Deep Learning in Visual Recognition

Yadong Mu
Machine Intelligence Lab
Institute of Computer Science & Technology
Peking University
Email: myd@pku.edu.cn
Website: www.muyadong.com

Outline

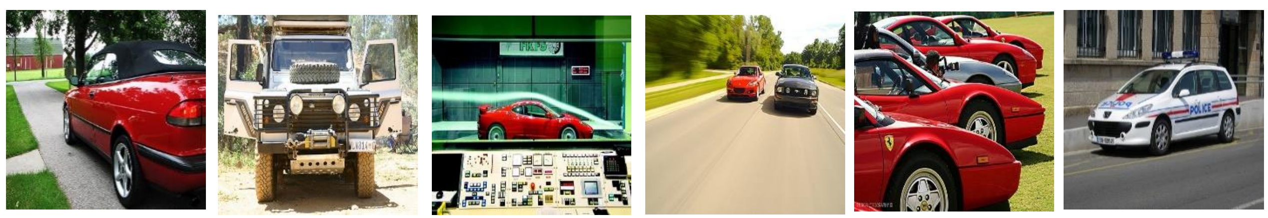
- **Visual Recognition: Task Definition and Challenges**
- **Bag-of-words Models**
- **Improving BoW Models**
- **Why DL suddenly works? (AlexNet, 2012)**
- **Can it go deeper? (ResNet, 2015)**

Visual Recognition / Image Classification

- Give a binary label to indicate whether an object is present

Does this image contain a car?

cars



not cars

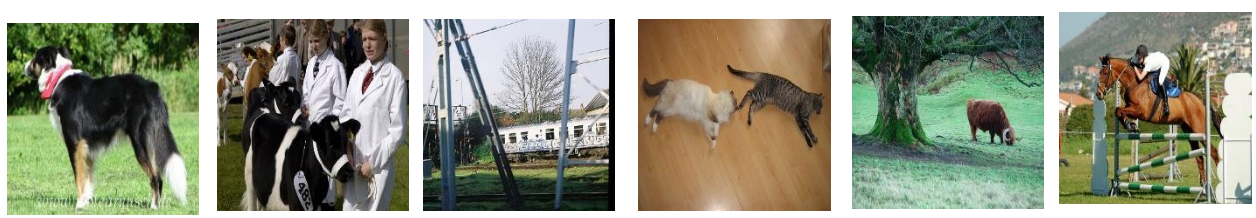


Image Classification

- 1st challenge: **semantic gap**

骑士神射里程碑夜受伤：很庆幸 没有更糟糕

2017年02月26日 15:42:58

来源：凤凰体育

0人参与 0评论

凤凰体育讯（记者朱俊清克里夫兰前方报道）科沃尔在今晚对阵公牛的比赛中出战24分钟，三分球7中4，拿下14分。并且在今晚超过了卡特和克劳福德，成为生涯命中三分球数榜的第5名。在今天的比赛中，他第四节膝盖受伤下场。

赛后他在更衣室接受了采访，记者都很关心他的伤势，他表示：“我感觉还不错，挺幸运的，今晚本可能会更糟糕，我落地时腿绷直了，然后膝盖有一些疼，你们都见过很多更糟糕的受伤，还好我没有太大的问题。”他表示自己明天会接受检查，但是应该不会错过周一的比赛，他表示自己不想错过任何一场比赛。记者问到关于在生涯命中三分球数榜上升一名有什么感想。他说：“有一些兴奋吧，我当然会享受这样的时刻。我希望继续在这个榜上提升名次。然后我也很骄傲自己能够在这个榜上取得这样一个名次，算是给自己这么多年一直不断的投篮的一个认可。”

汤普森也接受了简短的更衣室采访。记者问到：“当你知道詹姆斯今晚将会缺席时，你们有什么样的准备？”他说：“就是下一个球员顶上，准备好上场。赛季很长，总会有球员缺席，受伤或者生病，这只是比赛的一部分。我们只需要时刻准备好比赛。”他也评价了德隆的到来，他说：“德隆来对球队是一个很大的补充，他能投能突，他会让我们球队更强。”

Image Classification

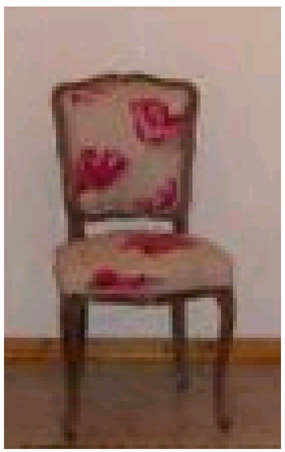
- 1st challenge: **semantic gap**

To the computer, images are essentially a 3-D array

Image Classification

- 1st challenge: **semantic gap**

This is a chair



Find the chair in this image

Output of normalized correlation

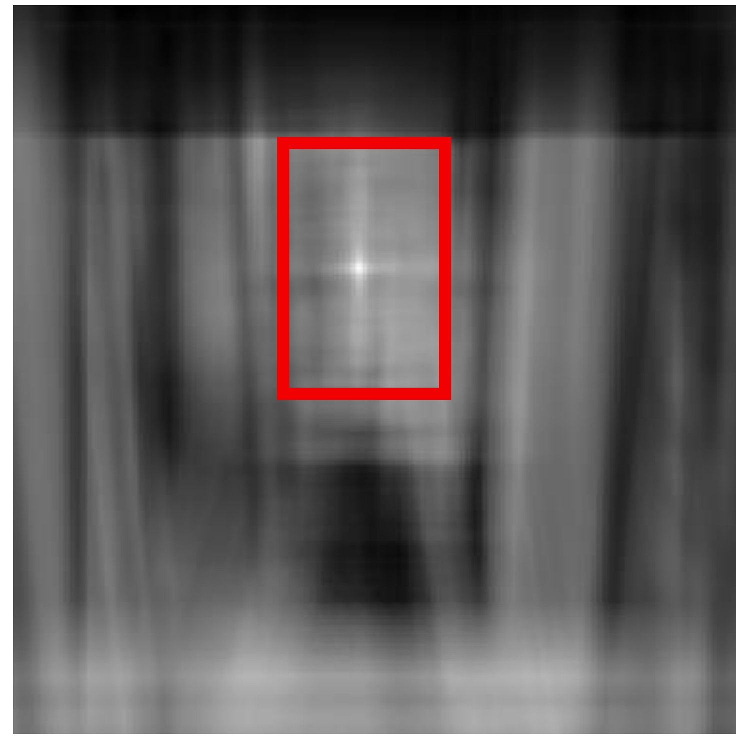
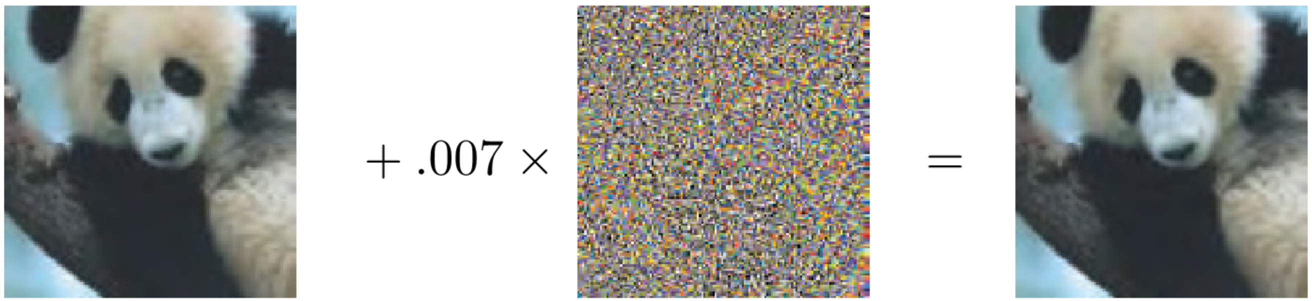


