Fusion Based Deep CNN for Improved Large-Scale Image Action Recognition

Yukhe Lavinia Holly H. Vo Abhishek Verma

CALIFORNIA STATE UNIVERSITY
FULLERTON

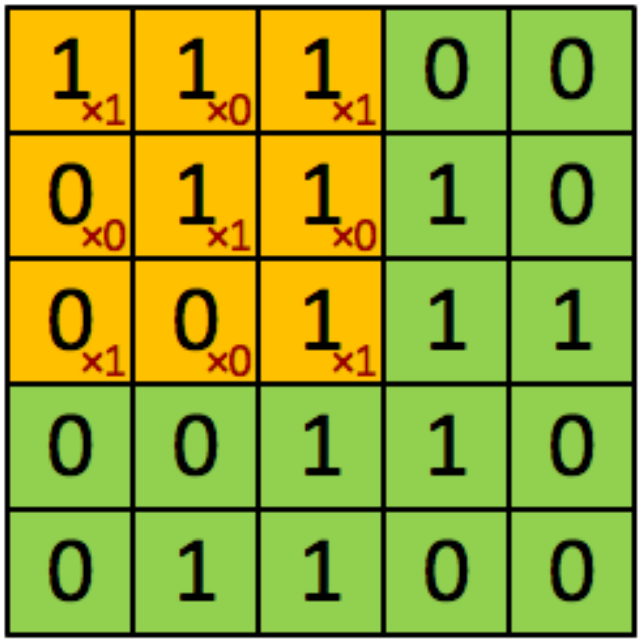
Agenda

Action Recognition
Convolutional Neural Networks (CNN)
GoogLeNet
VGGNet
Residual Net (ResNet)
Dataset
Pre-processing
Methodology
Experimental Setup
Experimental Results

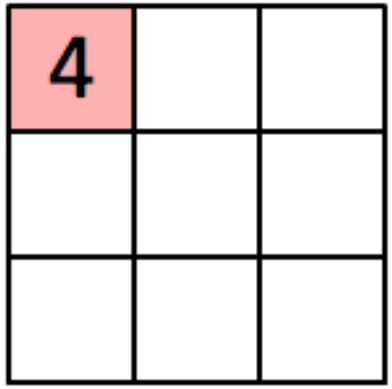
Action Recognition

- Process of labeling actions in video and still images
- Action representation: video (sequence of frames) and still images
- Benefits of still image action recognition: reduce amount of video frames, image retrieval
- Broad applications: security surveillance, child and elder-care monitoring, human-computer interaction

