Machine Learning - Deep Learning fundamentals
(69152) **CNNs**

Master in Robotics, Graphics and Computer Vision
Ana C. Murillo

Universidad Zaragoza

Reminders

- **Reading Lab (Lab 4)**
  - Pick a paper (tomorrow)
  - read it
  - prepare a presentation for your lab session

Today

- Open problems with deep learning?
- More well-known architectures

Compiled “open problems” from moodle answers

Ethics and Fairness : Is it possible to avoid bias? “Responsible AI”? External seminar
Explainability : finer-grain, probabilities, …
Generalization - Adaptability : Incremental, “foundation-models”
Efficiency (time-memory) : efficient architectures, ENERGY?
Efficiency (data-requirements) : less supervision
Multi-modal
Meta-learning
Robustness

Open or interesting problems related to deep learning?

perform more complex tasks?

More architectures …

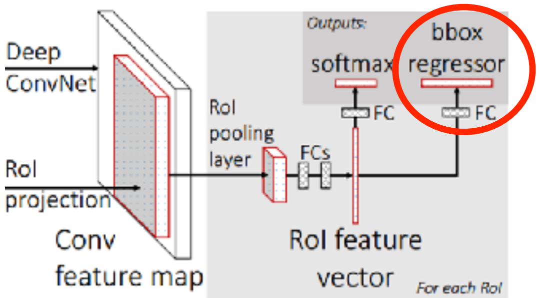
- Not only classification —> detection?

Deep Learning && Regression

- **Detection**

predict normalized b-box coordinates

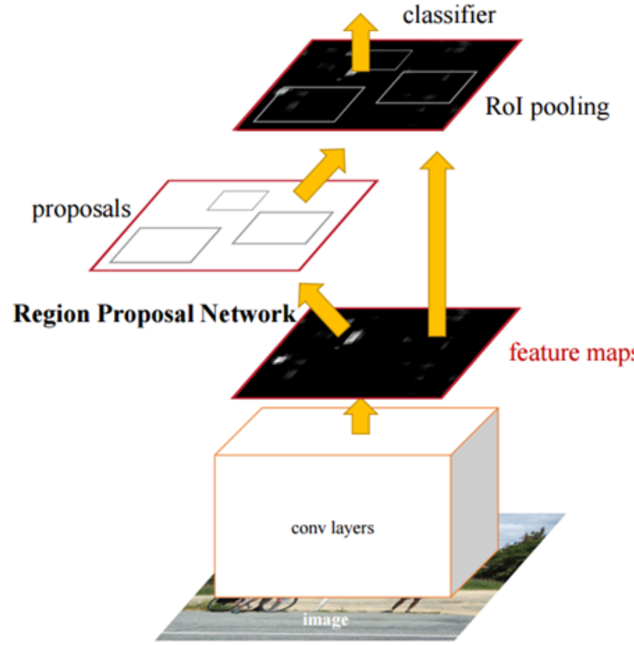
$p = (x_1, y_1, x_2, y_2)$



Fast R-CNN. Ross Girshick. 2015

Different versions of Region-CNN:

- R-CNN - 2013
- Fast R-CNN - 2015
- Faster R-CNN - 2016



CLASSIFICATION + DETECTION

Analyze feature maps (activations) to learn where the objects are

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
Shaoqing Ren, Kaiming He, **Ross Girshick**, and Jian Sun. 2016

Deep Learning && Regression

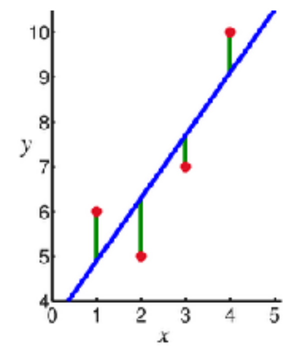
- Linear regression: relationship between a scalar (y) and one or more explanatory variables (x)

$$y_i=\beta_0 1+\beta_1 x_{i1}+\beta_2 x_{i2}+ \dots +\beta_p x_{ip}+\varepsilon_i = \mathbf{x}^{\mathsf{T}}_{\mathbf{i}}\boldsymbol{\beta}+\varepsilon_i, \quad i=1, \dots ,n$$

“map” to a continuous output (regression)

VS

“map” to a discrete output (classification)



