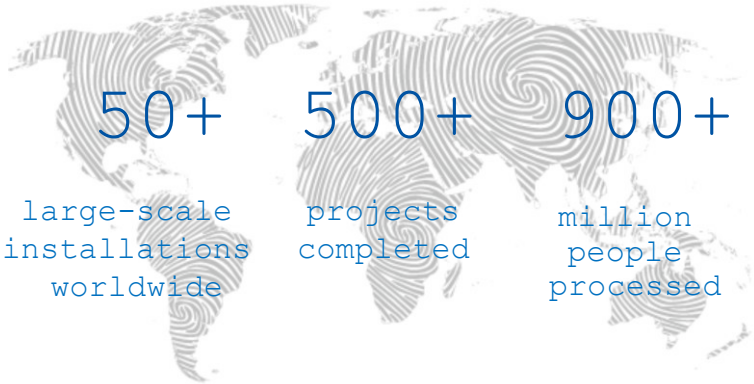
Machine Learning for Face Detection & Recognition

Marián Beszédeš @ ML Meetup

About me

- Student @ FEI @ STUBA
  - Signal processing @ Department of Telecommunications
  - M.Sc. – Neural Network Face Detection
    - Slovak Academy of Sciences award
  - Ph.D. - Face and Facial Features Detection
    - Werner von Siemens Excellence Award
- Computer Vision Team Leader @ Innovatrics
  - IFace SDK – Face biometrics
  - 4 Years
  - 8 team members
  - Brno / Bratislava

About Innovatrics



- 70+ Employees
- 68 countries
- 20 industry awards

- The World's fastest AFIS
  - Automated Fingerprint Identification System
  - 720+ million fingerprint matches in second
  - Awarded accuracy / speed – NIST VPVTE / Minex III / PFT.
- Cutting edge face detection / recognition technology

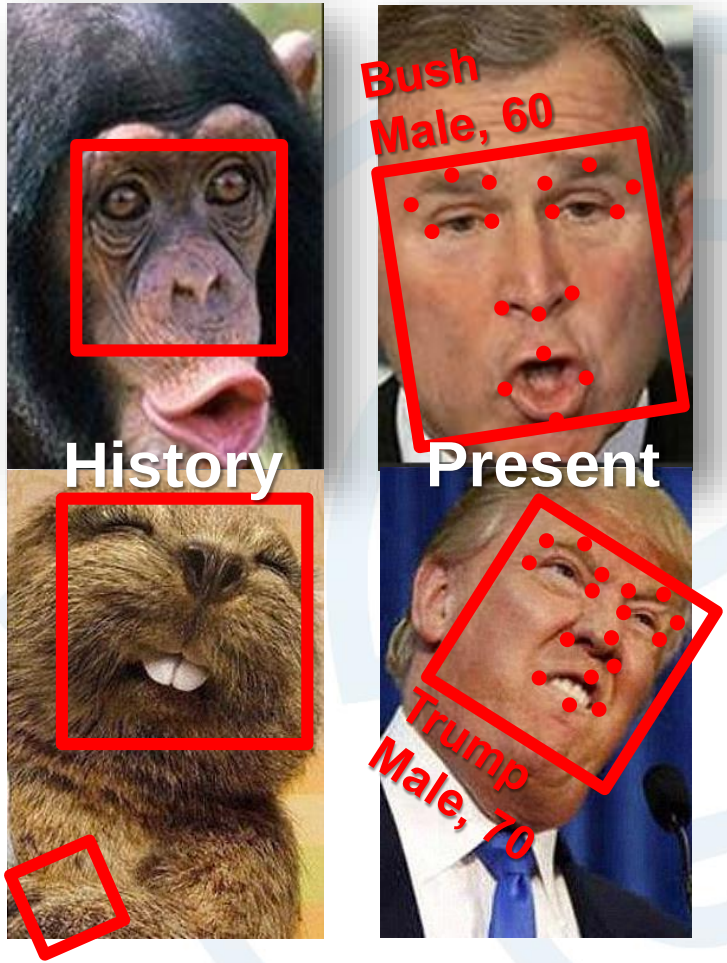
About IFace SDK

- Face detection
- Facial landmarks detection
- Face attributes recognition (mouth status, eyes status …)
- Face tracking (people counting / tracking)
- Face segmentation
- Face recognition
  - Age
  - Gender
  - Verification / Identification

Demo app

Agenda

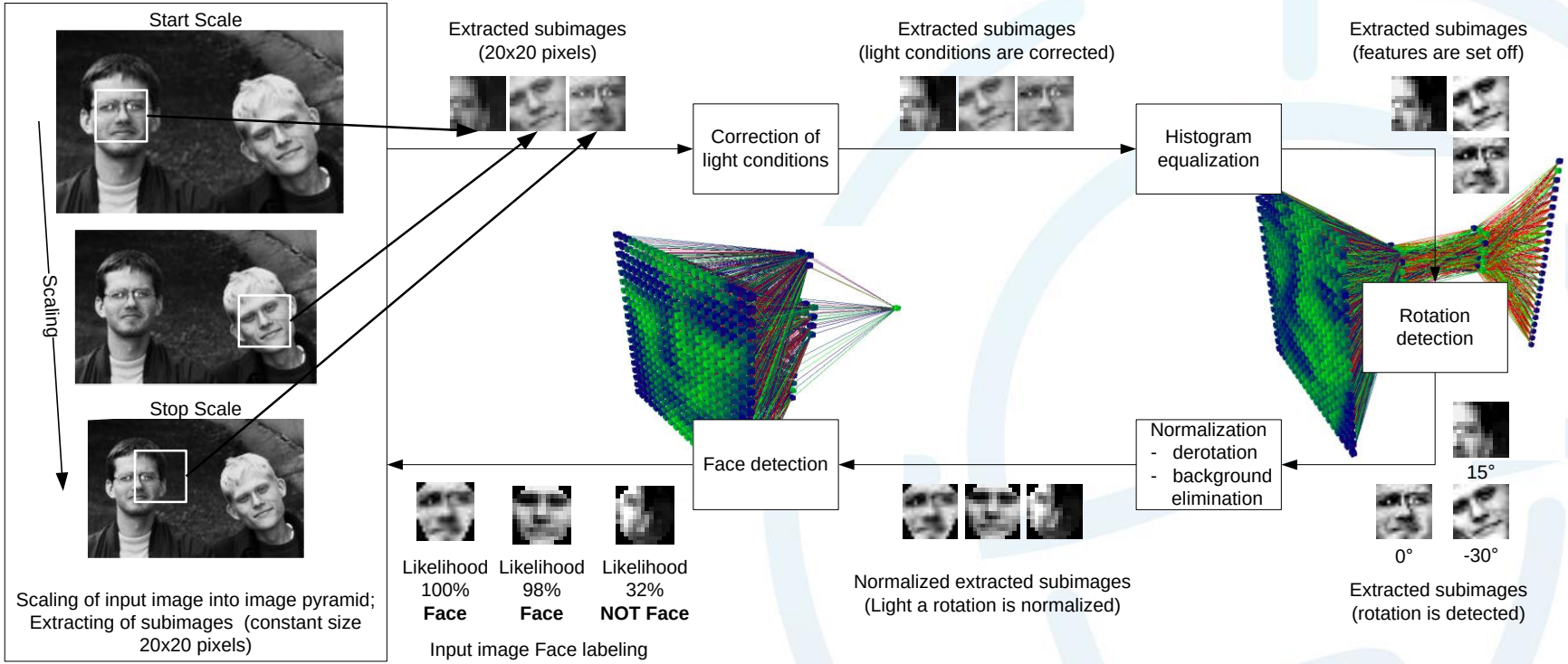
- Face detection history
- Face detection present
- Face recognition



Face detection is not easy

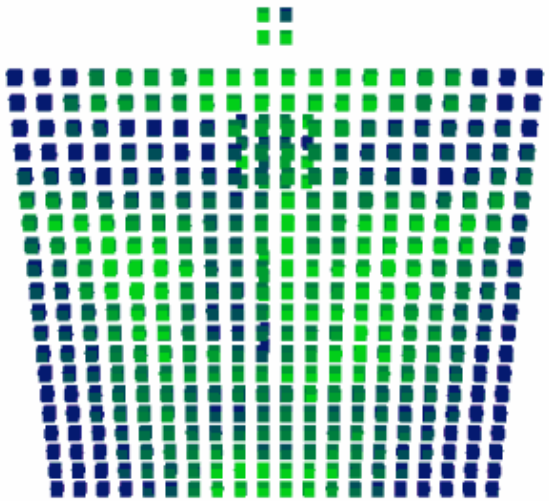
- Faces appearance change:
  - Count
  - Size
  - Orientation – Roll / Yaw / Pitch
  - Occlusions
  - Glasses
  - Face expression
  - Facial hair
  - Face appearance
    - Ethnicity
    - Age / Gender
    - Fake faces

Rowley & Baluja [1998] – NN based face detection



- Rowley, Baluja, Kanade: “Neural network-based face detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, CMU

Rowley & Baluja [1998] - NN face classifier

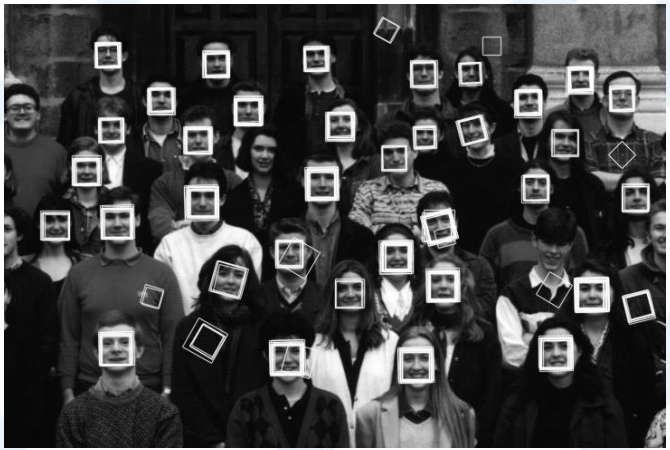


- Input layer:
  - 20×20 = 400 neurons
- Hidden layers:
  - three subnetworks
  - 4+16+6 = 26 neurons
  - analysis of different image parts
- Output layer:
  - Output <0,1>
  - 1 (Face) / 0 (NotFace)
- Training:
  - Positive samples – normalized faces
  - Negative samples – anything else, bootstrapping for hard negatives