Convolutional Layer

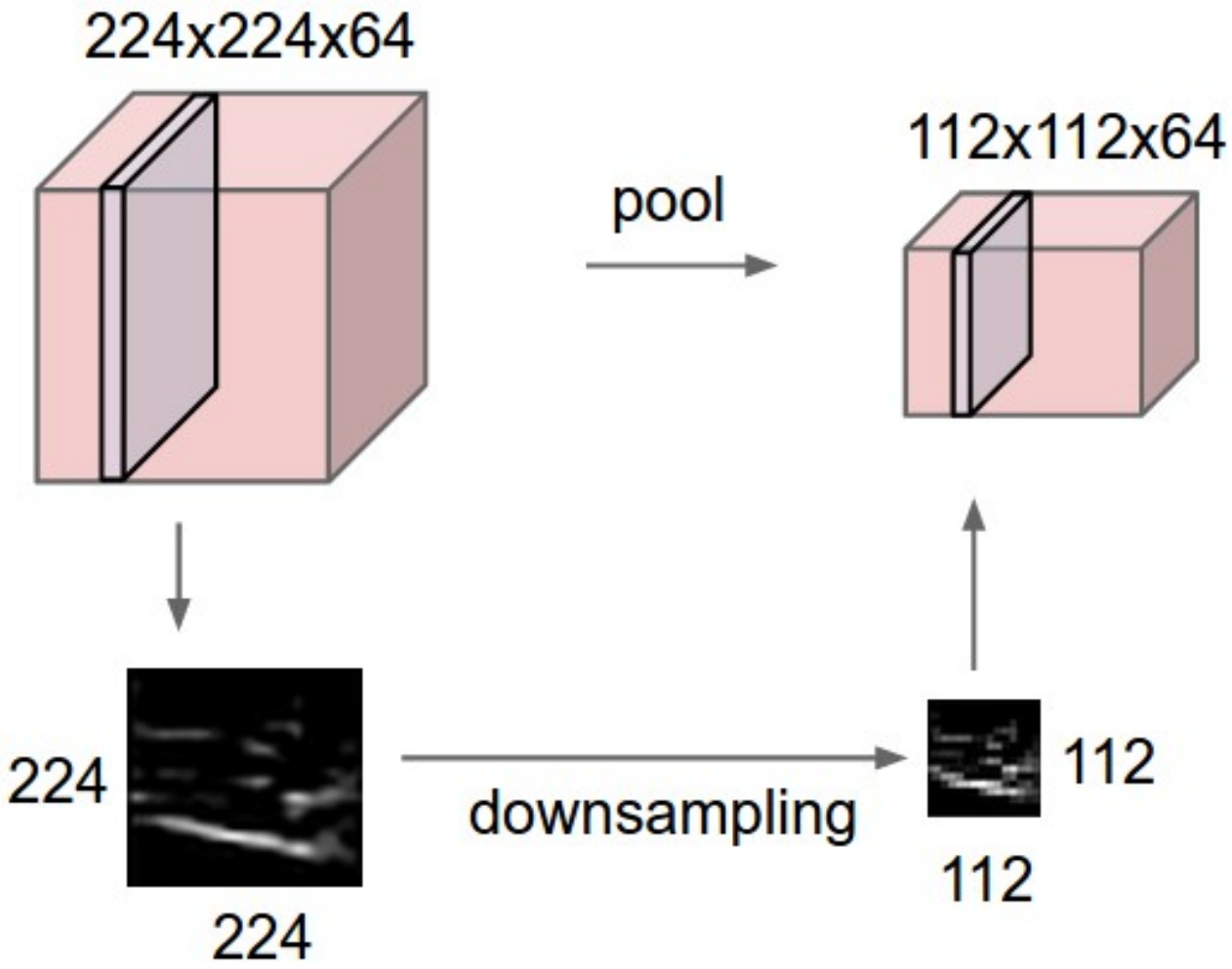


Image



Convolved Feature