Image Classification

- Another example: adversarial training



$$\boldsymbol{x} \quad + .007 \times \quad \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)) \quad = \quad \boldsymbol{x} + \epsilon\, \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

$y =$"panda" w/ 57.7% confidence

"nematode" w/ 8.2% confidence

"gibbon" w/ 99.3% confidence

Figure 7.8: A demonstration of adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Reproduced with permission from Goodfellow et al. (2014b).

Image Classification

- 2^{nd} challenge: visual variations

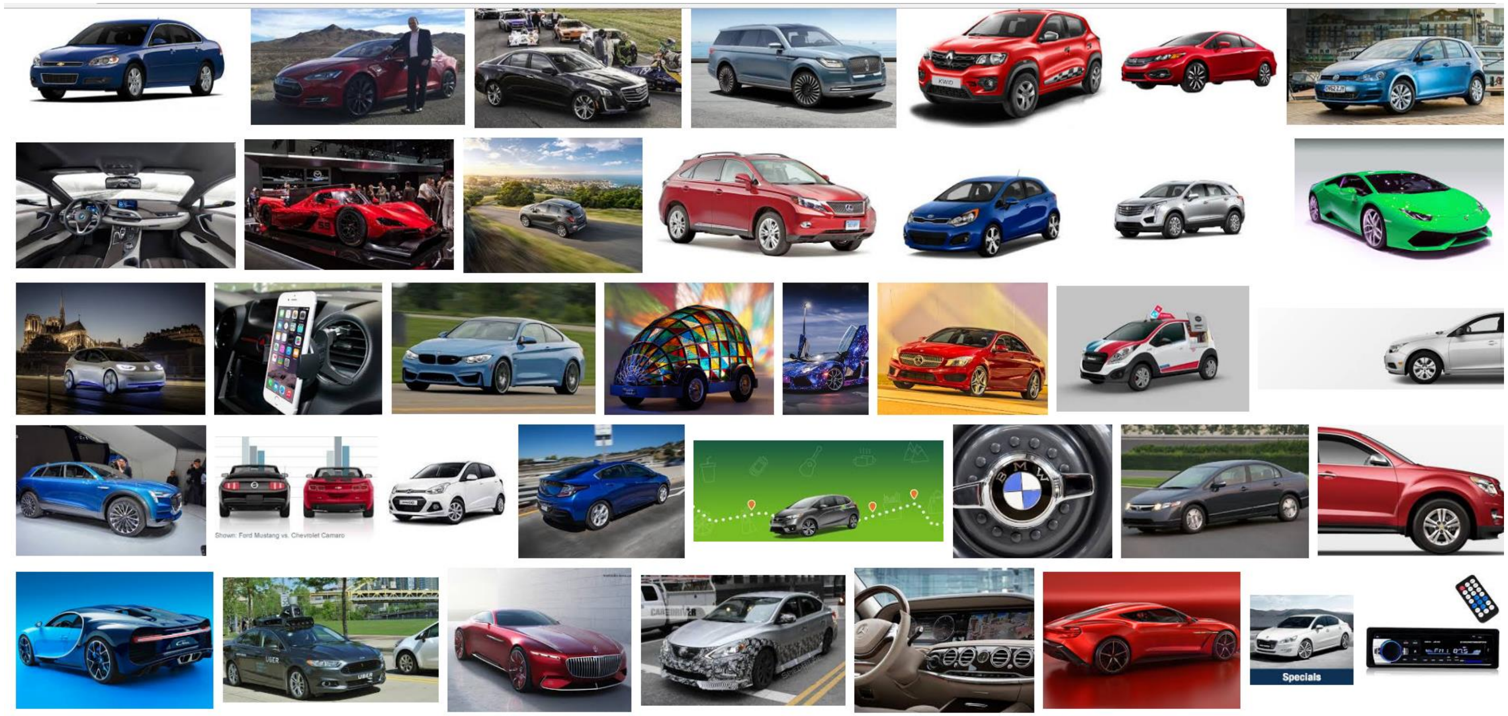


Image Classification

- 3^{rd} challenge: multiple labels

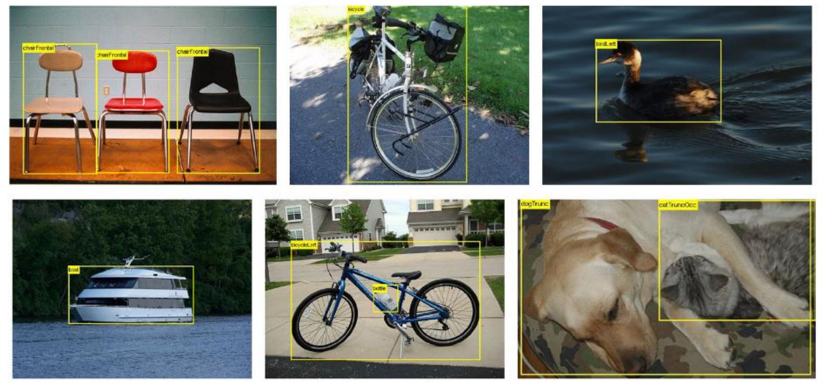
What kind of objects do you see in this image?