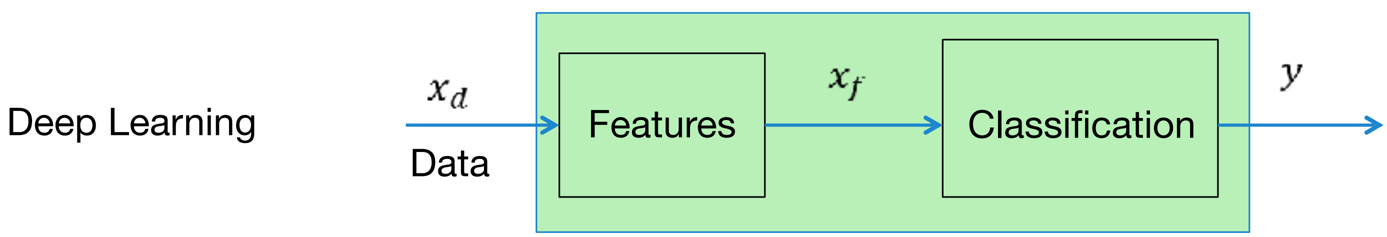
Wikipedia:
File:Linear least squares example2.png

Deep Learning && Regression

- Regression of **Pose** - *PoseNet*
  - Train the network to output 3D position (x) and orientation (q) (7 dims). Loss function:

$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2$$

  - Modify GoogLeNet:



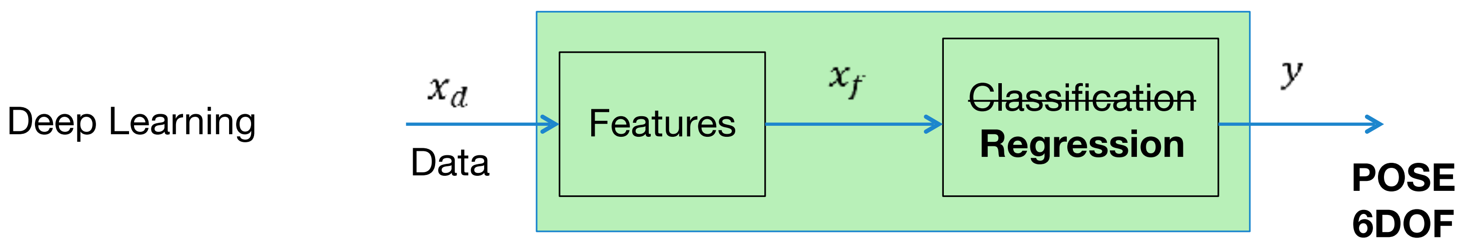
PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.
Alex Kendall, Matthew Grimes and Roberto Cipolla. ICCV 2015

Deep Learning && Regression

- Regression of **Pose** - *PoseNet*
  - Train the network to output 3D position (x) and orientation (q) (7 dims). Loss function:

$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2$$

  - Modify GoogLeNet. Key change: replace softmax classifiers with affine regressors (each final fully connected layer now outputs a pose vector of 7-dims).