Pooling Layer



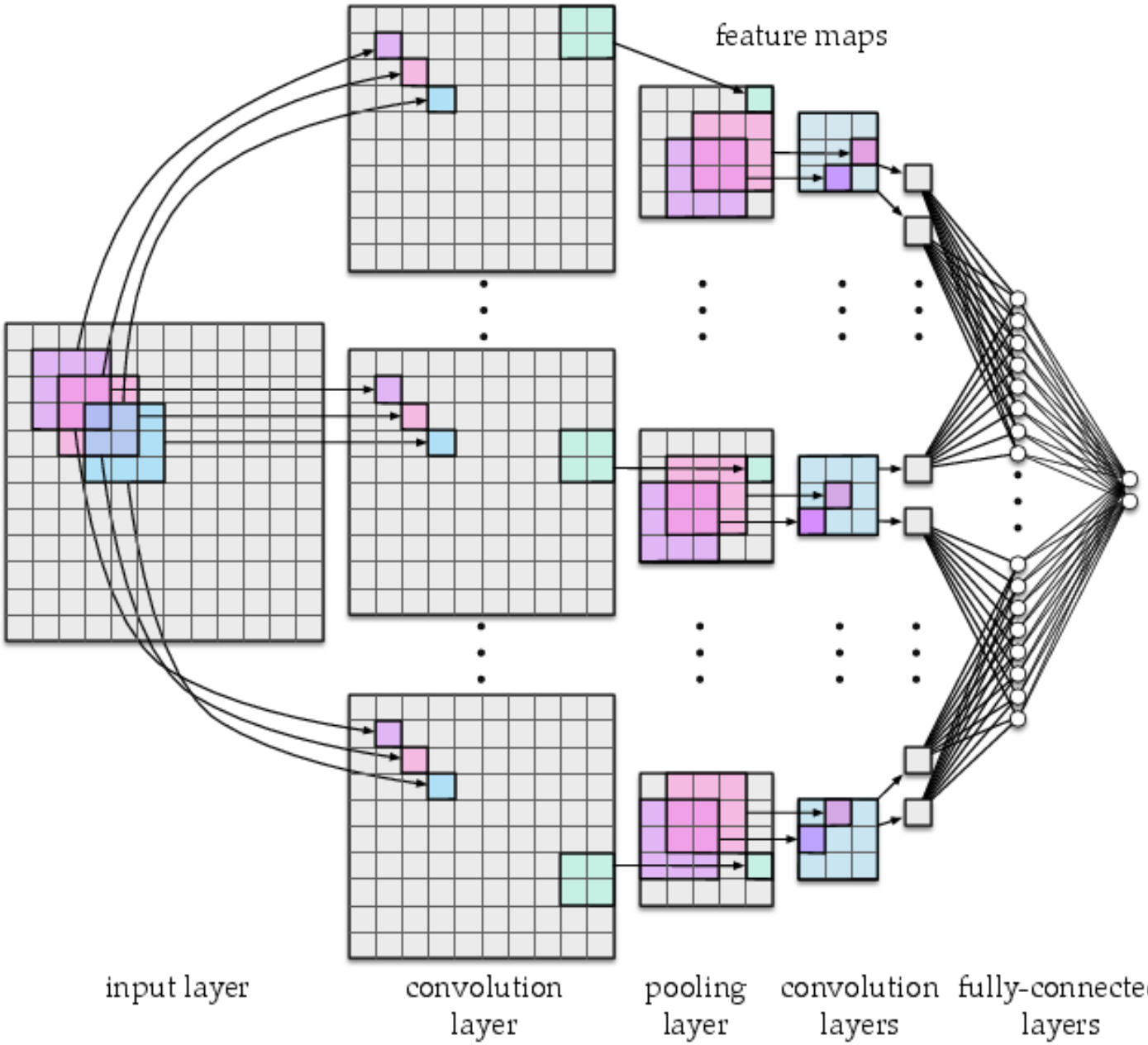
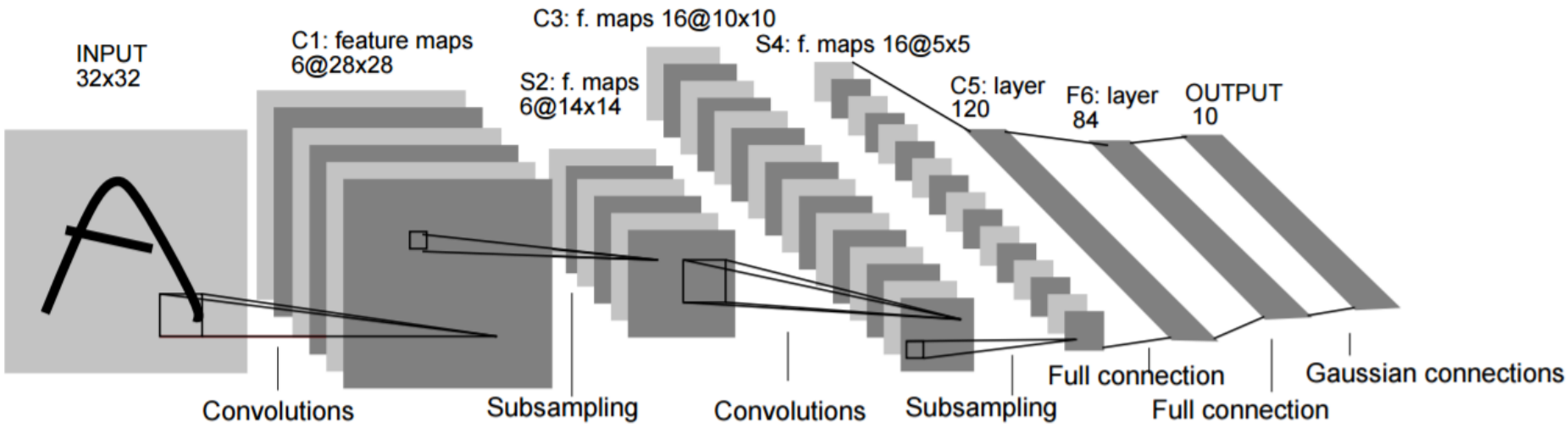