Image Classification @ CV Community

- Pascal VOC Challenges 2005-2012

Table 1 The VOC classes

Vehicles	Household	Animals	Other
Aeroplane	Bottle	Bird	Person
Bicycle	Chair	Cat	
Boat	Dining table	Cow	
Bus	Potted plant	Dog	
Car	Sofa	Horse	
Motorbike	TV/Monitor	Sheep	
Train			



- Large Scale Visual Recognition Challenge (ILSVRC) 2010-2016

1000 synsets for Object classification/localization

kit fox, Vulpes macrotis
English setter
Australian terrier
grey whale, gray whale, devilfish, Eschrichtius gibbosus, Eschrichtius robustus
lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens
Egyptian cat
ibex, Capra ibex
Persian cat
cougar, puma, catamount, mountain lion, painter, panther, Felis concolor
gazelle
porcupine, hedgehog
sea lion

http://www.image-net.org/challenges/LSVRC/2016/index

IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications CoolStuff About

Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.
Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

http://www.image-net.org/

Image Classification @ Multimedia Community

- Image tagging: YFCC100M

Labeled Tags:	Predicted Tags:
sky	sky:0.073
sunset	sunset:0.069
boat	boat:0.051
bridge	bridge:0.039
water	lake:0.033

Labeled Tags:	Predicted Tags:
car	car:0.085
sky	sky:0.079
building	bus:0.055
street	street:0.039
ocean	water:0.026

Labeled Tags:	Predicted Tags:
tree	tree:0.081
grass	grass:0.070
plants	plants:0.069
animal	cow:0.047
horse	horse:0.032

Labeled Tags:	Predicted Tags:
tower	tower:0.080
bridge	bridge:0.078
water	water:0.049
clouds	sky:0.029
sky	clouds:0.025

Labeled Tags:	Predicted Tags:
boat	boat:0.075
tower	street:0.067
ocean	ocean:0.049
water	water:0.038
tree	cloud:0.025

Labeled Tags:	Predicted Tags:
sky	sky:0.073
clouds	clouds:0.06
person	person:0.05
ocean	lake:0.039
water	ocean:0.033