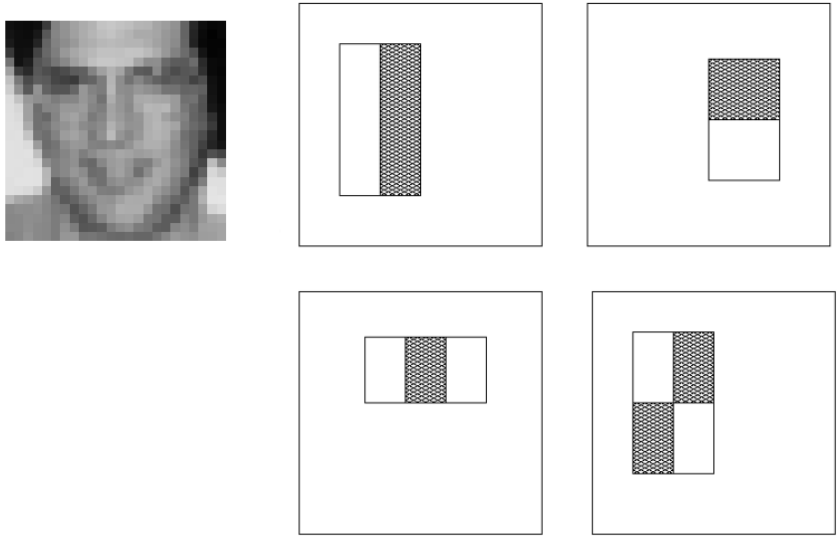
Rowley & Baluja [1998] - Resume

- Pros
  - It works !!!
  - Good False Reject rate

- Cons
  - Too many False Accepts
  - Slooooooooooow = mins / image
  - Hard negative mining & iterative training is necessary



Viola & Jones [2001] – Boosted Cascade Detector

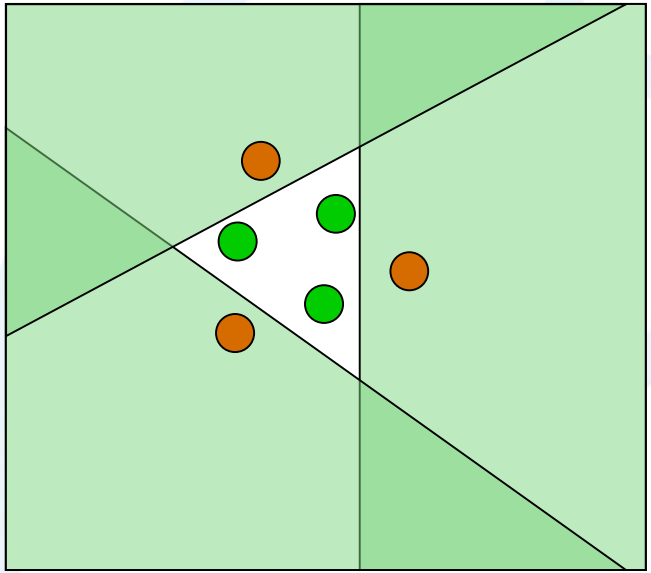


- Haar features:
  - Very, very fast (using Interal Image)
  - Very simple
  - 60k possible features for 24x24 image
- Weak classifier:
  - Lets find a the feature which differentiate between faces & non-faces well
- Strong classifier:
  - Lets combine many weak classifiers effedtively
  - Done by **AdaBoost**
  - 200 weak learners = Good classifier

- Viola, Jones: "Rapid object detection using a boosted cascade of simple features“, *CVPR,* 2001

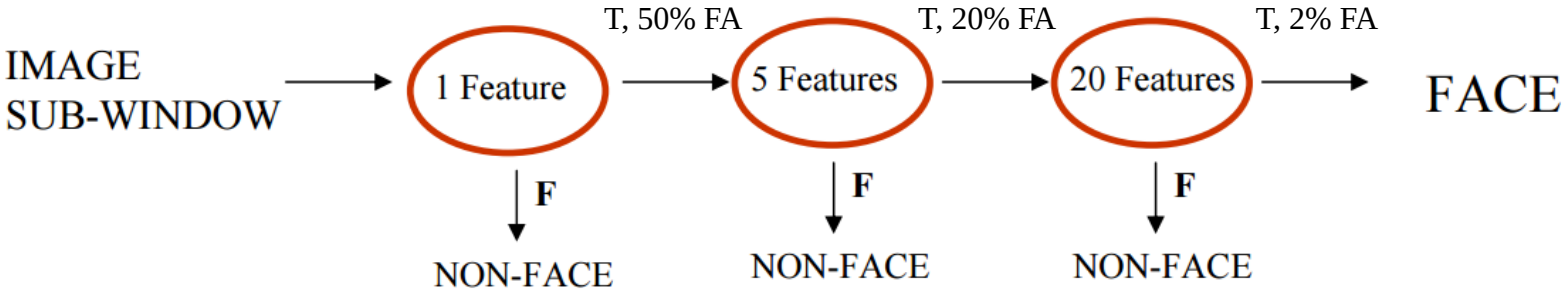
Freund & Schapire [1995] - ADABoost

1. Find weak classifier
2. Add it strong classifier
3. Re-weight samples



- Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." icml. Vol. 96. 1996.

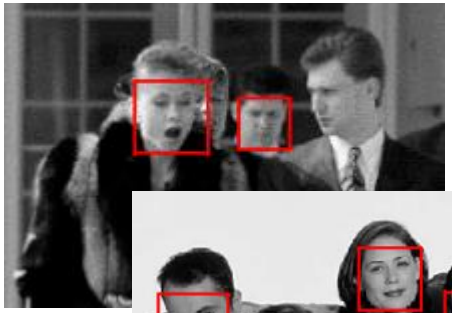
Viola & Jones [2001] - Cascade

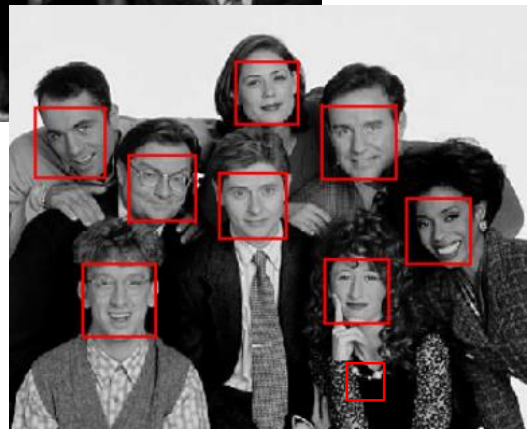


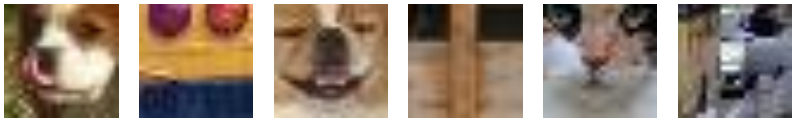
- 1 Cascade level = 1 Strong Classifiers
- Cascade position defines strong classifier properties
  - Weak classifiers count
  - Acceptable False Accept (FA) rate
  - False Reject Rate should be as low as possible
- Image areas with no faces are evaluated very fast

Viola & Jones [2001] - Example

Viola & Jones [2001] - Resume







- Pros:
  - It works really fast
    - Pc 30-60 Fps / Embeded 10-20 Fps
  - Use variations of AdaBoost
  - Use variations of features (LBP, …)
  - Many commercial applications
- Cons:
  - Do not work for not-frontal faces
    - Each face rotation = dedicated detectors
    - N rotations = N detectors = N times slower
  - Still many false accepts
  - The more cascade levels, the more FRs

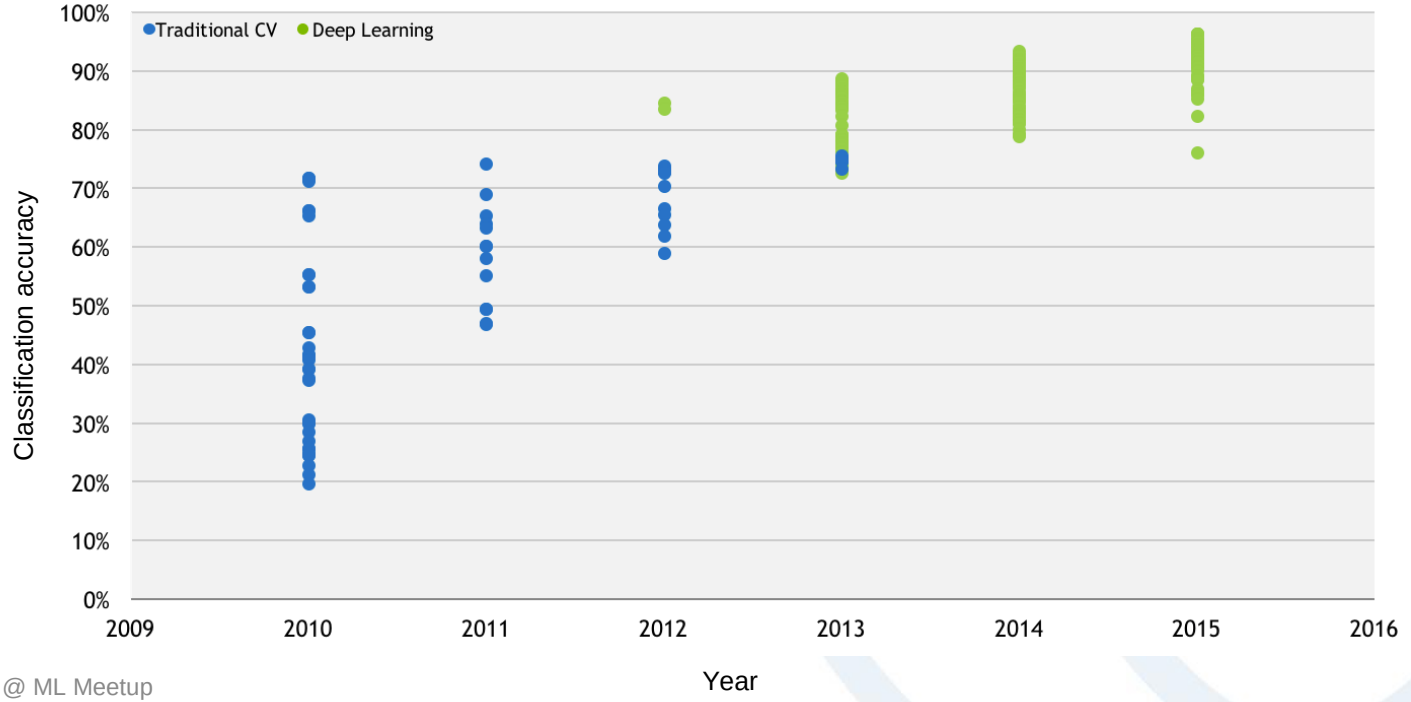
Imagenet [2010]