http://mi.eng.cam.ac.uk/projects/relocalisation/

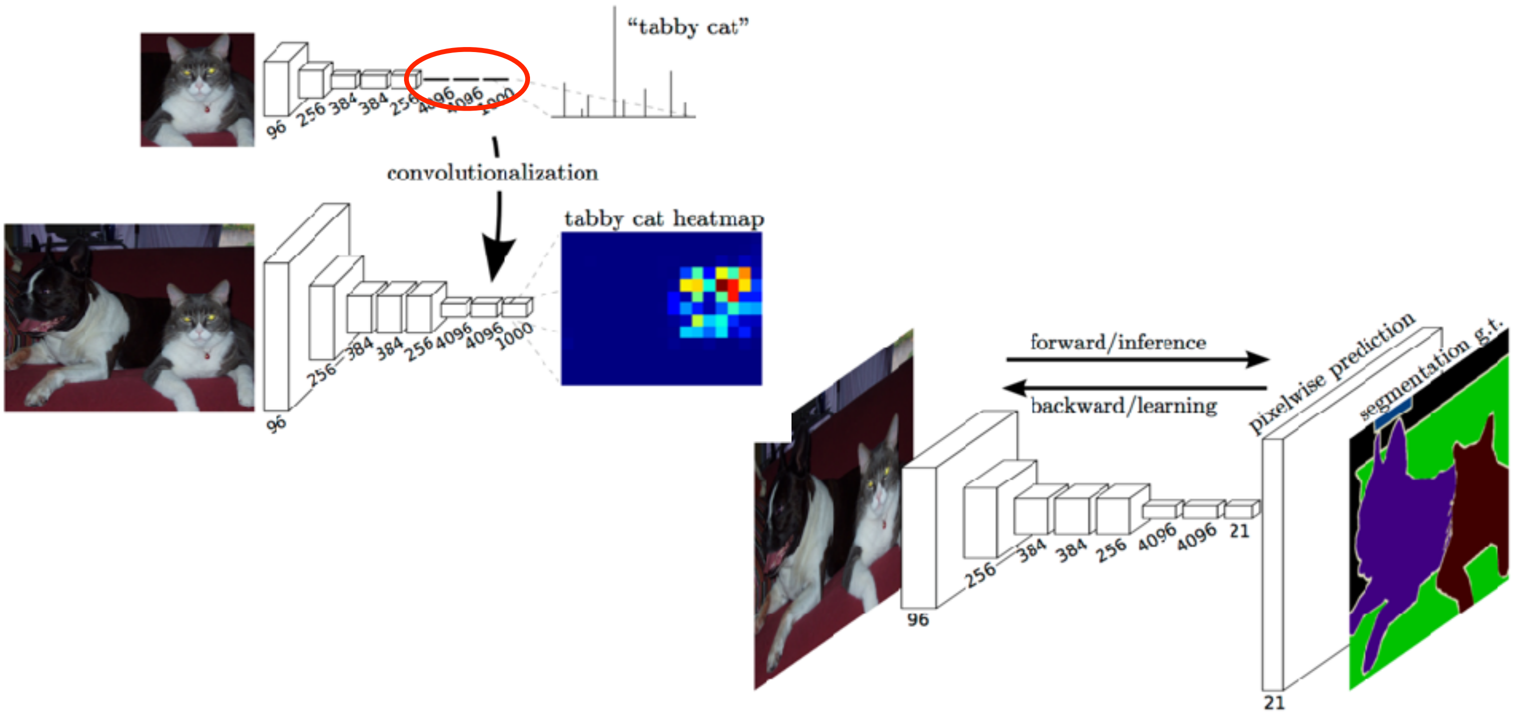
PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. Alex Kendall, Matthew Grimes and Roberto Cipolla. ICCV 2015

More architectures …

More accurate image understanding?

Deep Learning && pixel classification

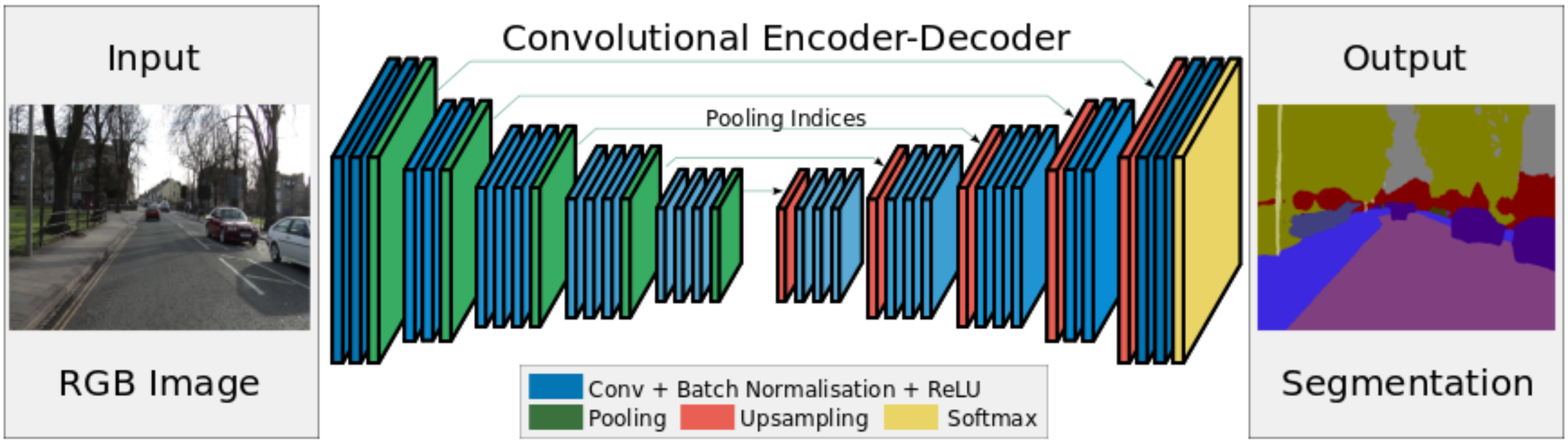
- Dense labeling/**Segmentation** - Fully Convolutional Net (FCN)