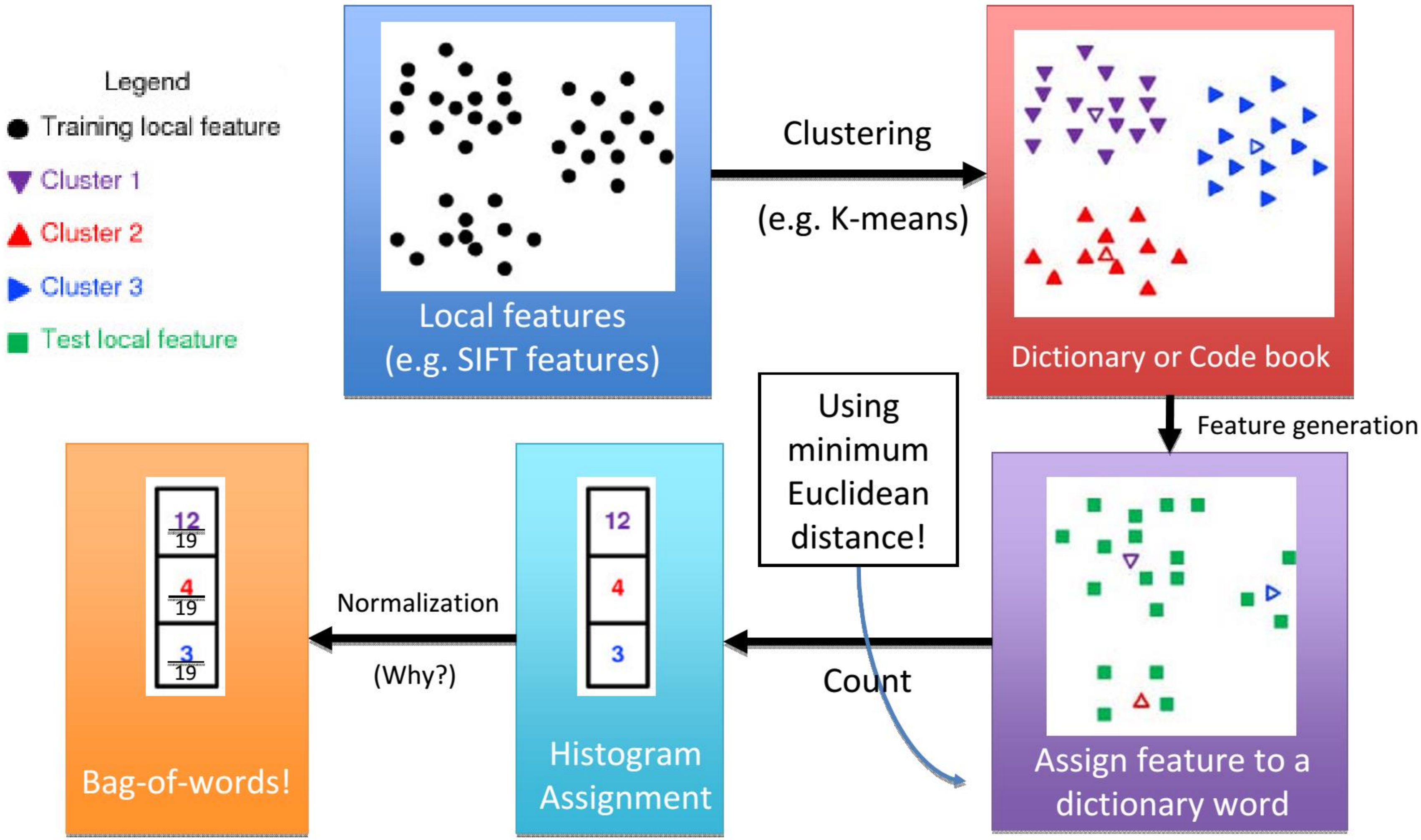
- Concept detection: TRECVID SIN task

Ontology design -> Image collecting -> Image annotation -> Concept detection

LSCOM LARGE SCALE CONCEPT ONTOLOGY FOR MULTIMEDIA

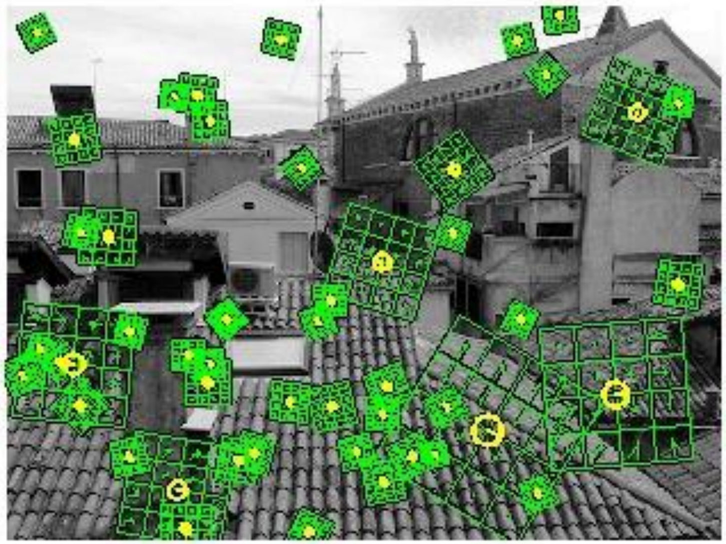
Object → **Bag of ‘words’**

Pipeline of Visual Bag-of-Words Model



Local Features – the "Words"

- SIFT (Scale-invariant feature transform)



http://www.vlfeat.org/overview/sift.html

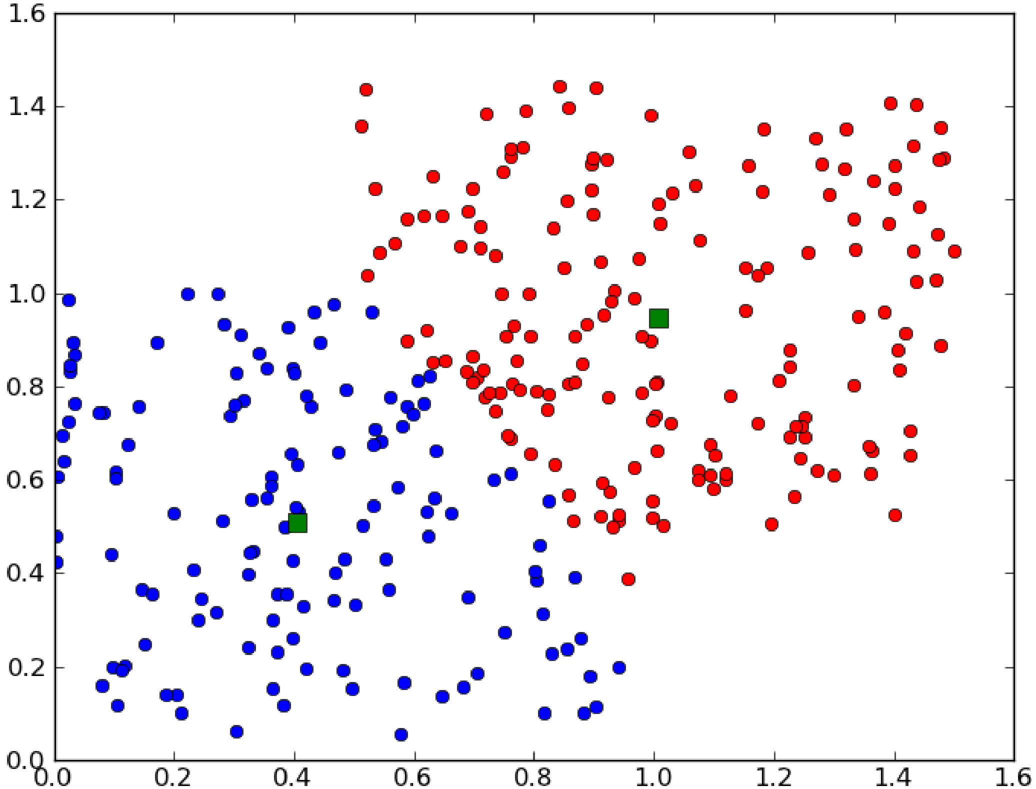
- And its variants!
  - SIFT (sift)
  - HueSIFT (huesift)
  - HSV-SIFT (hsvsift)
  - OpponentSIFT (opponentsift)
  - *rg*SIFT (rgsift)
  - C-SIFT (csift)
  - RGB-SIFT(rgbsift)

http://koen.me/research/colordescriptors

Dense SIFT

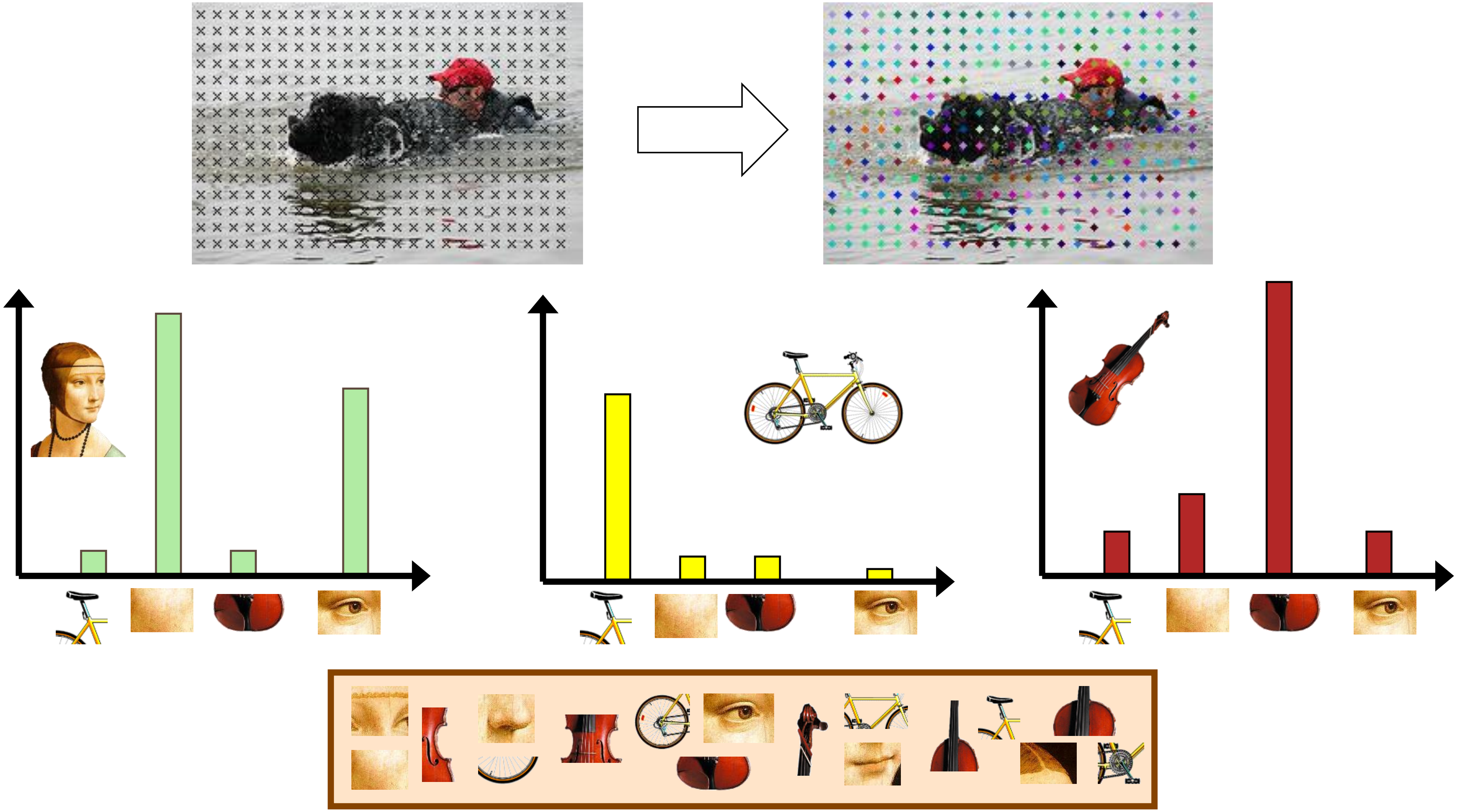
Visual Feature Clustering – the “Dictionary”