- Task
  - Assign the correct class label to the whole image

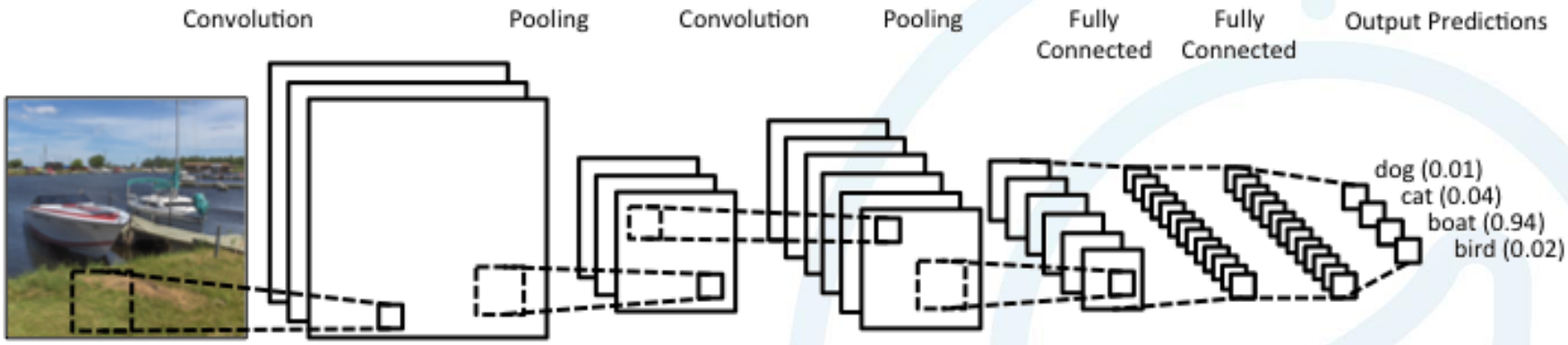
- Dataset
  - 1 image = 1 object
  - Classes: 1000
  - Training set: 1.2M
  - Testing set: 0.15M
  - Res: 224 x 224

Deep Conv. NN (DCNN),2012-present

- ILSVRC Top-5 Classification Accuracy on ImageNet
  - Development form 2010 to 2015



Deep Convolutional NN, Imagenet Classifier

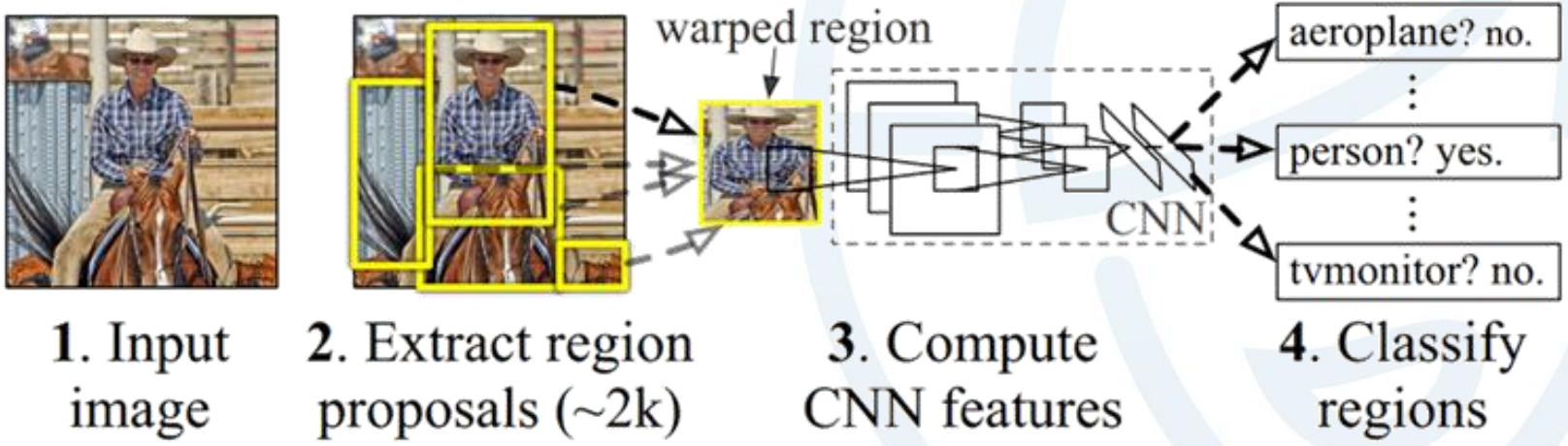


- Constant image size
- Layers:
  - Convolution + Relu
  - Pooling
  - Fully connected
  - Decision

- Supervised learning
  - Stochastic Gradient Descent
  - Error backpropagation
  - Loss function
    - How good we are in classification / detection / regression?
    - Selection I crutial

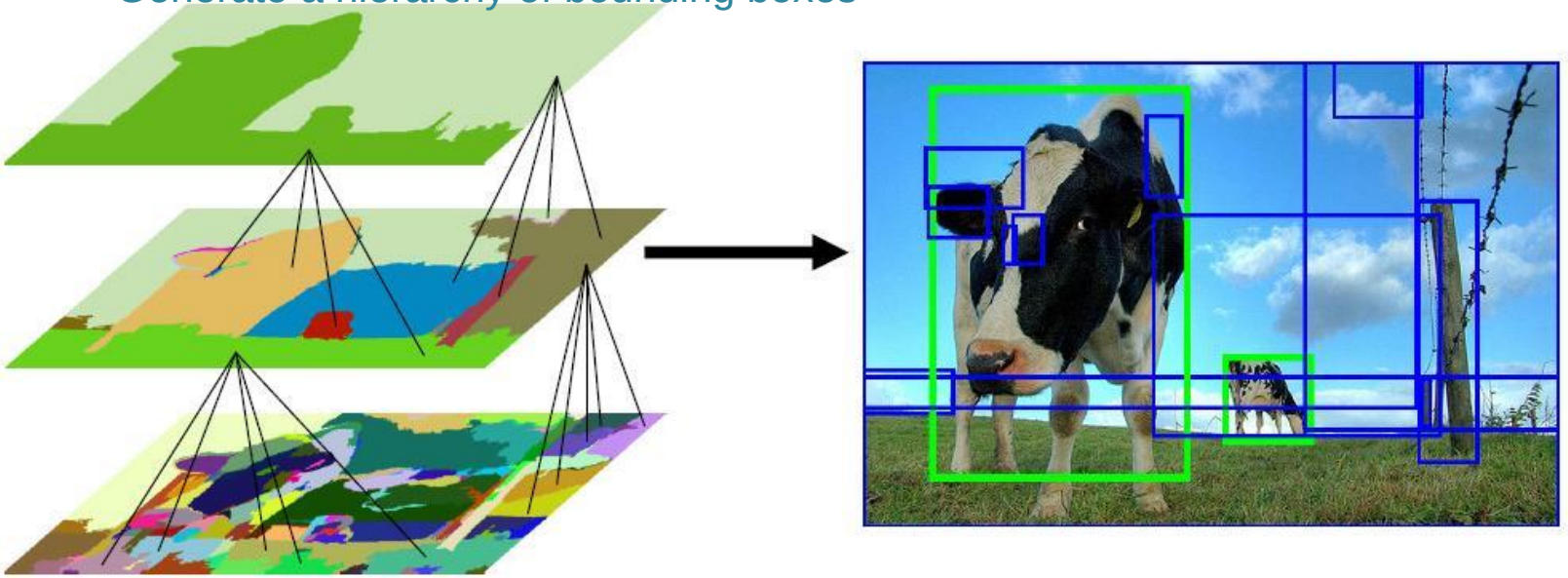
R-CNN[2014], DCNN Object detection

- Girshick, Ross: "Rich feature hierarchies for accurate object detection and semantic segmentation“, *CVPR,* 2014, Berkeley
  - Region proposals: Selective search
  - CNN features: from pre-trained “ImageNet classification NN”
  - Candidates for classification & position adjustment



R-CNN [2014], Selective Search

- Uijlings & Sande. "Selective search for object recognition." International journal of CV
  - Do over-segmenttion
  - Group regions with highest similarity
  - Generate a hierarchy of bounding boxes



R-CNN [2014], Classification & Position adjustment

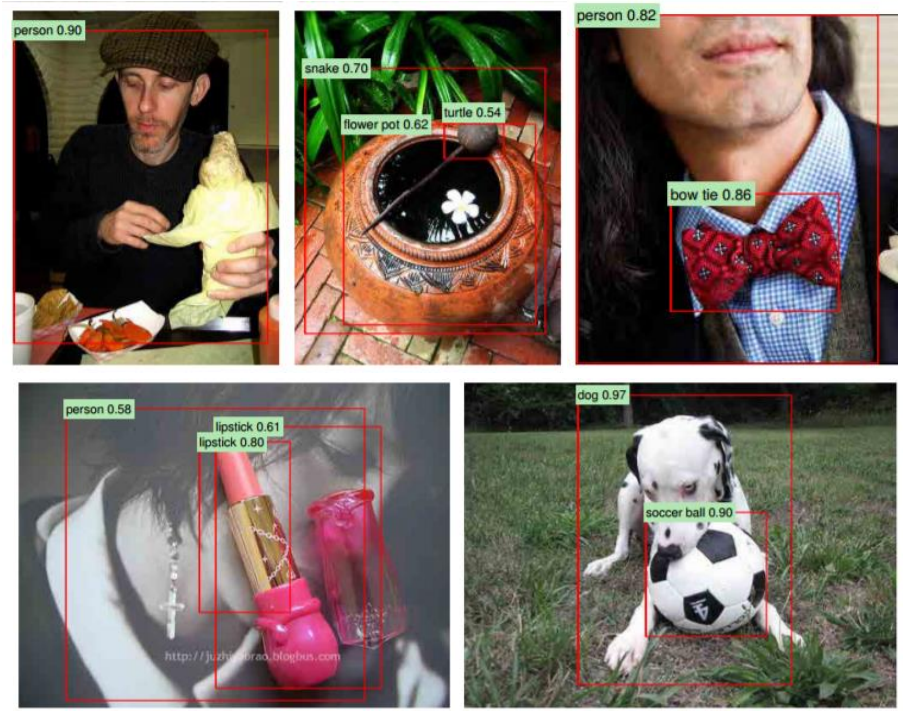
- Proposal CNN features
  - Classification into 1000 + 1 (background) classes
    - Multinomial logistic loss
  - Regression of position adjustment = (x,y,width,height)
    - Euclidean Loss