Fully Convolutional Networks for Semantic Segmentation
J. Long*, E. Shelhamer* and T. Darrell CVPR 2015 and PAMI 2016

Deep Learning && pixel classification

- Dense labeling/**Segmentation** - Encoder-Decoder



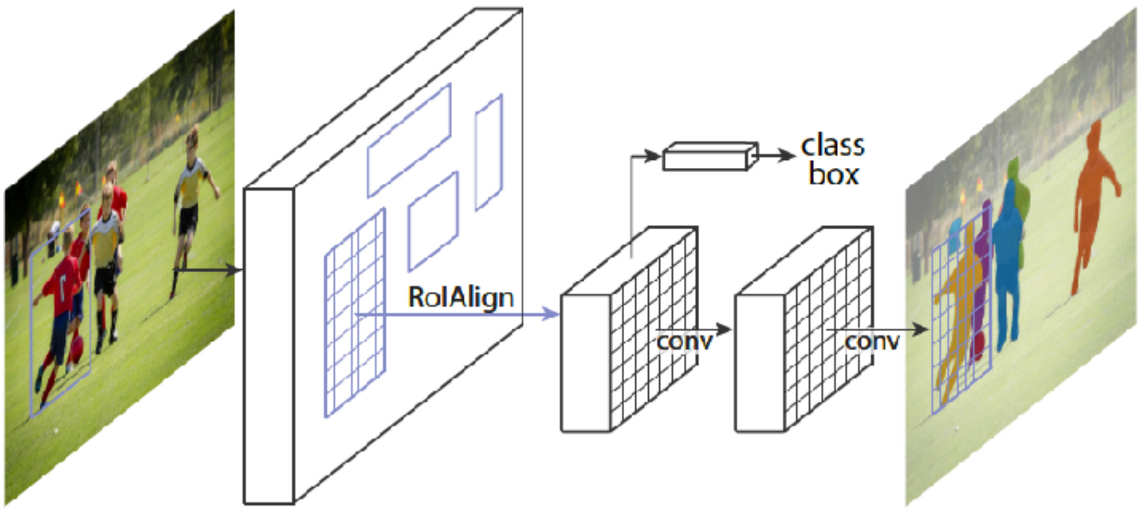
http://mi.eng.cam.ac.uk/projects/segnet/

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.
Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla. PAMI, 2017.

Deep Learning && pixel classification

- **Detection** + **Instance Segmentation**
  Mask R-CNN - 2017

(R-CNN + Semantic Segmentation)



Mask R-CNN.
He, K., Gkioxari, G., Dollár, P., & Girshick, R. ICCV 2017.

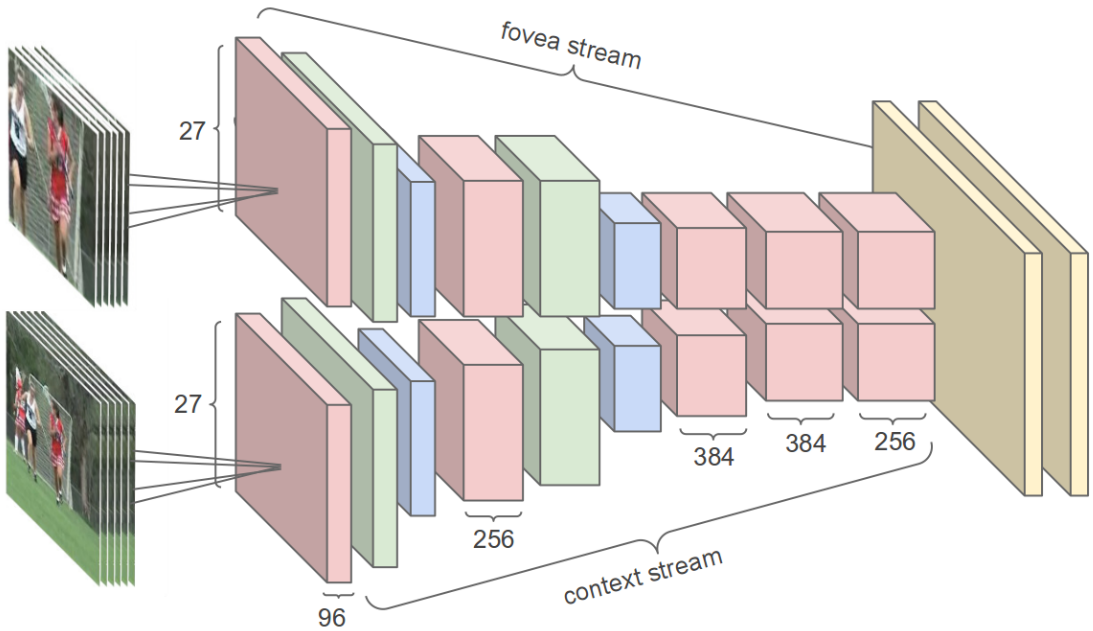
More architectures …

- More details? Interaction? Time?
- More “complex” input data

Deep Learning && Video

- Fuse multiple frame info

- high-resolution center crop + low-resolution full image (reduced input size, more efficient training)