- Iterate between “assignment” and “center-updating” steps
  - Each sample is assigned to the most similar center
  - Update centers by averaging all assigned samples



Local Feature Quantization

- Each SIFT descriptor is quantized into a visual word using the nearest cluster center.



How can we improve the BoW model

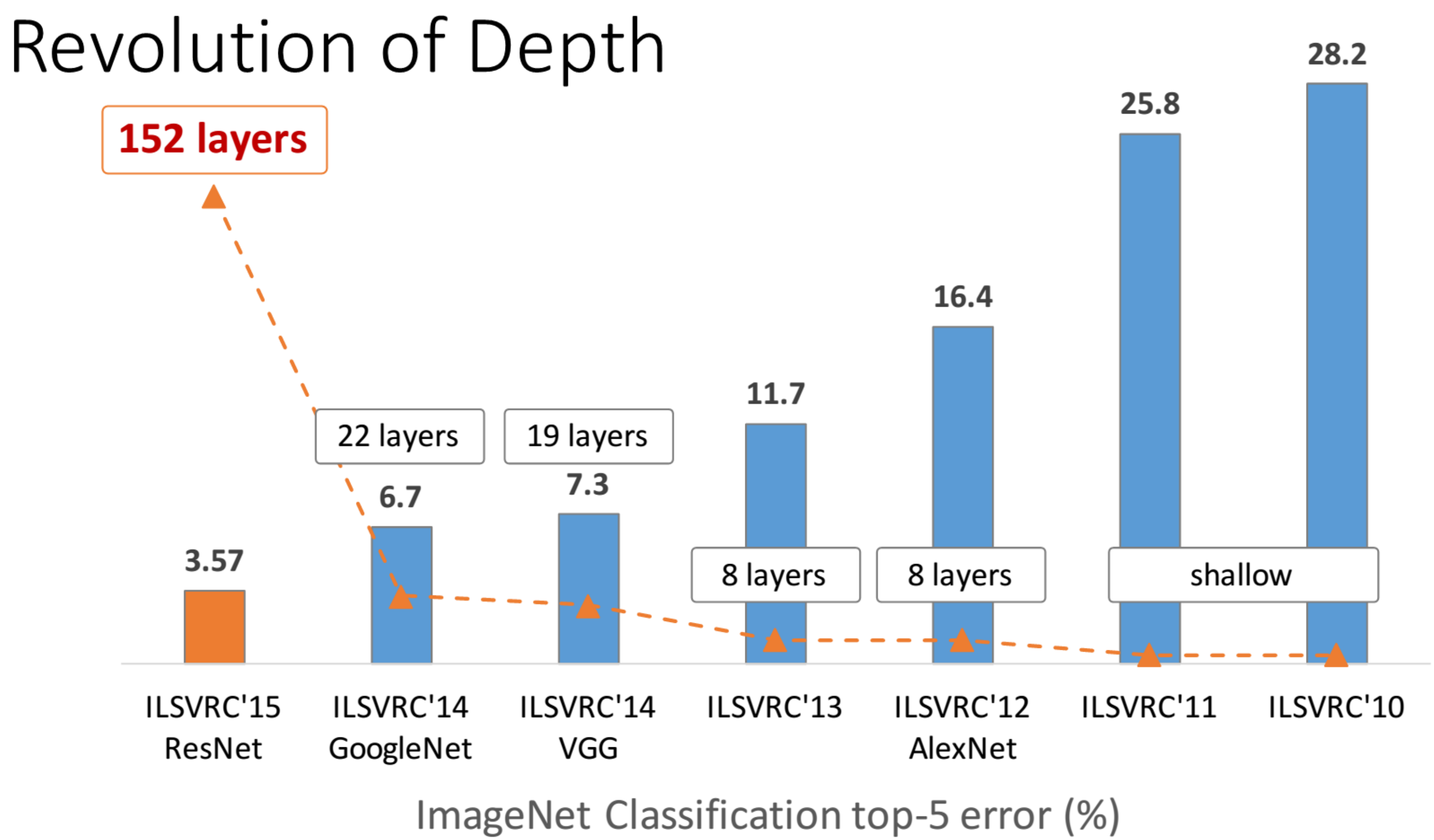
- **(Locality): Spatial Pyramid Matching & Pyramid Match Kernel**