(0, 0, 0, 0)
Proposal is good

(.25, 0, 0, 0)
Proposal too far to left

(0, 0, -0.125, 0)
Proposal too wide

R-CNN [2014], Resume



- Pros:
  - It works for detection of 1000 object classes
- Cons = SLOW
  - VGG16 + NvidiaK40 GPU = 10 – 50 s / Image
  - Selective search is slow
  - 2k RoIs is too much:
    - CNN feature calculation & evaluation for every RoI

Face Detection inspired by R-CNN

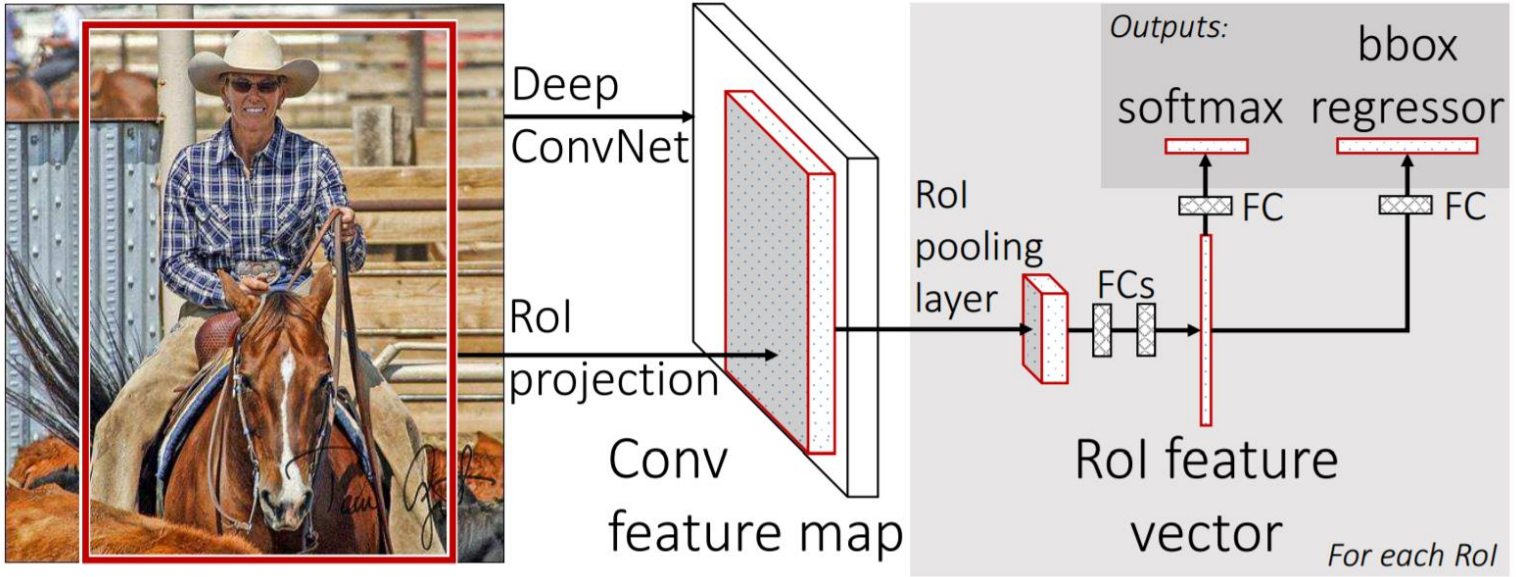
Boosted cascade region proposal

CNN – region validation normalization

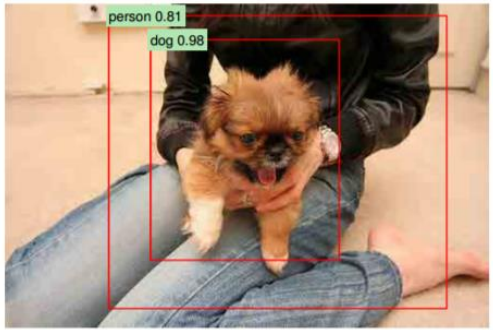
- Boosted Cascade region proposal
  - Few false rejects
  - Possible false accepts
  - Rough position and size
  - Much faster then Selective search
- Validation
  - Face / Not-face ? => Confidence score
- Normalization
  - Normalization Size / Position / Orientation

Fast R-CNN [2015]

- Girshick, Ross. "Fast R-CNN." *CVPR* 2015, Microsoft
  - Still Selective Search for RoI
  - DCNN with no full connect produce feature maps for input images of any size
  - RoI pooling = resize of feature maps

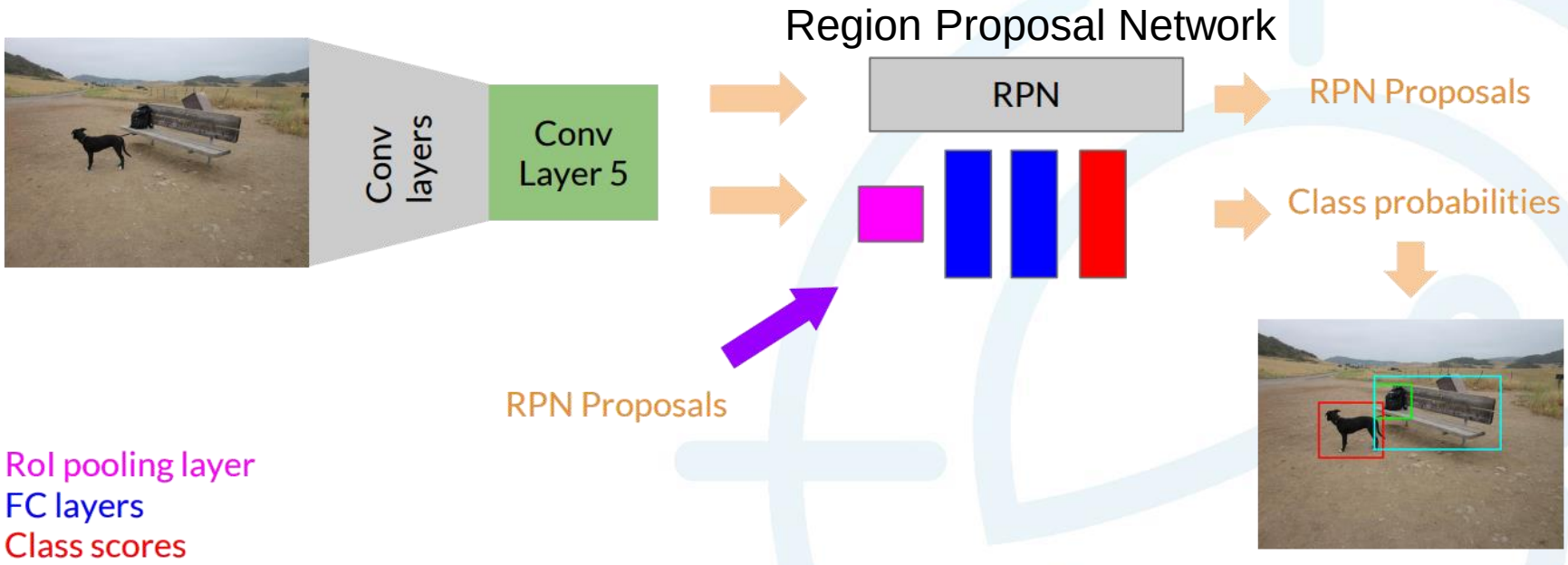


Fast R-CNN [2015], Resume



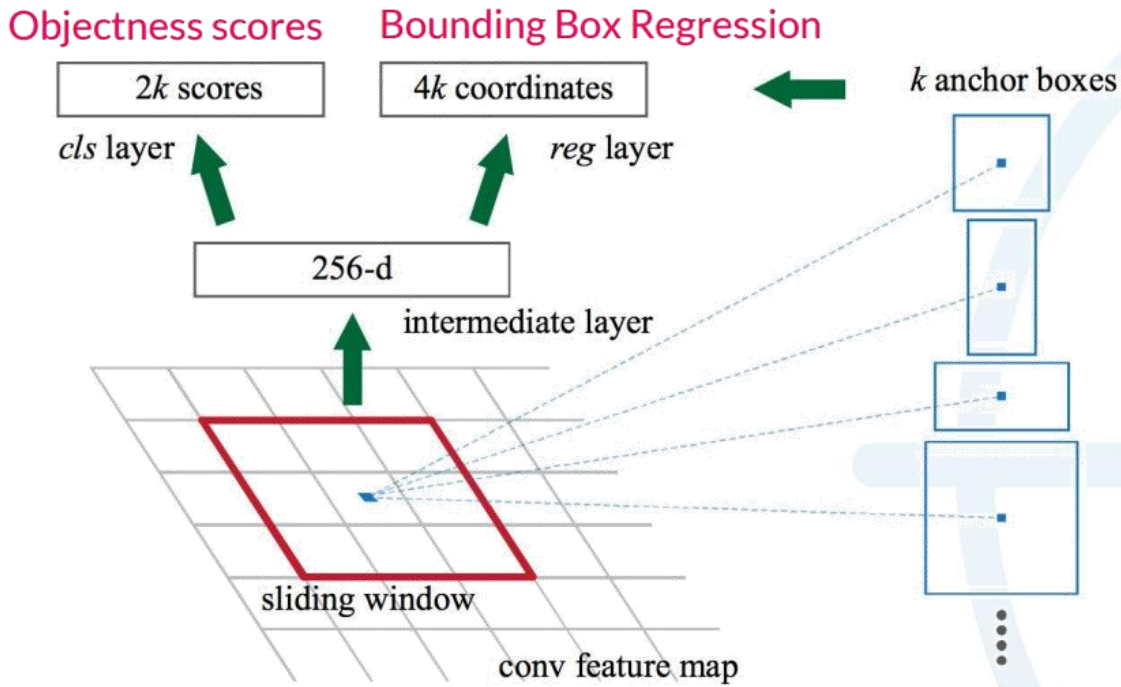
- Pros
  - 80x faster than R-CNN
  - VGG16 + NvidiaK40 GPU = 1 - 5 Fps
  - Similar accuracy
- Cons
  - Still too slow for CPU real-time
  - Doesn't detect small objects

