Spatio Temporal Feature Extraction for Action Recognition in Videos

Lucas Mahler

EDAG Engineering GmbH

University of Applied Sciences Ulm

October 13, 2020

Overview

Motivation

EDAG CityBot:

Figure: CityBot showcase.[2]

Motivation

- EDAG CityBot requires perception of surrounding scene
- Scene contains complex actors $\rightarrow$ humans
- In order to interpret environment, actions of humans need to be recognized and detected

Goals

Problem Statement:

- How to accurately extract spatio-temporal features?
- Which architecture allows to do so efficiently?

Fundamentals

Fundamentals

Convolutional Neural Networks:

- Class of deep neural network
- Applied to grid-like data e.g. images
- Consists of layers of neurons
- Each neuron convolves over the input with its associated convolution kernel

Biological Inspiration

Hierarchical Organization of the Visual System

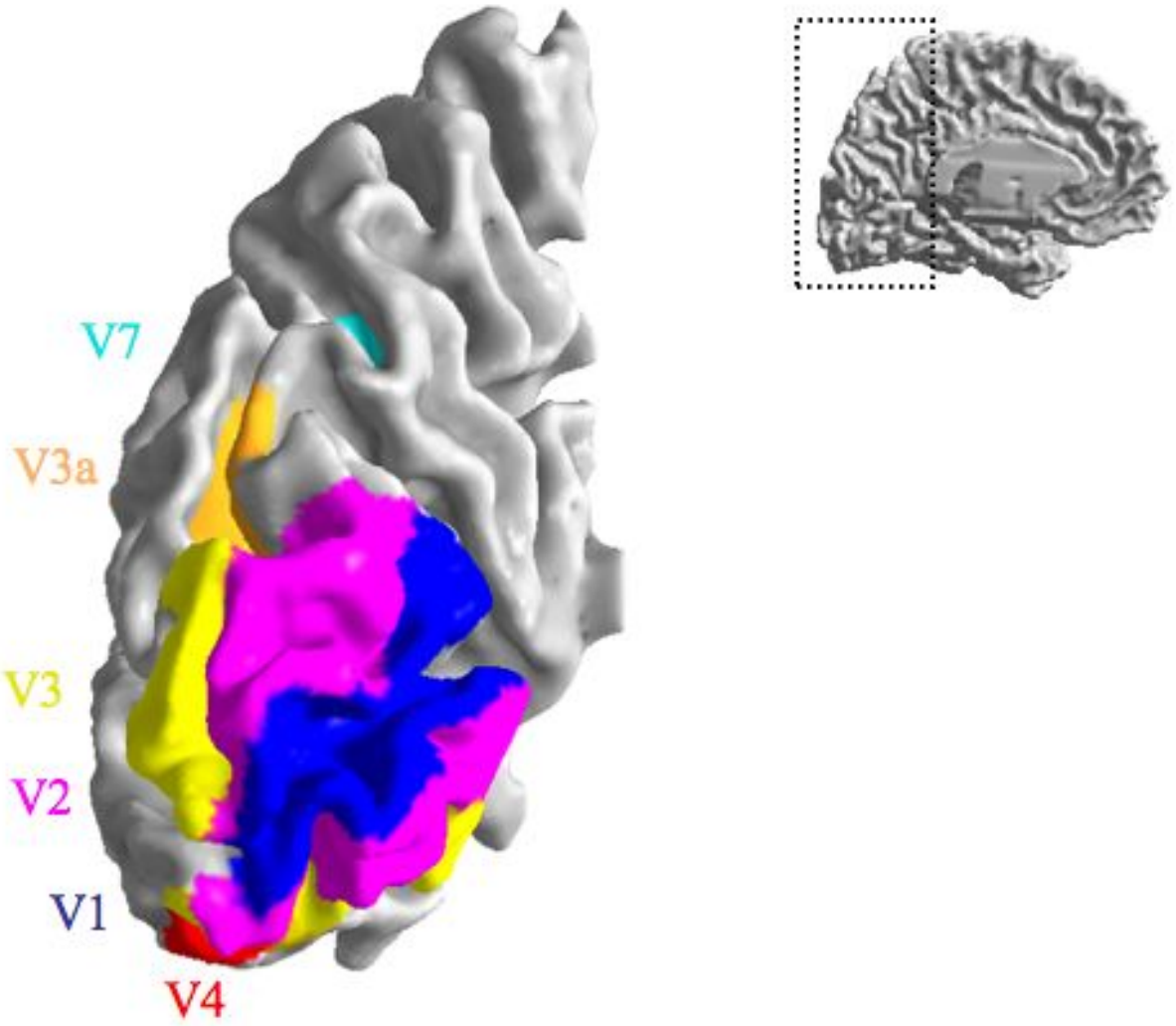


Figure: Areas of the visual cortex, from[3]

Functional Specialization

- Lateral geniculate nucleus (LGN) splits incoming information into two separate streams
- Specialized cells in LGN only respond to motion or structure/from stimuli
- Streams extend through the whole hierarchy of the visual system

Functional Specialization

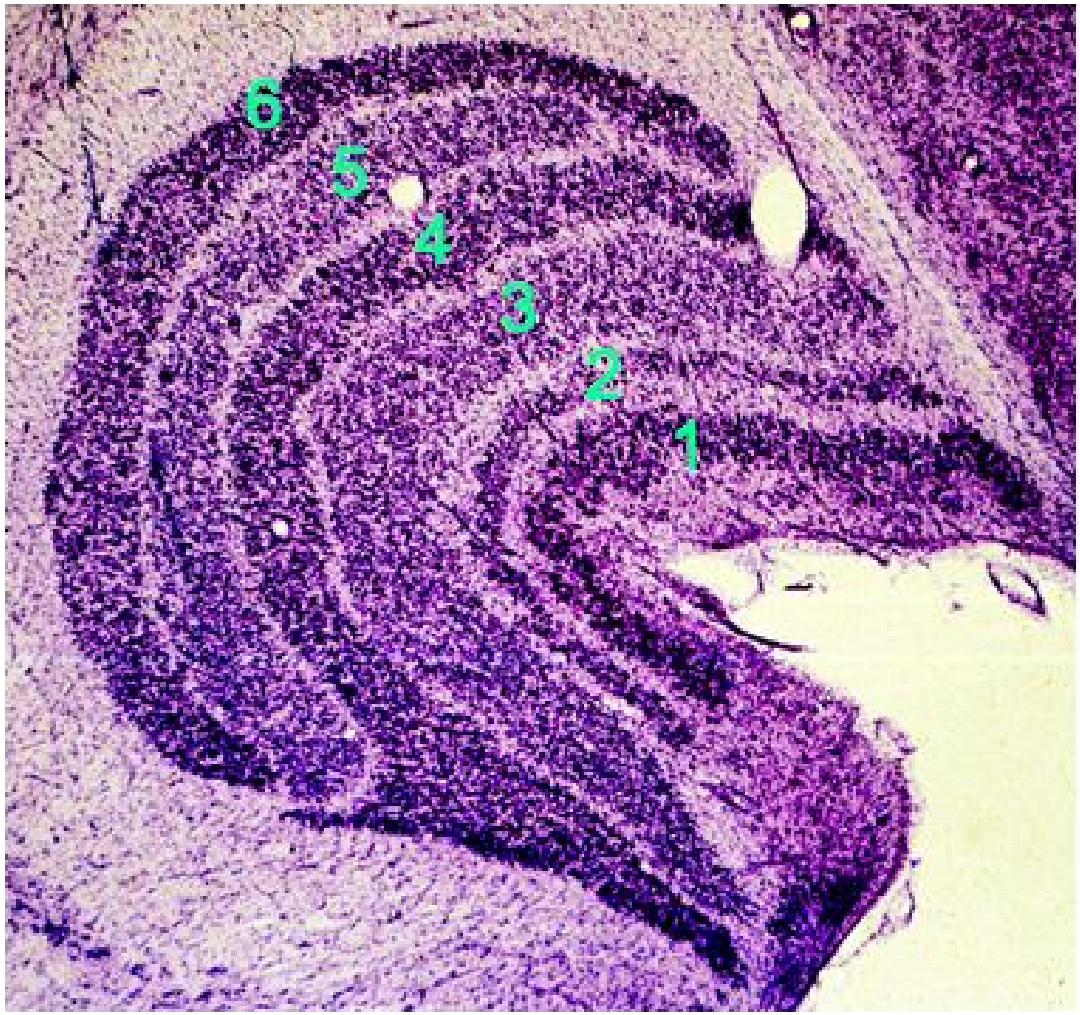


Figure: Laminar organization of the lateral geniculate nucleus from [3]. Layers 1-2 are magnocellular layers, layers 3-6 are parvocellular layers.