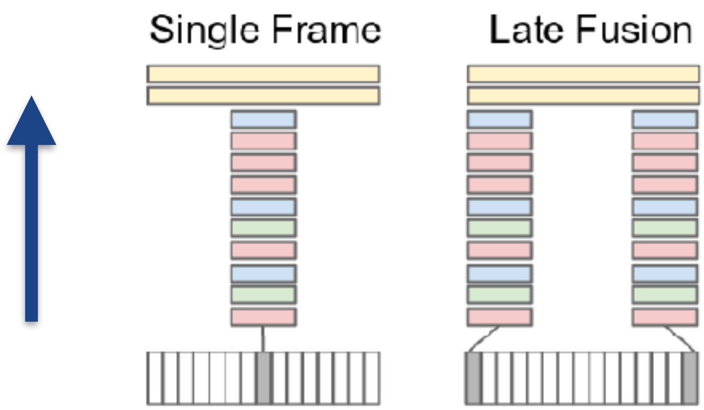


http://cs.stanford.edu/people/karpathy/deepvideo/

Large-scale Video Classification with Convolutional Neural Networks.
Andrej Karpathy et al. CVPR 2014.

Deep Learning && Video

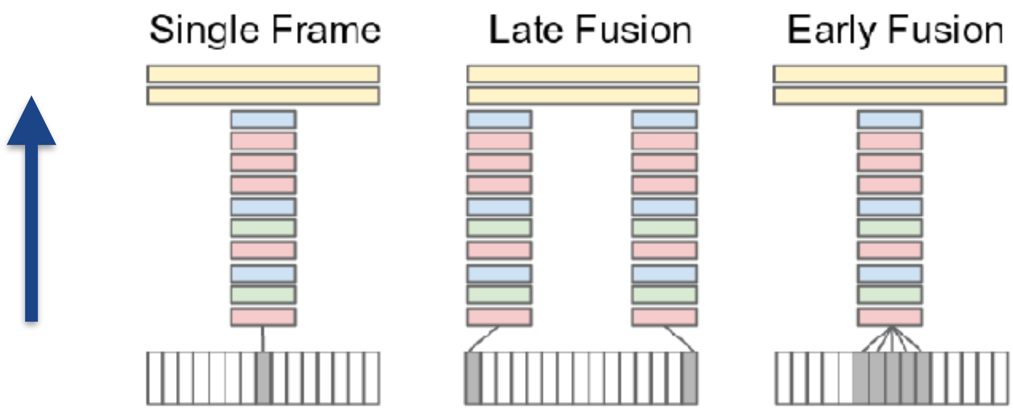
- Fuse multiple frame info. Many strategies



Large-scale Video Classification with Convolutional Neural Networks.
Andrej Karpathy et al. CVPR 2014.

Deep Learning && Video

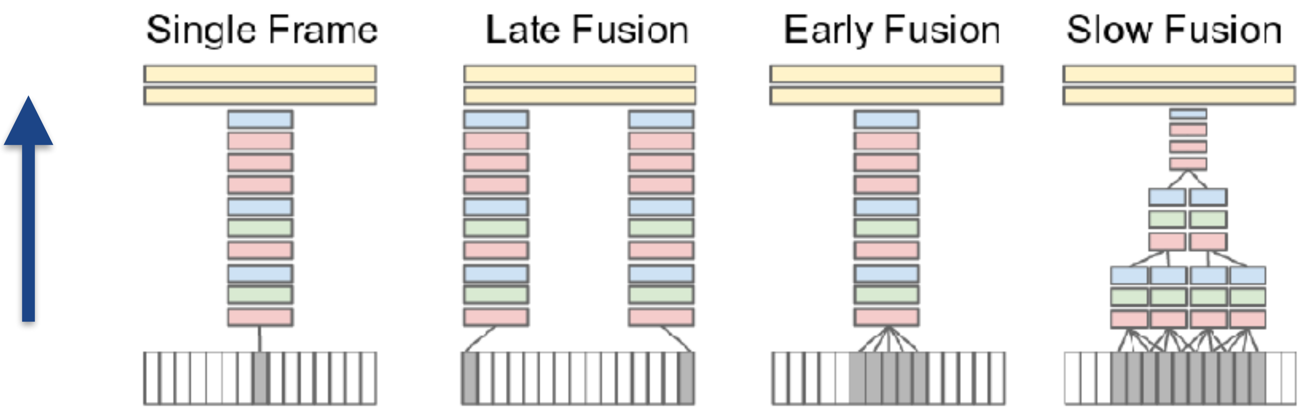
- Fuse multiple frame info. Many strategies