Faster R-CNN [2015]



- Ren, He, Girshick: “Faster r-cnn: Towards real-time object detection with region proposal networks”, *Advances in neural information processing systems,* 2015, Microsoft

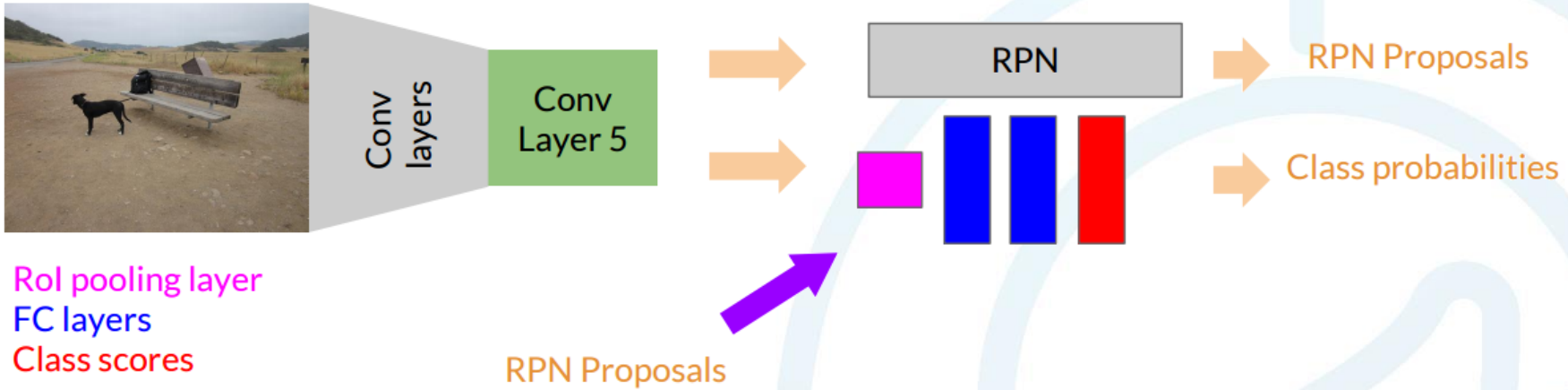
Faster R-CNN, Region proposal network



For every position in feature map:

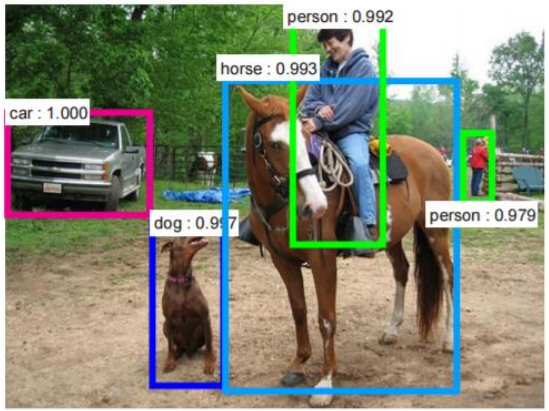
- Analyze region 3x3
  - Apply 256 convolutional (3x3) filters
- Evaluate potential objects relatively to **k** anchor boxes:
  - Objectness = presence or absence
    - cls layer - 2 classes, softmax loss
  - Improve localization
    - reg layer - 4 shape offsets, reg loss
- In practice, **k** = 9
  - 3 different scales
  - 3 aspect ratios

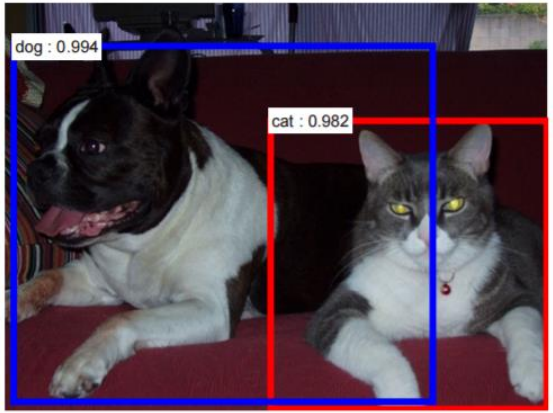
Faster R-CNN, 4-step training

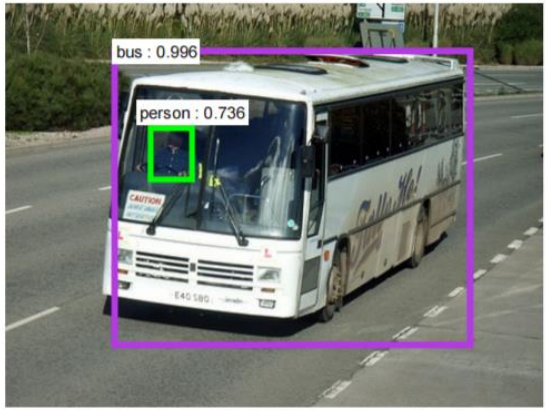


- Step 1: Train RPN initialized with an ImageNet pre-trained model
- Step 2: Train Fast R-CNN with learned RPN proposals
- Step 3: The model trained in 2 is used to initialize RPN and train again
- Step 4: Fine tune FC layers of Fast R-CNN using same shared convolutional layers as in 3.

Faster R-CNN, Resume







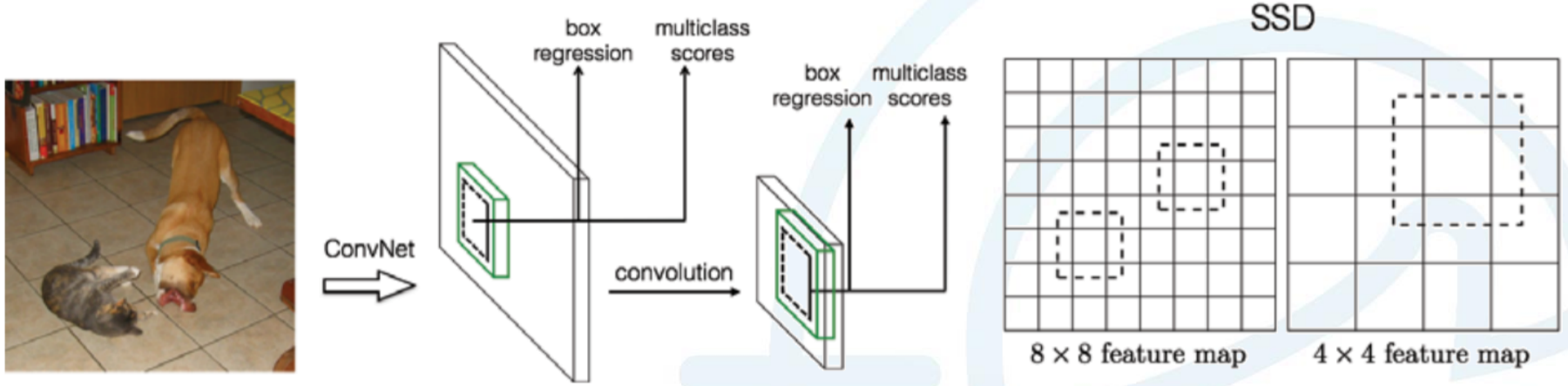
Pros:

- Higher detection accuracy than Fast R-CNN
- Really faster
  - RPN & Object Detection Network share convolutional features
- Speed: 5-17 fps

Cons

- Smaller objects are not detected
- Complicated training in 4 steps

Single Shot Detector (SSD) [2016]



- Liu, Anguelov, Erhan, Szegedy: "SSD: Single shot multibox detector. ECCV October 2016, Google

SSD Framework

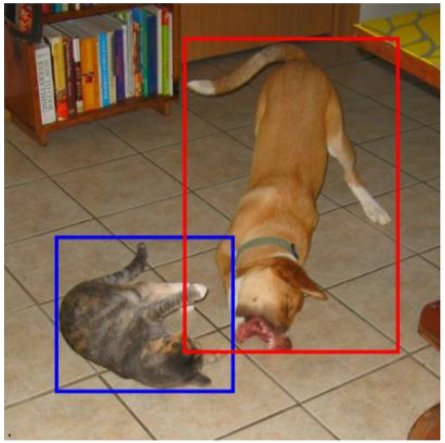
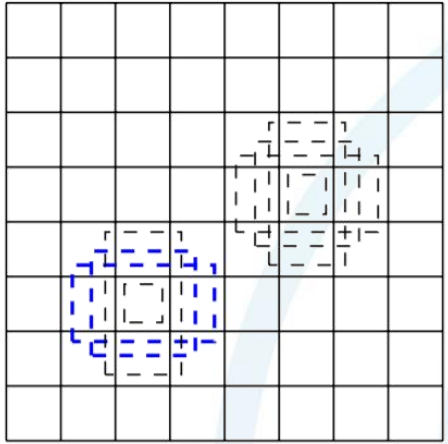
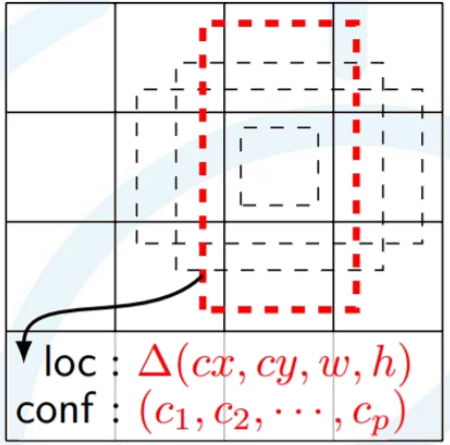