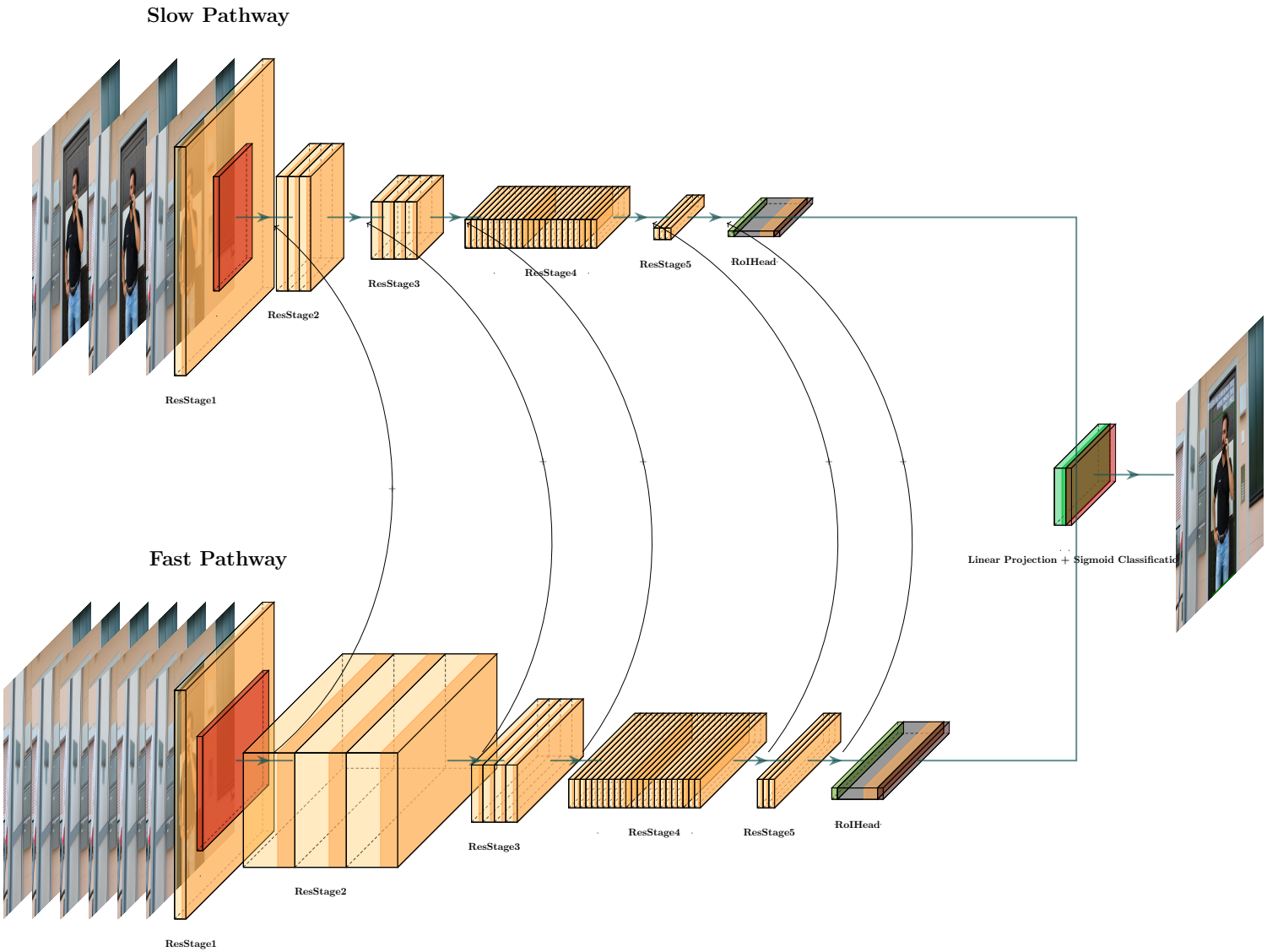
SlowFast

Design Considerations

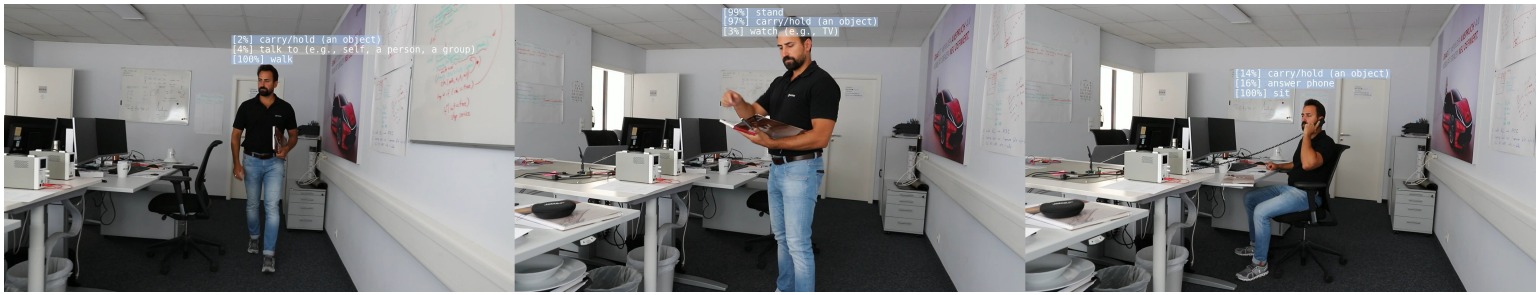
- Depth of the Architecture
- Hierarchical Structure
- Spatial and Temporal Streams
- Specialized Neurons

→ SlowFast Architecture proposed by Feichtenhofer et al.[1]

SlowFast Architecture



Action Recognition with SlowFast



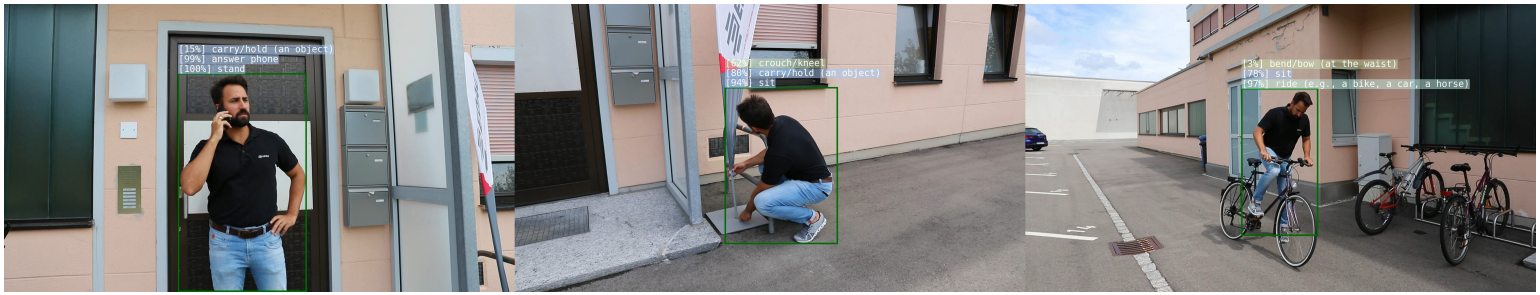
Action Detection with SlowFast

For human action detection, modifications to the recognition architecture are necessary.

- Off-the-shelf person detector computes bounding boxes for persons in an input video
- Bounding boxes of detected persons are fed to last stage of SlowFast
- SlowFast only extracts features that are relevant to current bounding box and forms predictions based on each bounding box

$\rightarrow$ RoI-Align block is added after final ResStage

Action Detection with SlowFast



Demo Video ...