- **(Aggregation): VLAD, Fisher Vector, Soft Assignment**

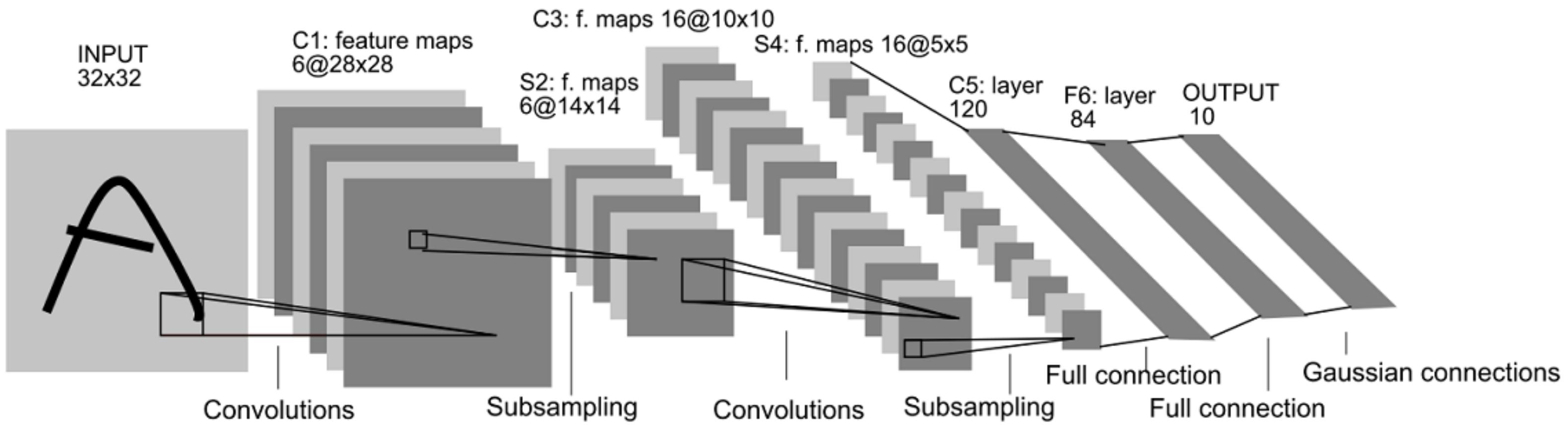
- **(Dictionary): Vocabulary Tree, Sparse Coding**

AlexNet

- Named after Alex Krizhevsky, proposed in 2012



LeNet-5

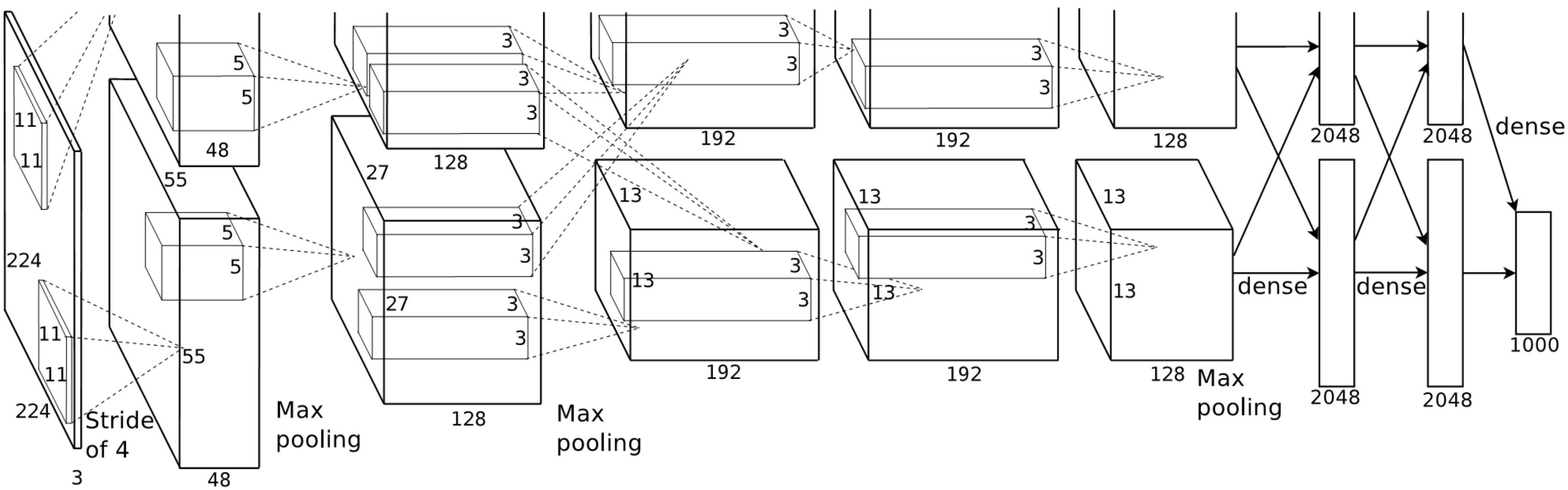


- Input: 32x32 pixel image. Largest character is 20x20
  (All important info should be in the center of the receptive field of the highest level feature detectors)
- Cx: Convolutional layer
- Sx: Subsample layer
- Fx: Fully connected layer
- Black and White pixel values are normalized:
  E.g. White = -0.1, Black =1.175 (Mean of pixels = 0, Std of pixels =1)

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, november 1998.

AlexNet

- Much larger than LeNet-5
- Trained on two GTX 580 GPU
- Largest networks at its time
- Utilize multiple engineering tricks (dropout, ReLU)



Alex Krizhevsky et al., ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

Why DL Suddenly works?

…It may be that the primary barriers to the success of neural networks were psychological (practitioners did not expect neural networks to work, so they did not make a serious effort to use neural networks)…

-- Goodfellow et al. “deep Learning”

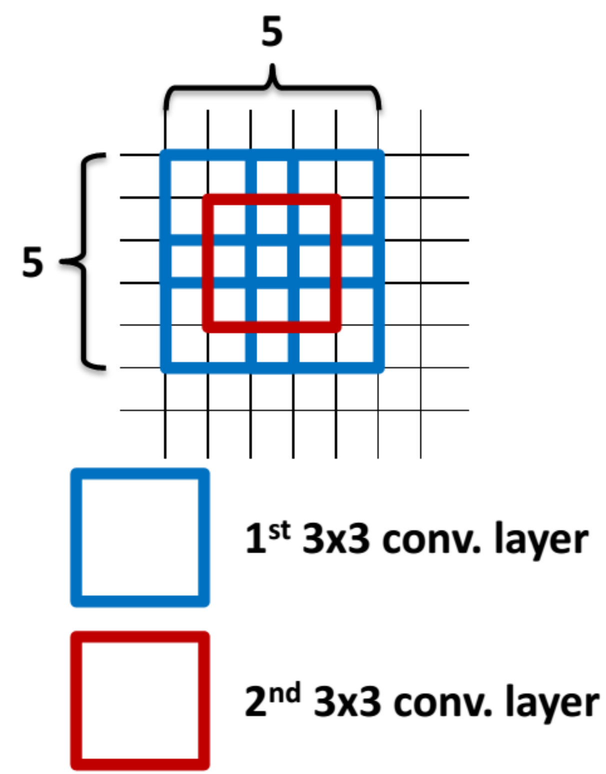
Why DL Suddenly works? – My Two Cents

- Emerging of big visual data
- GPU -> large network
- New engineering tricks (dropout, ReLU etc.)