Image source: http://www.cs.put.poznan.pl/kkrawiec/wiki/uploads/Site/fig3.png

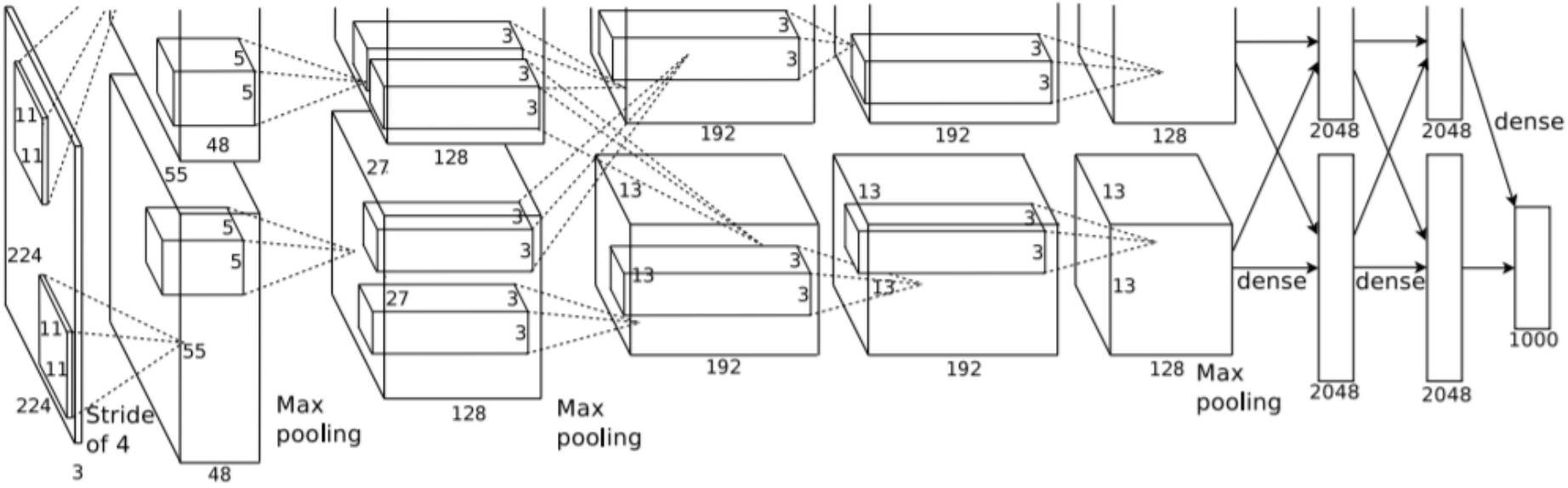
LeNet-5



Input 32x32

CONV1-POOL1-CONV2-POOL2-CONV3-FC

AlexNet



Input 224x224

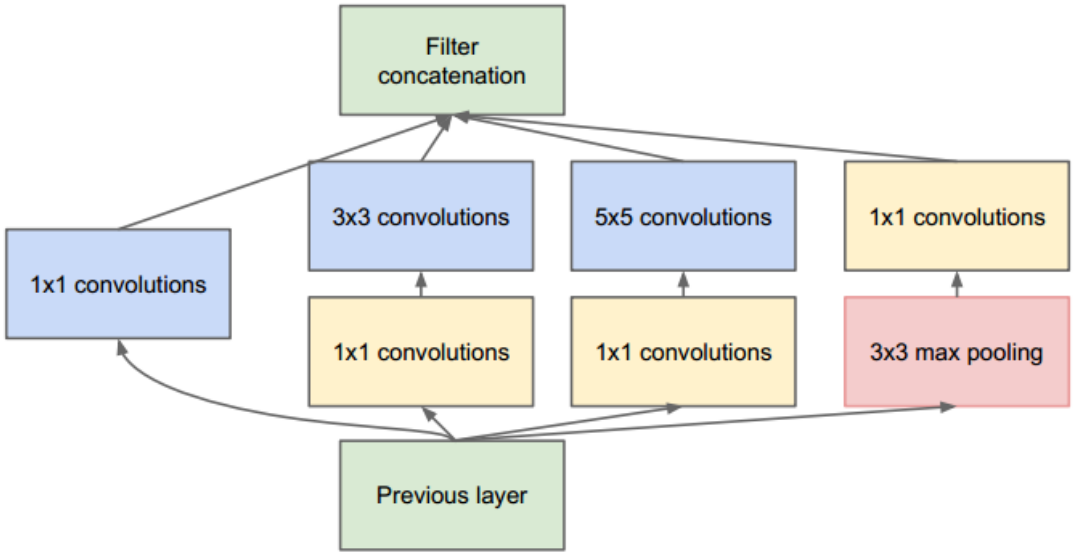
CONV1-POOL1-CONV2-POOL2-CONV3-CONV4-CONV5-FC