Large-scale Video Classification with Convolutional Neural Networks.
Andrej Karpathy et al. CVPR 2014.

Deep Learning && Video

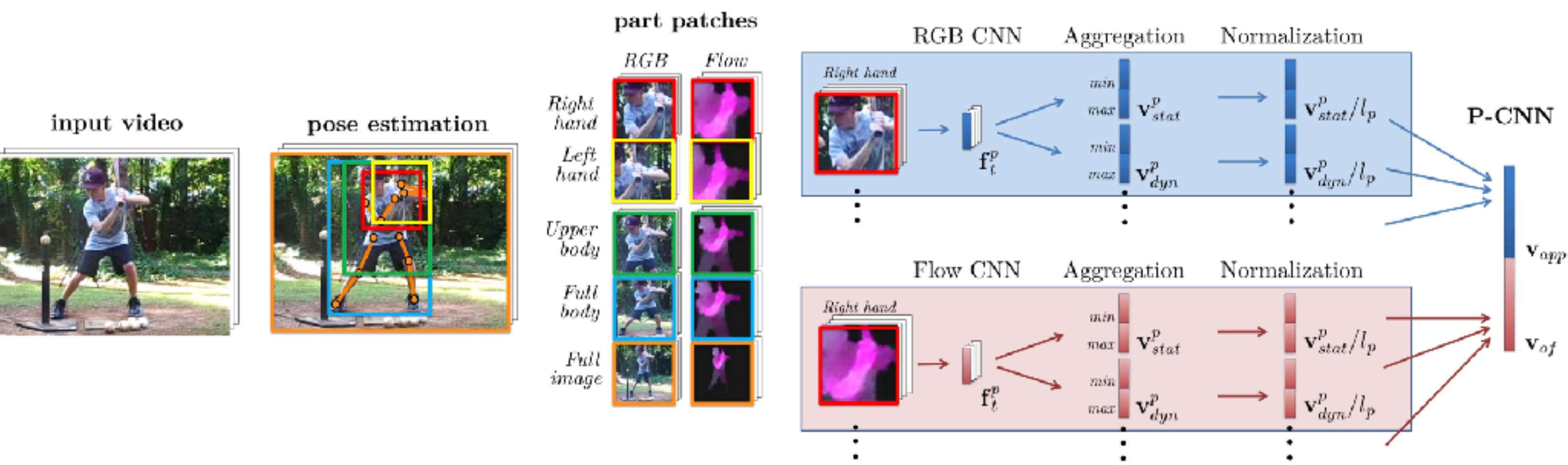
- Fuse multiple frame info. Many strategies



Large-scale Video Classification with Convolutional Neural Networks.
Andrej Karpathy et al. CVPR 2014.

Deep Learning && Multi-Modal

- Multi-modal input from images
  - Deep learning features: Pose-based CNN
  - Combine **parts**, **pose** and **flow** with CNNs (aggregates motion and appearance information)



P-CNN: Pose-based CNN Features for Action Recognition
G. Chéron, I. Laptev and C. Schmid; in Proc. ICCV'15

Deep Learning && Video

Do you see any problem with this way of treating sequential/video data?

Deep Learning && Video

Do you see any problem with this way of treating sequential/video data?