Image with GT boxes



8×8 feature map

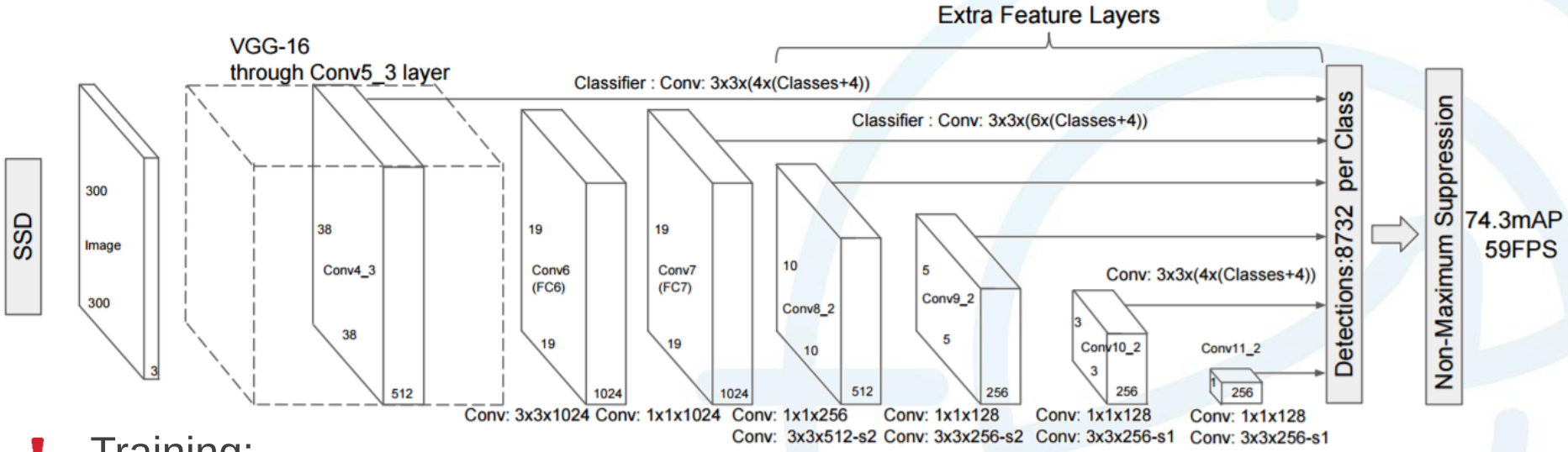


4×4 feature map

- Several feature maps with different scales
- Each location of feature map is evaluated
  - 3x3 neighborhood
  - different aspect ratios anchor boxes

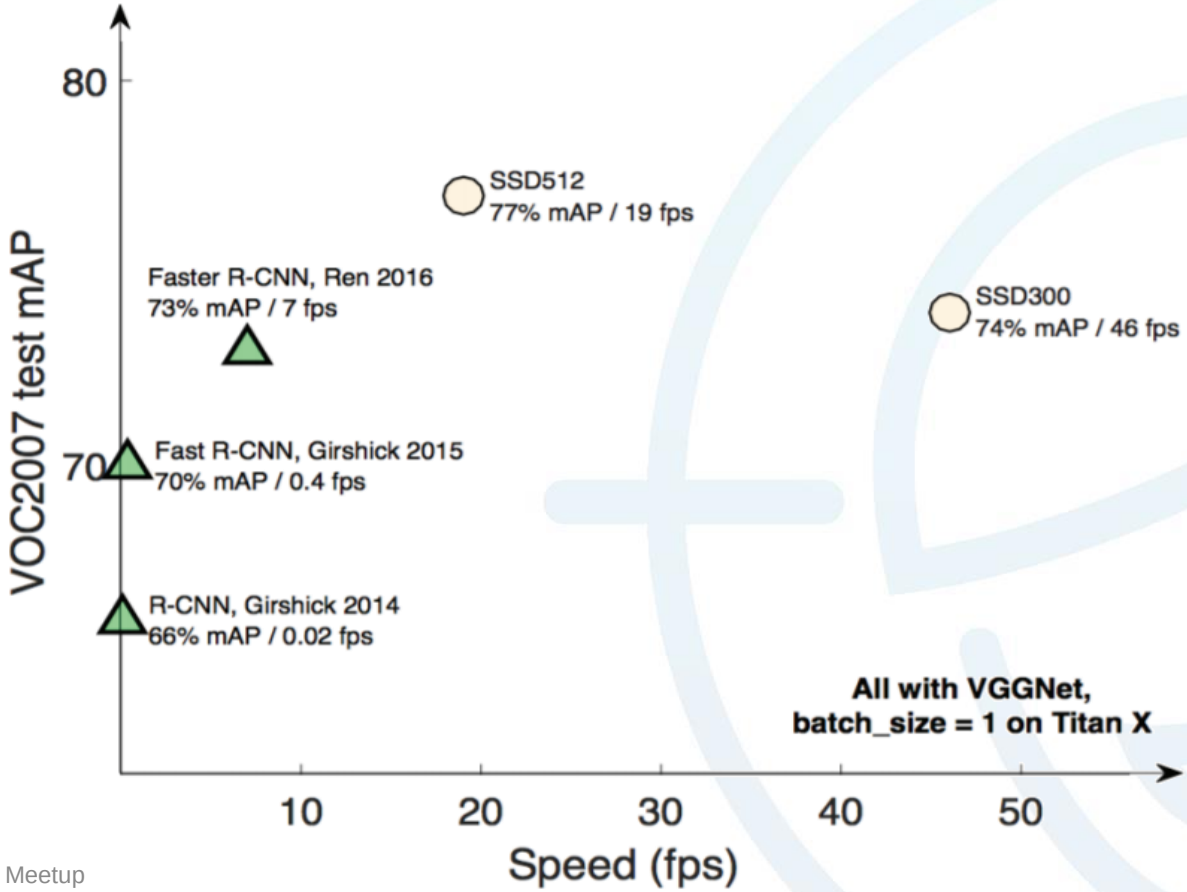
- For each anchor box is predicted:
  - 4 coordinates shape offset (loc)
  - N+1 object category scores (conf)

SSD DCNN / Training



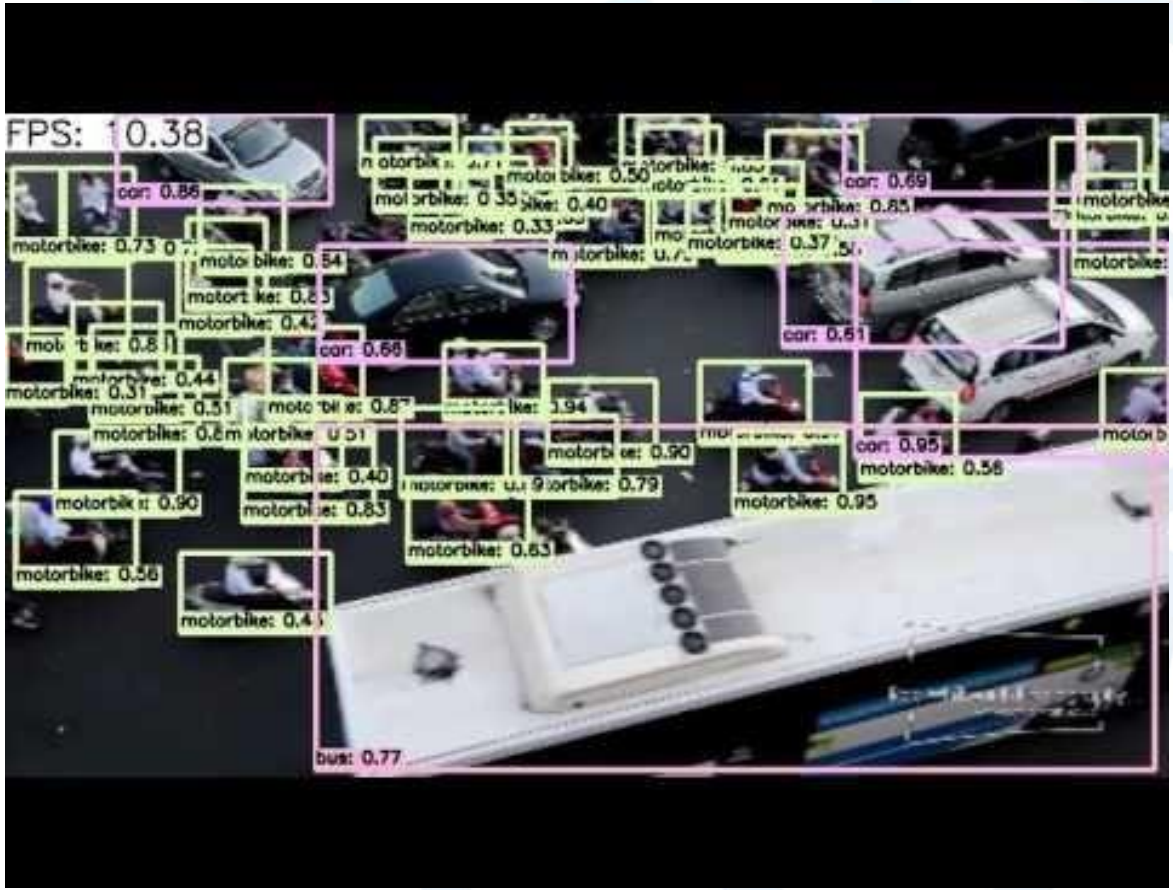
- Training:
  - End-2-End
  - Loss = weighted sum of localization loss (Smooth L1) and confidence loss (Softmax)
  - Negative samples is much more => Hard Negative Mining (Positive : Negative = 1 : 3)
  - Data augmentation helps : horizontal flip / random crop & color distortion / random expansion

SSD vs. R-CNN vs. Fast(er) R-CNN



SSD, Resume

- Pros:
  - No pixels / features **resampling** for bounding box hypotheses (as *R-CNN)
  - Very fast
  - End-to-end training
- Cons:
  - Almost as accurate as Faster R-CNN
  - Still not realtime on embeded