ILSVRC Winners

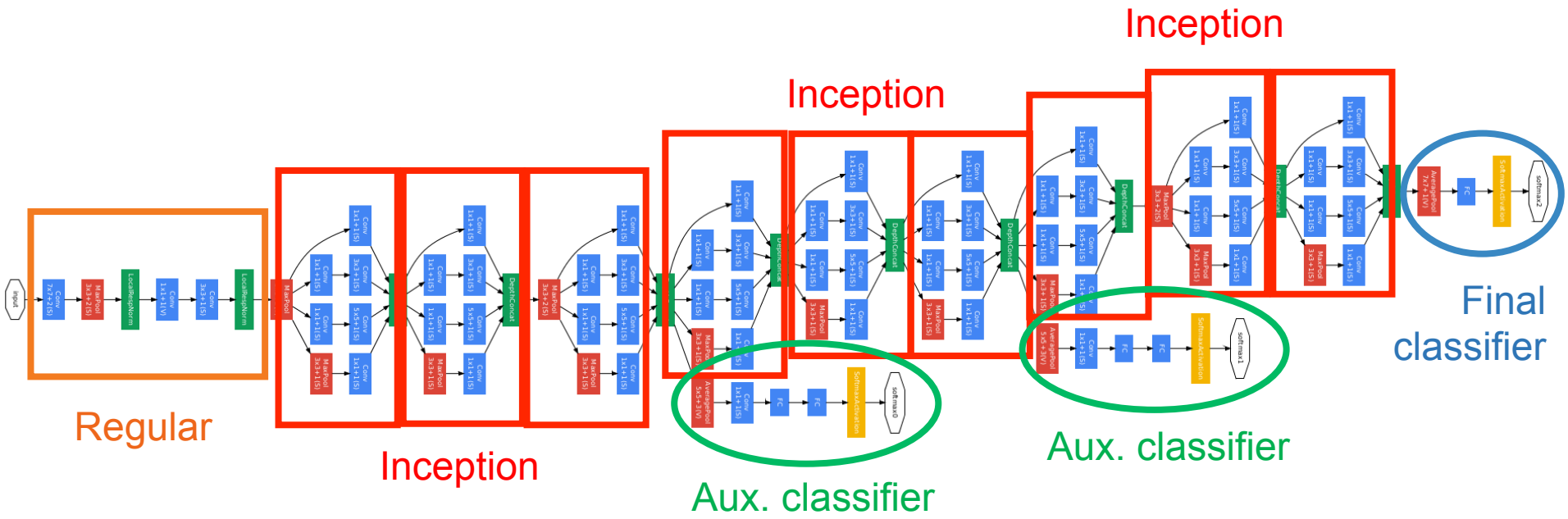
- ImageNet Large Scale Vision Recognition Challenge
- Advances in image classification and object detection
- Alexnet (2012): popularized deep CNN
- Winners: GoogLeNet, VGGNet, Residual Net

GoogLeNet

- Inception-v1
- Wide, parallel 1x1 conv, 3x3 conv, 5x5 conv, max pooling
- Reduced dimension through 1x1 conv
- Auxiliary classifiers
- Fast

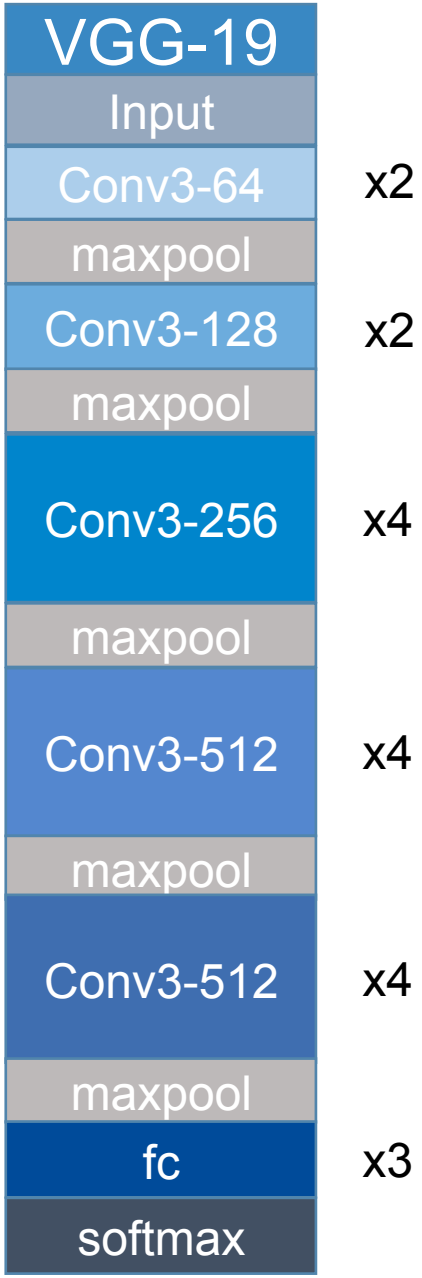


GoogLeNet


Inception
Inception
Inception
Regular
Inception
Aux. classifier
Aux. classifier
Final classifier