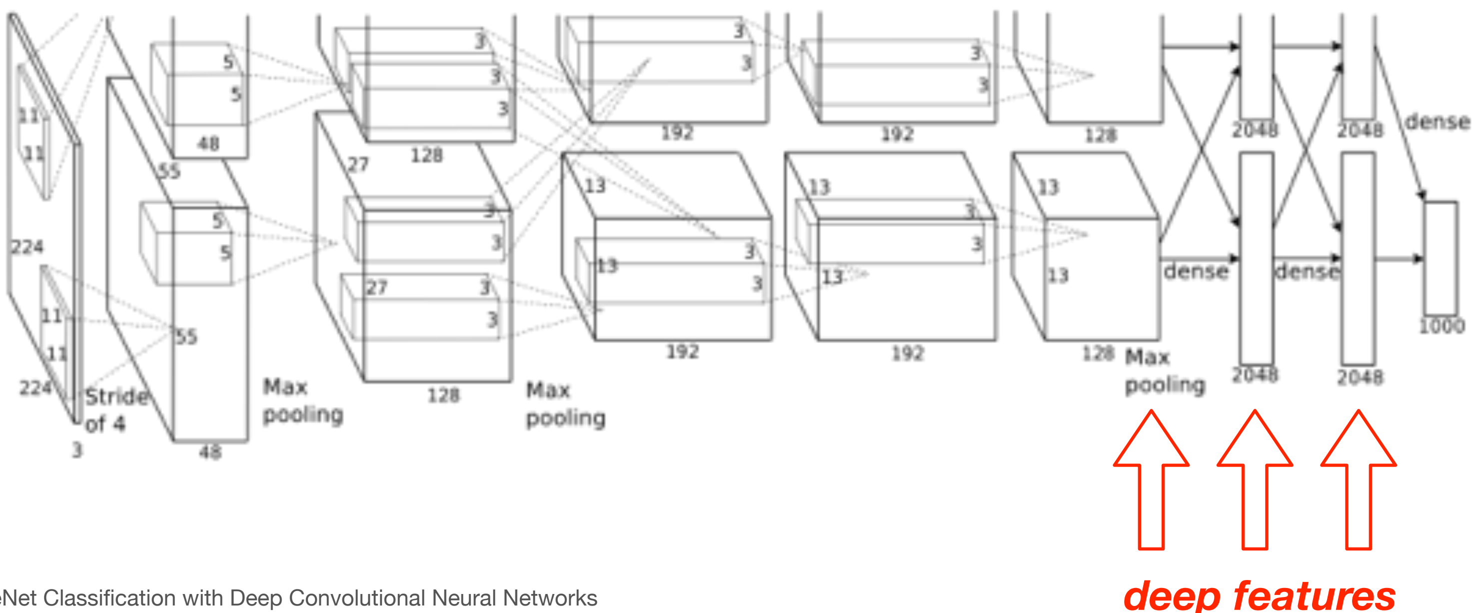
Later in the course: RNN, TCN, Transformers, …

(More modern-adequate architectures to deal with sequential and multi-modal data)

Summarizing: CNNs & Transfer Learning

- CNNs are able to generalize well!
  - great **features**
  - **fine-tuning**

Lab 3
you'll practice
some of this

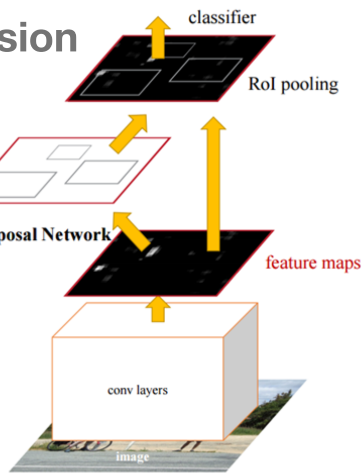


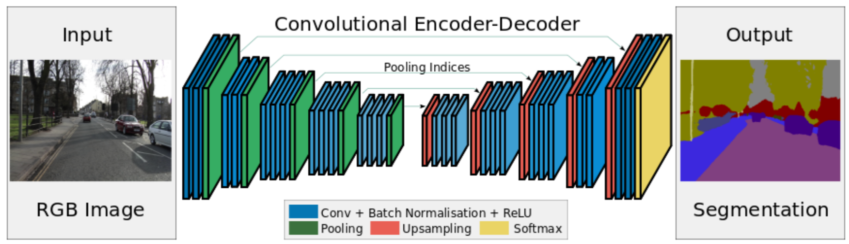
ImageNet Classification with Deep Convolutional Neural Networks
Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. NIPS 2012