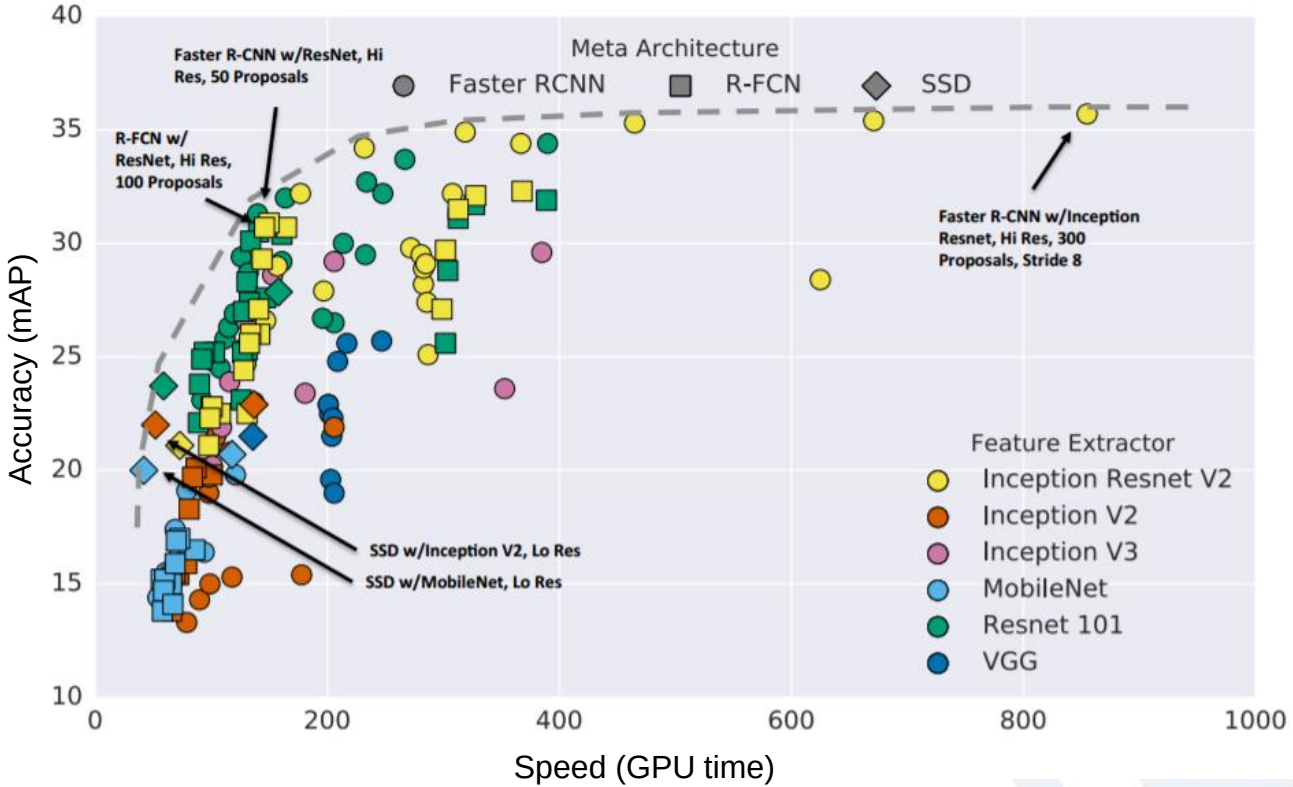


Object detection vs. Face detection

INNOVATRICS
YOUR BIOMETRIC PARTNER

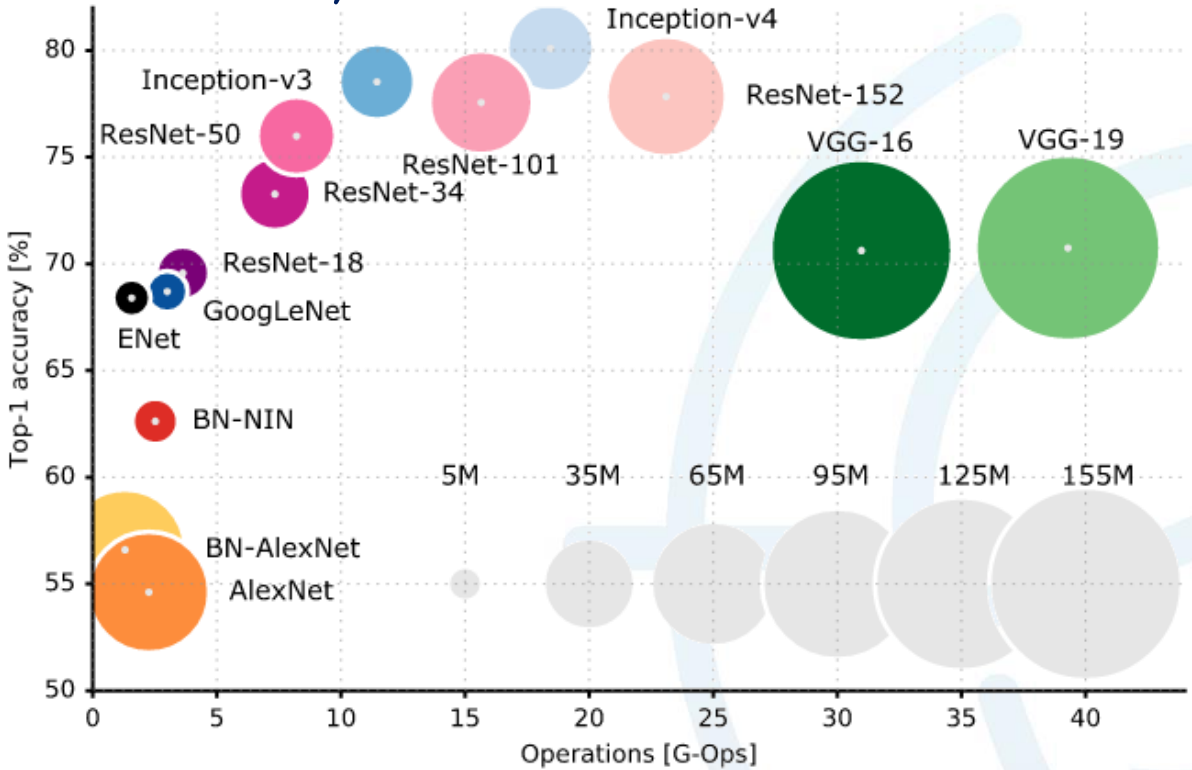
- Object detection
  - Many classes + Background
  - No surrounding context
  - Complex task needs complex features => deep slow network
- Face detection
  - Few face orientations + Background
  - If there is face, there should be body => face neighborhood analysis helps
  - Simpler task needs simpler features => shallow fast network

Detection. Architecture selection



- Huang, Jonathan, et al.: "Speed/accuracy trade-offs for modern convolutional object detectors." *CVPR 2017,* Google

Classification, Architecture selection



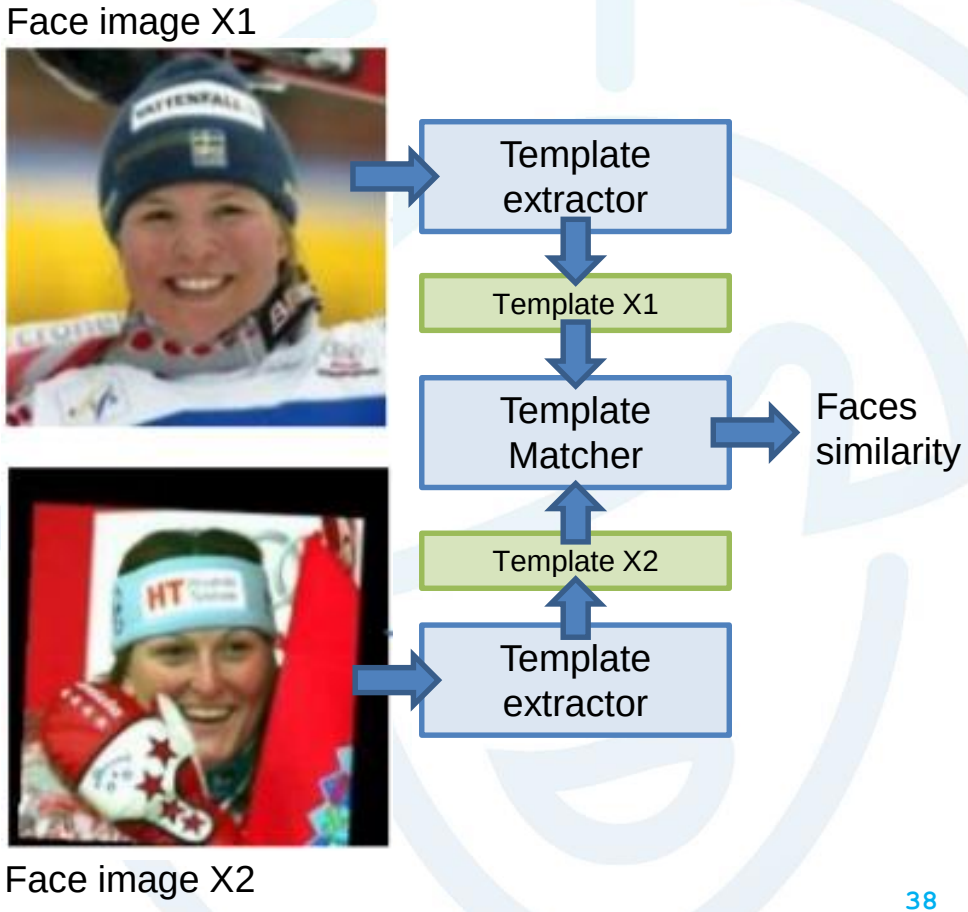
- Canziani, Paszke, Culurciello: "An Analysis of Deep Neural Network Models for Practical Applications." *arXiv preprint, 2017*

Face verification

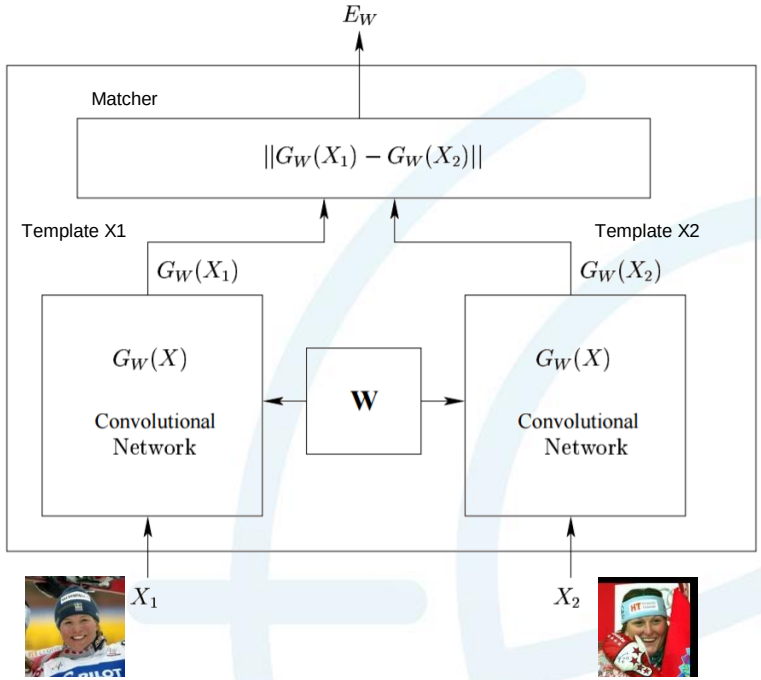
- Two facial images contain face of the same person or not
- Is it the same person ?
  - Yes, Janica Kostelic (LFW dataset)
- We need to measure the similarity !