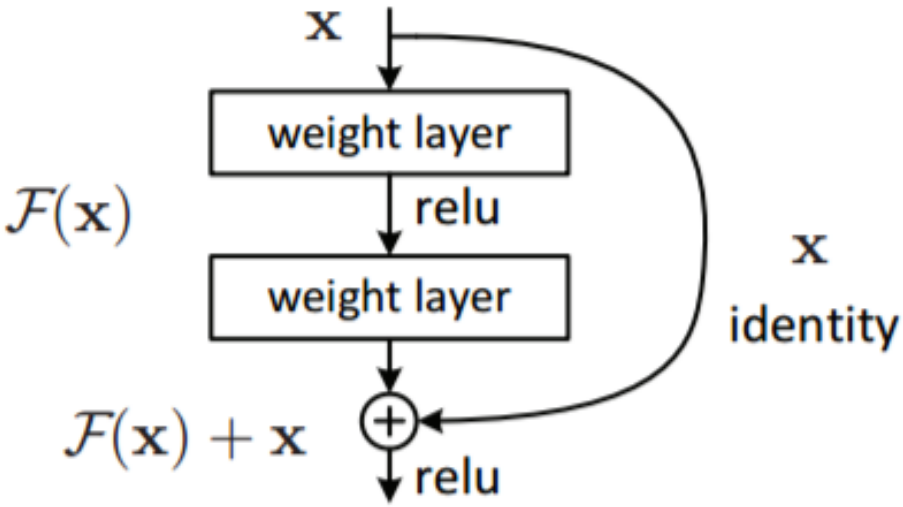
VGGNet

- Small filter
- Exhaustive sweep
- Homogenous architecture
- 3x3 conv, stride 1, maxpool 2x2
- 16 or 19 weight layers



ResNet

- Current state-of-the-art
- Residual learning
- Skip connections
- Batch normalization
- Deeper network



Dataset

Stanford 40 Action

40 classes
9532 total images
3200 training images
800 validation images
5532 test images

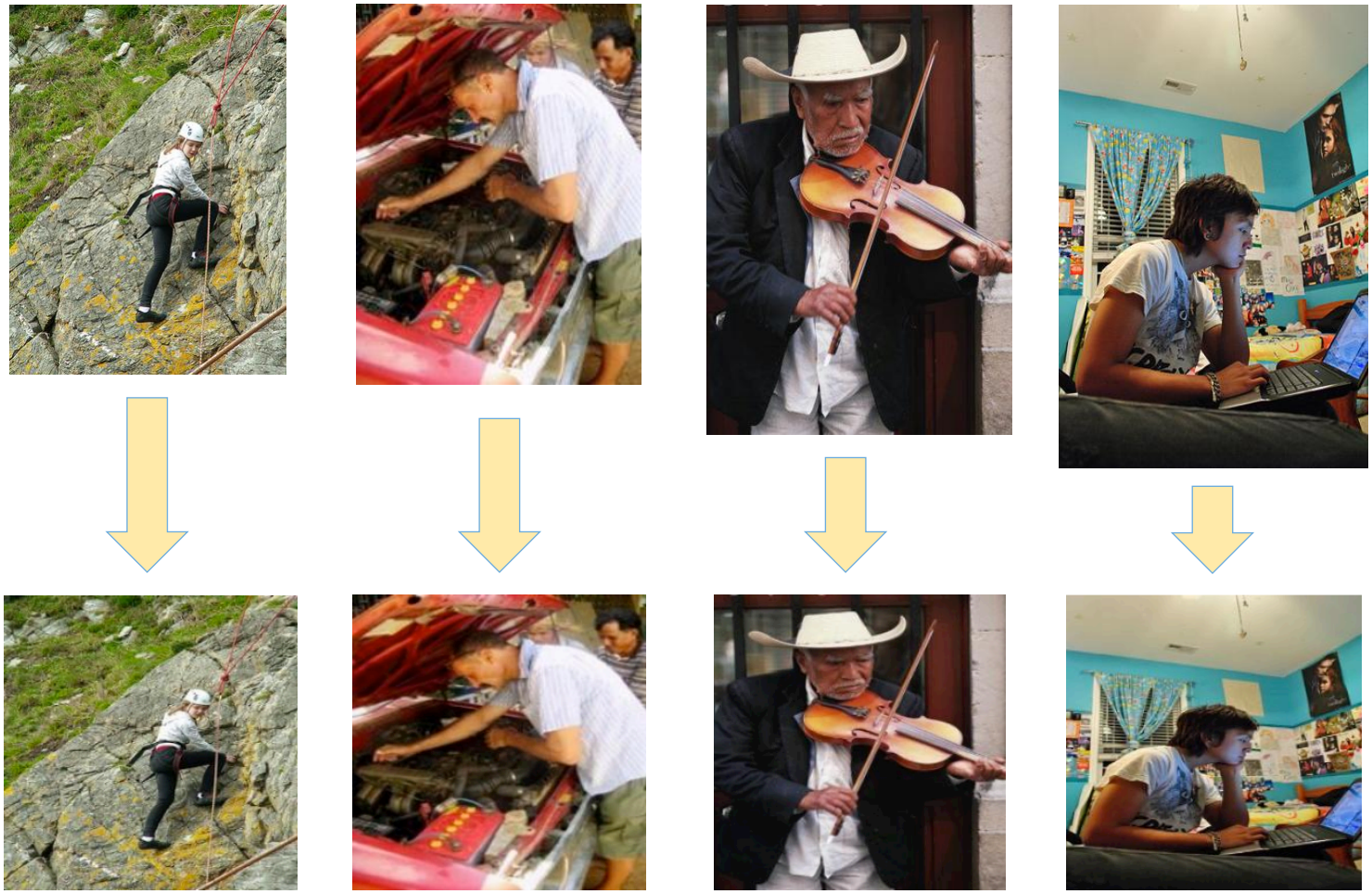
Dataset: Stanford 40 Action

- One of the hardest still image action recognition datasets
- Background clutter, various visibility, various poses