Summarizing - More architectures …

- CNN - Not only image-classification

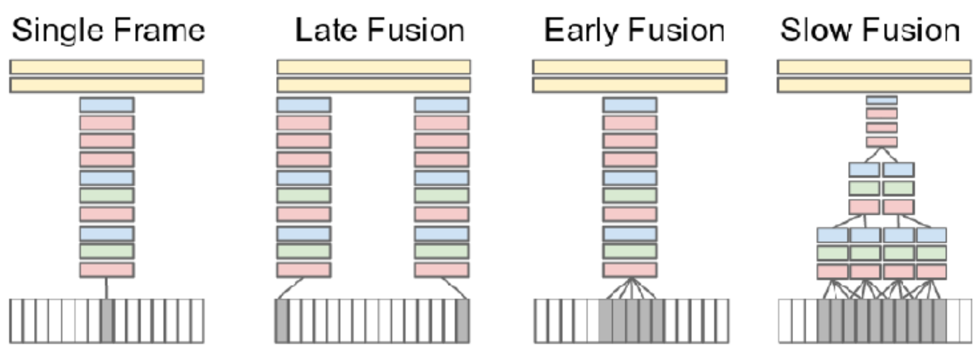
regression





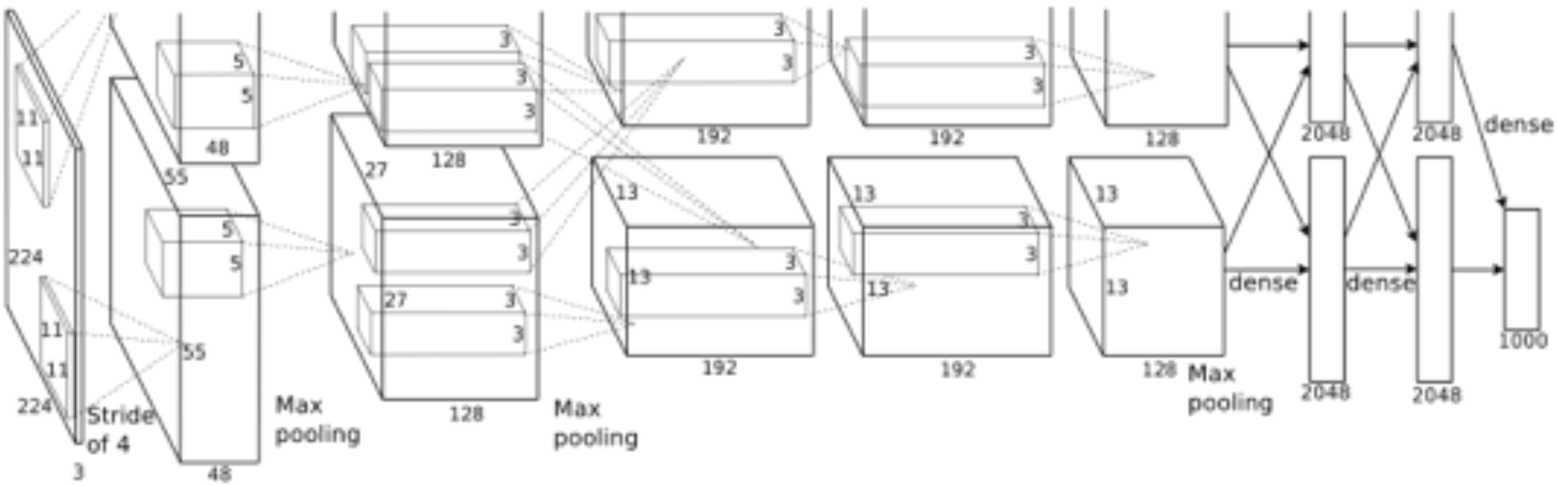
classification per pixel

fusion of multiple "sources"



Examples to understand CNNs

- **Params? feature map size?**
  - How many params does the 2nd conv. layer have? and the 5th?



  - Input of 240x240x3; Conv1 (48 kernels, 3x3) - Pooling (stride 2) - Conv2 (48 kernels, 3x3) —> size of feature map after Conv2?

Demos to understand CNNs

- karpathy : DEMO online, CNN
- Playground tensorflow
- visualisation (places CNN)

  http://people.csail.mit.edu/torralba/research/drawCNN/drawNet.html

Visualizing and Understanding Convolutional Neural Networks
Matthew Zeiler and Rob Fergus. **2013**

Later …

- More advanced models
- … and different supervision strategies
    - DRL
    - Unsupervised
    - Recurrent architectures

Bibliography - Resources for some of the materials today

- Stanford classes on deep learning for Computer Vision (http://cs231n.stanford.edu) and Deep Learning (https://cs230.stanford.edu/)
- Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org
- Deep Learning Summer School Montreal: https://mila.quebec/en/cours/deep-learning-summer-school-2017/