Face template extraction

- Face template
  - Representation appropriate for effective and accurate face verification
  - Feature vector
- Face template extraction
  - Feature vector extraction
- Face template matching
  - We need very fast measurement of similarity (distance) of two face template
  - Millions of matches per second on one PC CPU core
  - Euclidean distance ideal for matching
- Face similarity
  - If higher then threshold => same person

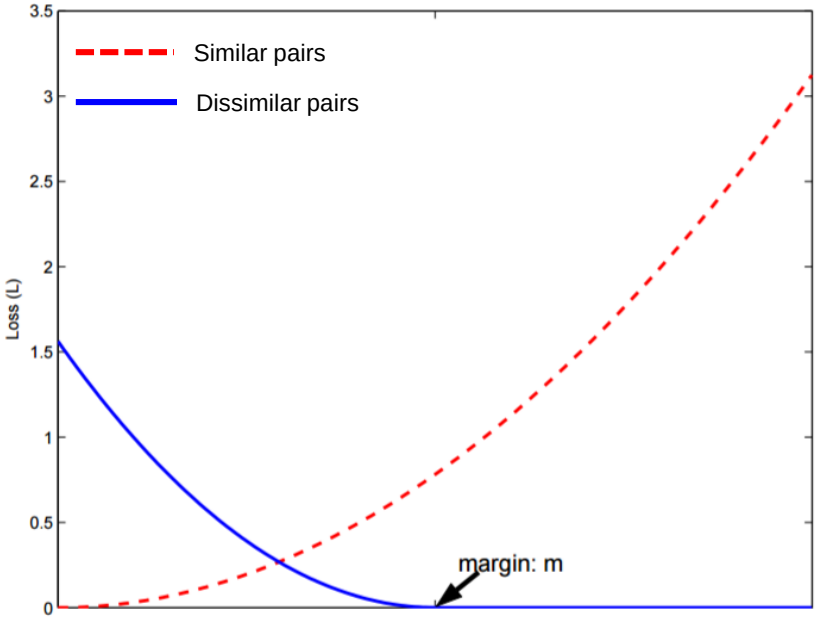


Siamese Network



- Chopra, Hadsell, LeCun: ”Learning a similarity metric discriminatively, with application to face verification”. *CVPR 2005*

Contrastive loss

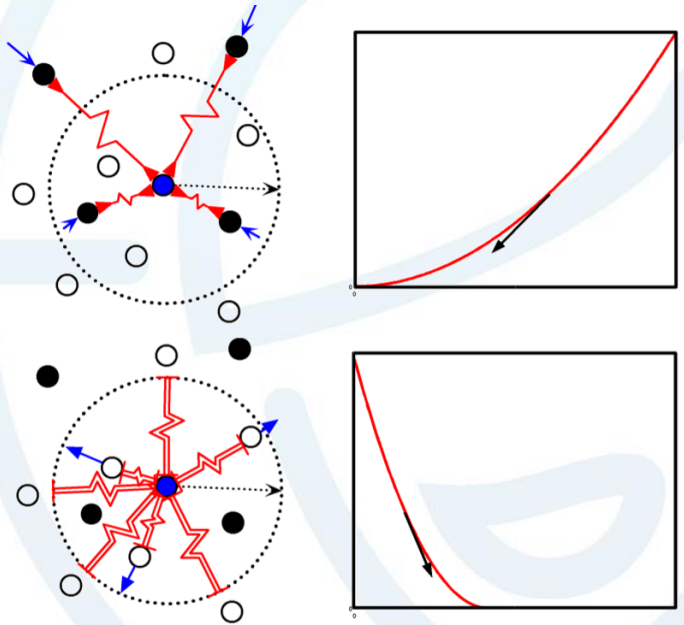


$$D_W(\vec{X}_1, \vec{X}_2) = \|G_W(\vec{X}_1) - G_W(\vec{X}_2)\|_2$$

$$(1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

Where

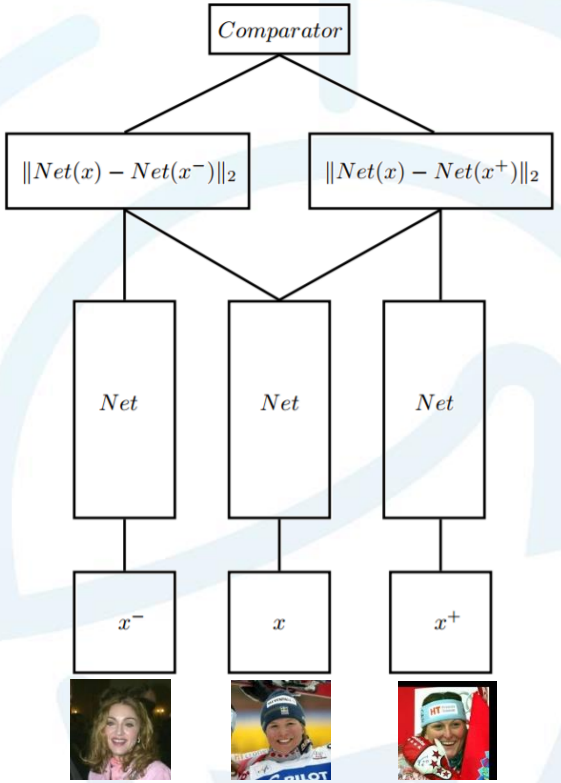
- Y = 0 for similar pairs
- Y = 1 for dissimilar pars



Triplet Network & Loss

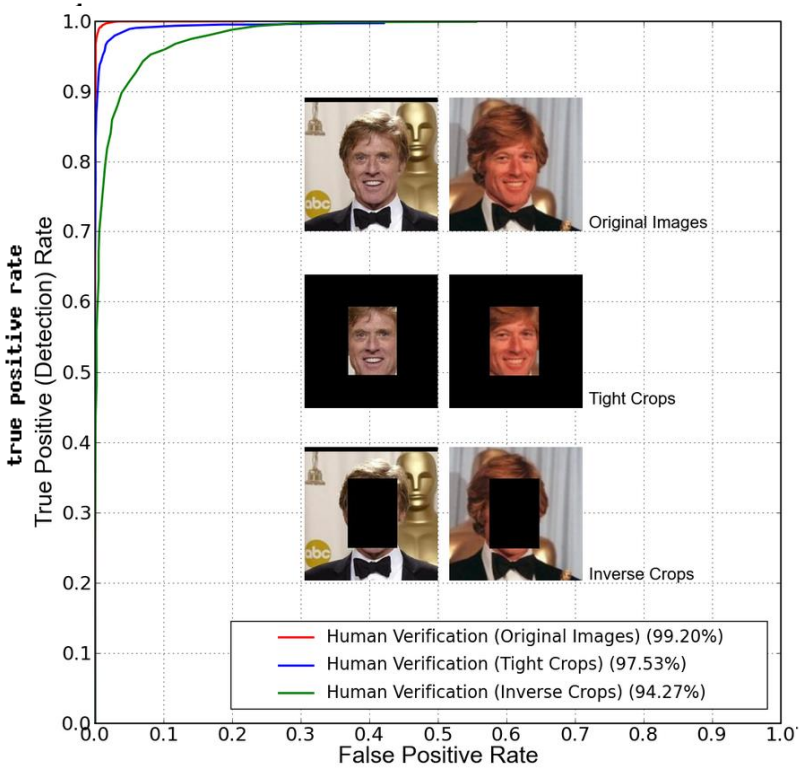
- Wang, Jiang, et al. "Learning fine-grained image similarity with deep ranking." *CVPR 2014, Google*
- Schroff, Kalenichenko, Philbin: “Facenet: A unified embedding for face recognition and clustering”, CVPR 2015, Google

$$L(x, x^{\pm}) = \left[\|Net(x) - Net(x^{+})\|_2^2 - \|Net(x) - Net(x^{-})\|_2^2 + \alpha \right]$$



- Training:
  - Millions of matching / non-matching faces
  - The harder negative face the better => hard negatives mining

Face Verification, Results



IFace – Face Recognition

- Template extraction modes
  - Accurate (more complex, slower, bigger) mode
  - Fast (lower accuracy, smaller) mode, optimized for high performance even on embedded platforms (100ms on IPhone)
- Trained on normalized images :
  - Face size / orientation / position (using info from face detection)
- Robust against various
  - lighting / quality / head position / expression / …
- Face verification
  - State of the art accuracy
  - Very fast matching – 1M matches / sec / core
- Tools
  - Caffe, Numpy, OpenCv







QUESTIONS ?

Thank you

Marián Beszédeš
Image Processing Team Leader at
Innovatrics

marian.beszedes@innovatrics.com
www.innovatrics.com

Sources

- https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/
- https://devblogs.nvidia.com/parallelforall/deep-learning-nutshell-core-concepts/
- http://www.eccv2016.org/files/posters/O-1A-02.pdf
- https://www.slideshare.net/xavigiro/faster-rcnn-towards-realtime-object-detection-with-region-proposal-networks
- http://vision.ia.ac.cn/zh/senimar/reports/Siamese-Network-Architecture-and-Applications-in-Computer-Vision.pdf