Source: B. Yao and L. Fei-Fei (2010)

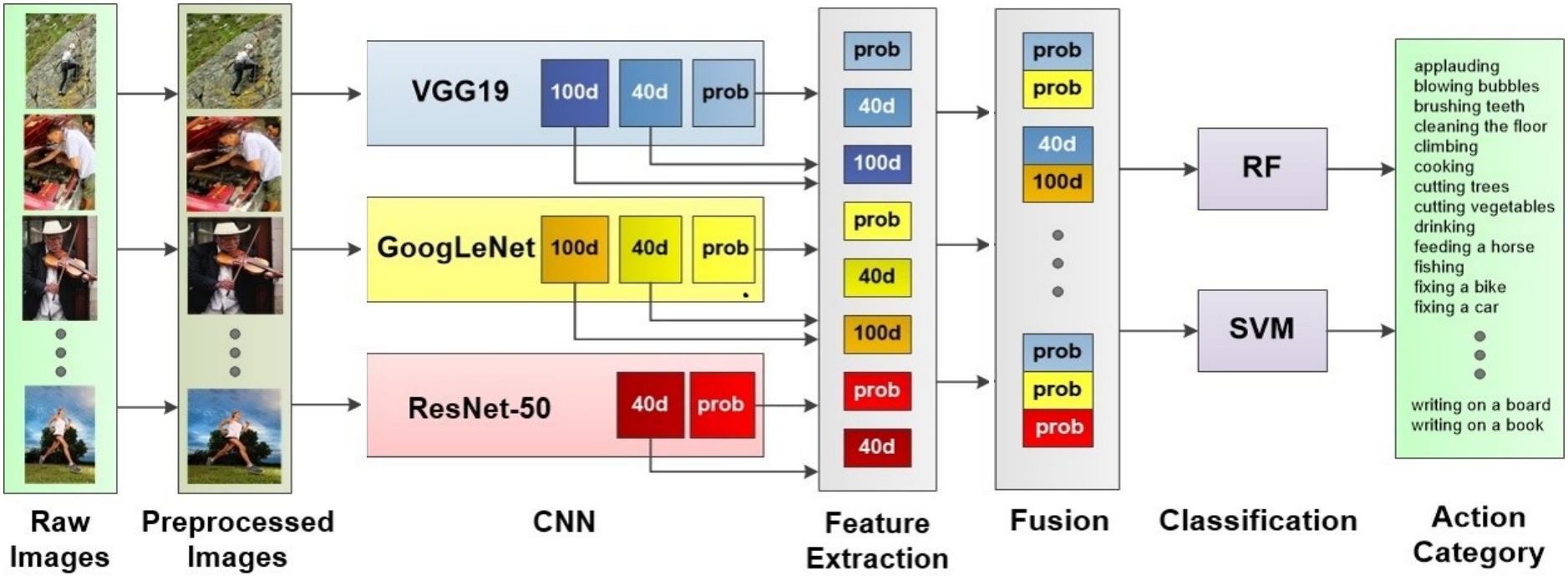
Pre-processing



Methodology

- Aim: collect more features using 3 networks
- Pre-trained weights
- Benchmark VGGNet (16 and 19), GoogLeNet, and ResNet on Stanford 40 Action
- VGG-19 performed slightly better than VGG-16