VGG Net

Why 3x3 layers?

- Stacked conv. layers have a large receptive field
  - two 3x3 layers – 5x5 receptive field
  - three 3x3 layers – 7x7 receptive field
- More non-linearity
- Less parameters to learn
  - ~140M per net



Network Design

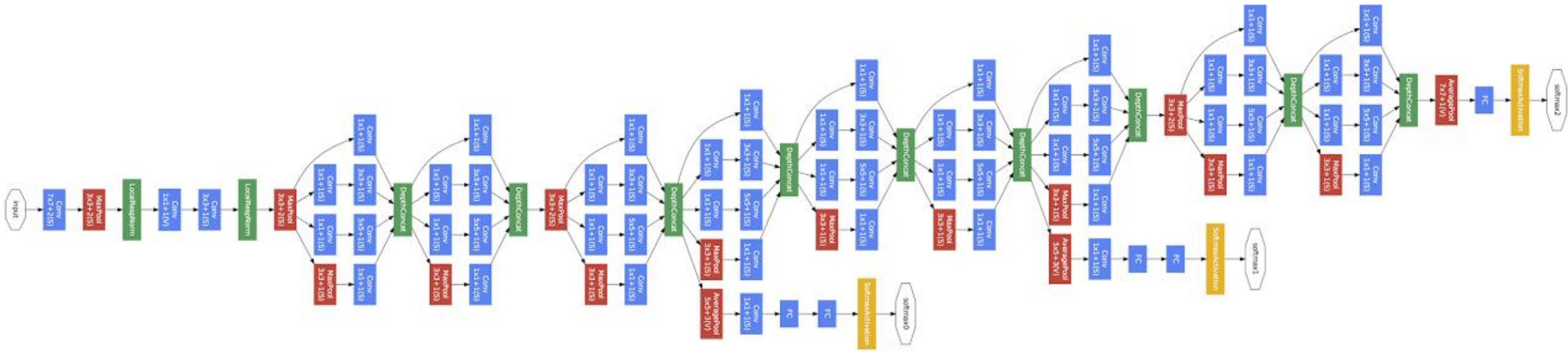
Key design choices:

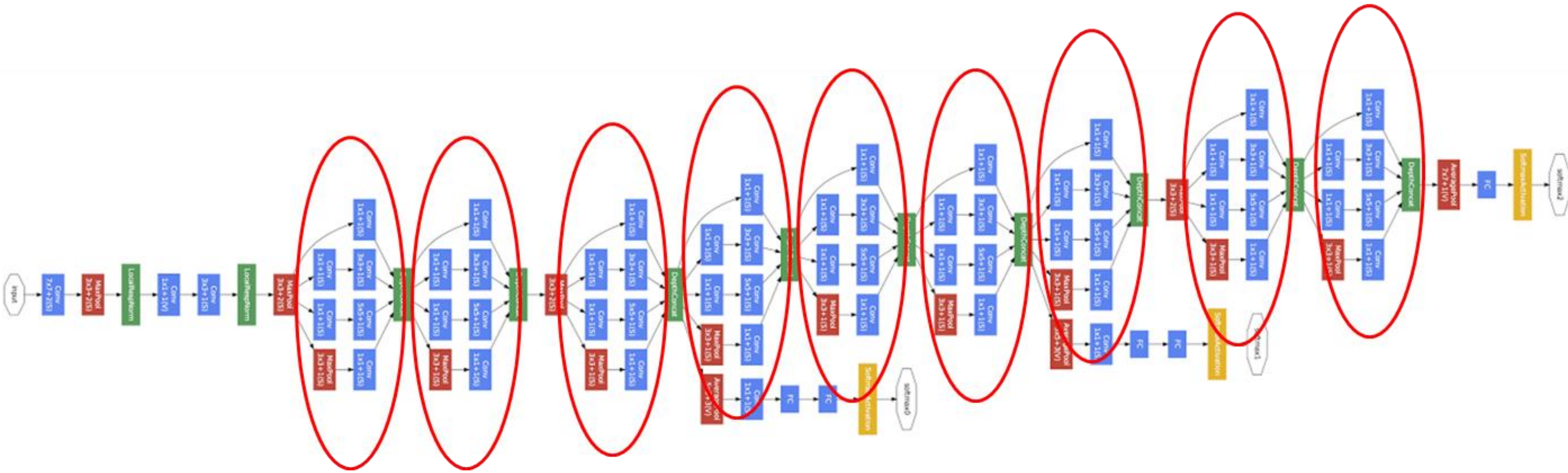
- 3x3 conv. kernels – very small
- conv. stride 1 – no loss of information

Other details:

- Rectification (ReLU) non-linearity
- 5 max-pool layers (x2 reduction)
- no normalisation
- 3 fully-connected (FC) layers