Methodology



Experimental Setup

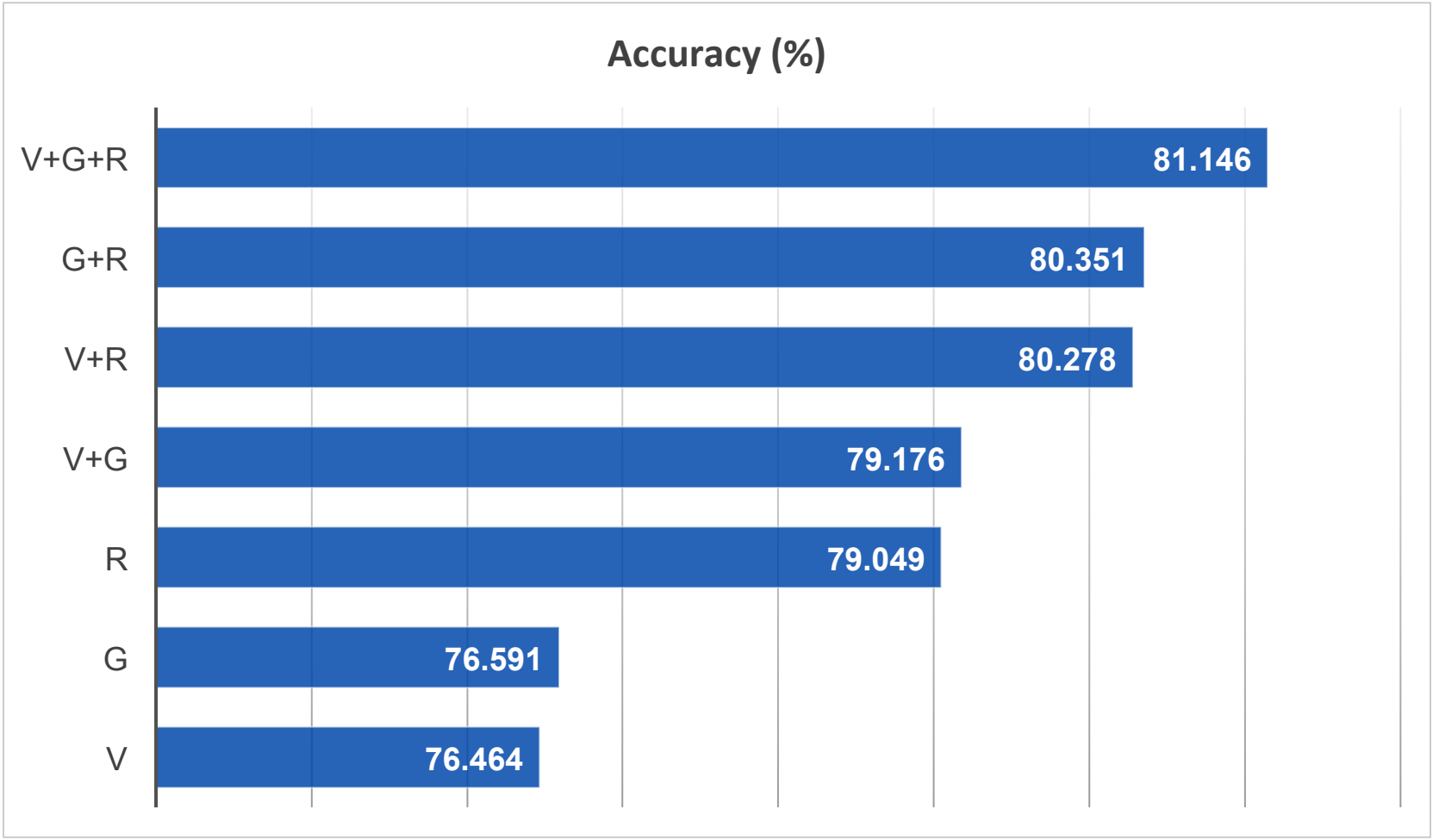
- Caffe
- NVIDIA DIGITS
- One NVIDIA GeForce GTX TITAN X GPU
- 12 GB VRAM
- Two Intel Xeon E5-2690 v3 2.60 GHz
- 48/24 logical/physical cores
- 256GB main memory

Source: Y. Jia et al. (2014), NVIDIA DIGITS Software (2015).

Experiments

	VGG-19	GoogLeNet	ResNet-50
Base LR	0.0002	0.0009	0.0009
Gamma	0.96	0.1	0.1
Batch Size	40	40	16
Epoch	50	30	30
Weight Decay	0.0005	0.0005	0.0005
Momentum	0.9	0.9	0.9
Train (mins)	147	9	58

Experimental Results



Summary

- Application of deep CNN on still image action recognition
- Fusion of two deep CNN models improved individual model accuracy
- Fusion of three deep CNN models further improved the two-nets fusion accuracy

Questions?

References

- Y. LeCun, et al., “Gradient-Based Learning Applied to Document Recognition,” *Proceedings of the IEEE*, 86(11):2278-2324, November 1998.
- A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," *Annu. Conf. on Neural Information Processing Systems (NIPS)*, Lake Tahoe, 2012.
- C. Szegedy et al., "Going deeper with convolutions," *Int. Conf. on Comp. Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015.
- K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Int. Conf. on Learning Reprsentation (ICLR)*, 2015.
- K. He et al., "Deep residual learning for image recognition," *Int. Conf. on Comp. Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016.
- B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," *Int. Conf. on Comp. Vision and Pattern Recognition (CVPR)*, San Franciso, CA, 2010.
- Y. Jia et al., “Caffe: Convolutional architecture for fast feature embedding, ” *arXiv preprint arXiv:1408.5093*, 2014.
- NVIDIA DIGITS Software. (2015). Retrieved April 23, 2016, from https: // developer.nvidia.com/digits.