GoogLeNet



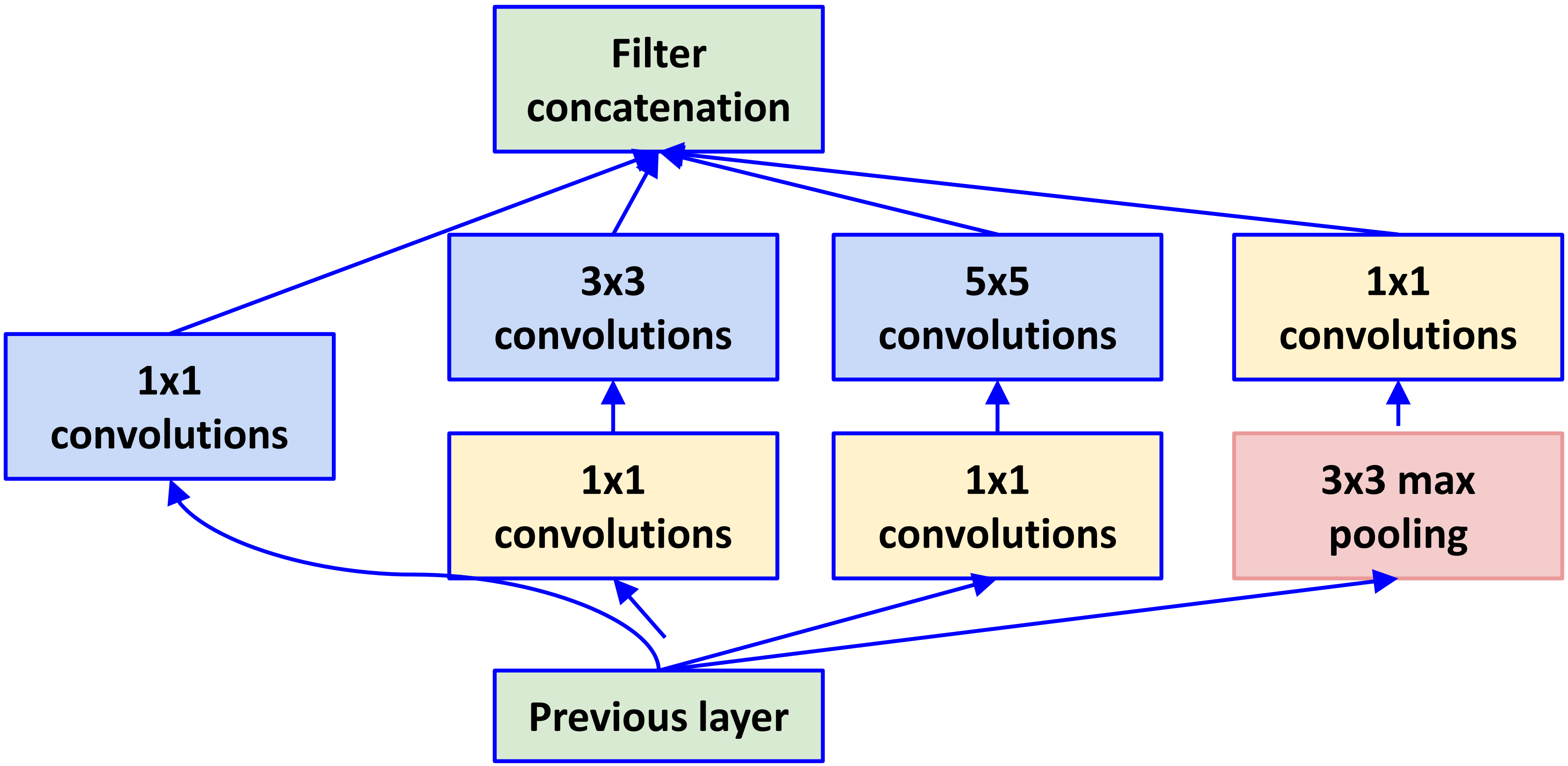


Inception

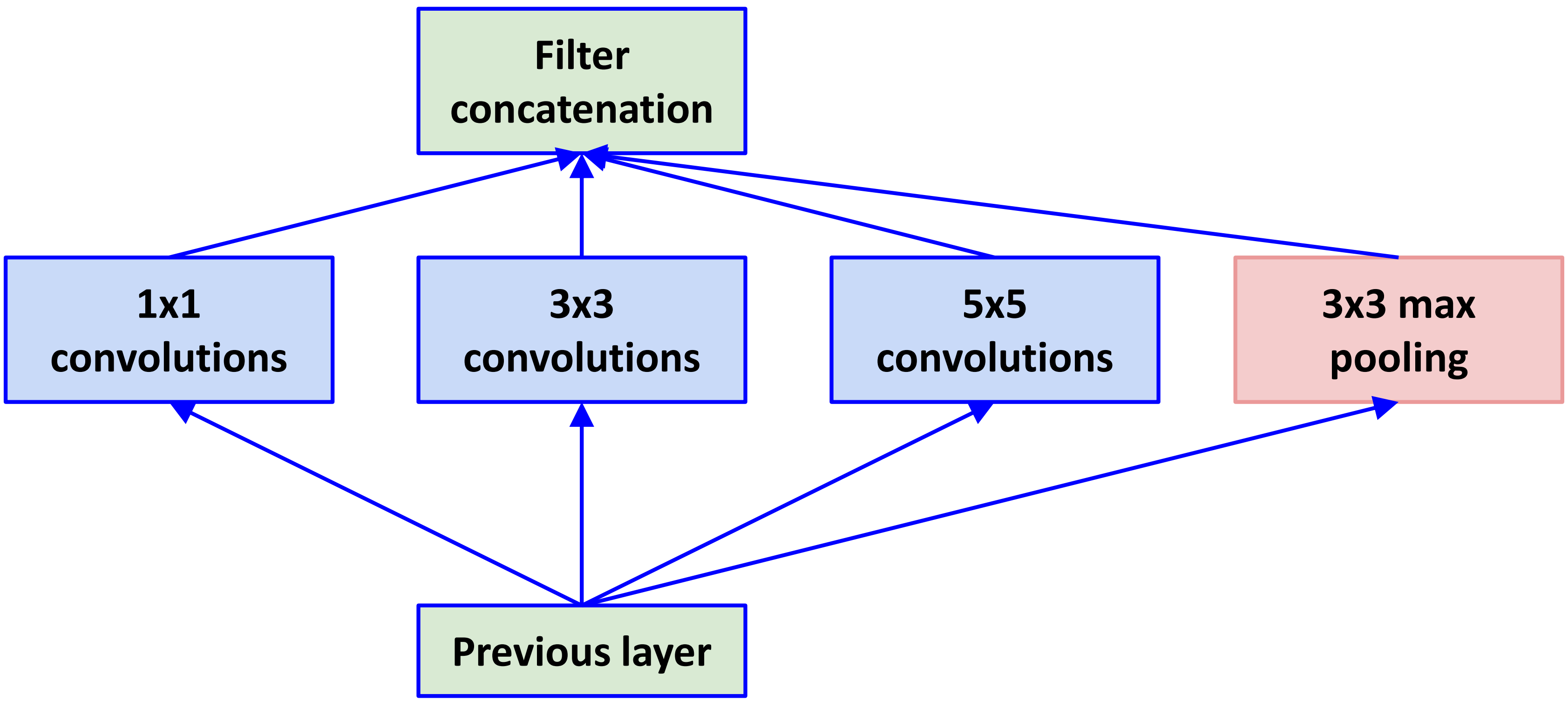
Network in a network in a network...

Convolution
Pooling
Softmax
Other

Inception module



Naive idea (does not work!)



ResNet

- See He Kaiming's ICML tutorial