Model Analysis

But is SlowFast better than previous methods?

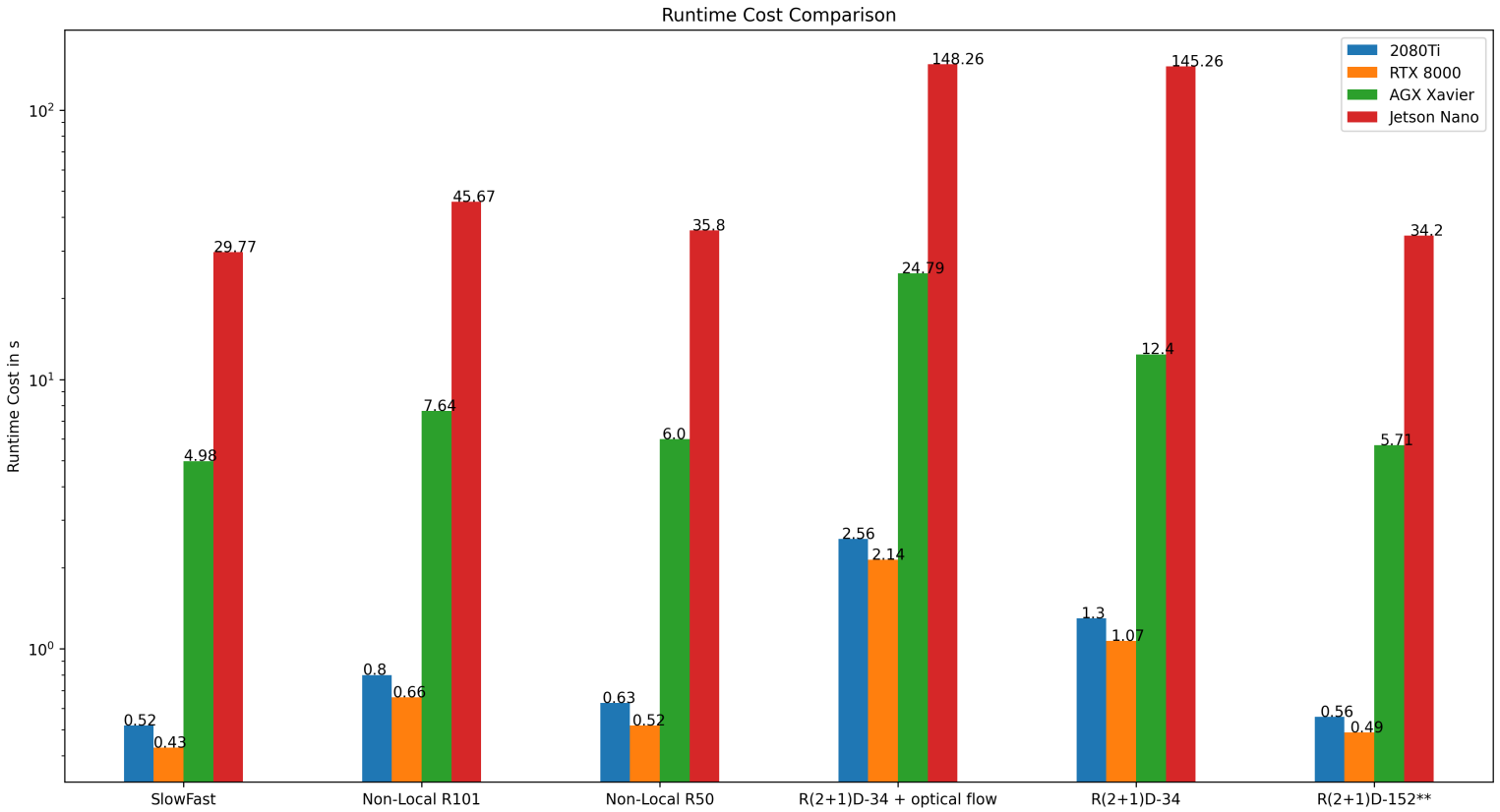
Practical Analysis

- Goal of new action recognition/detection methods: achieve state-of-the-art accuracy
- BUT: relevant factors for practical applications are often disregarded

Runtime Cost

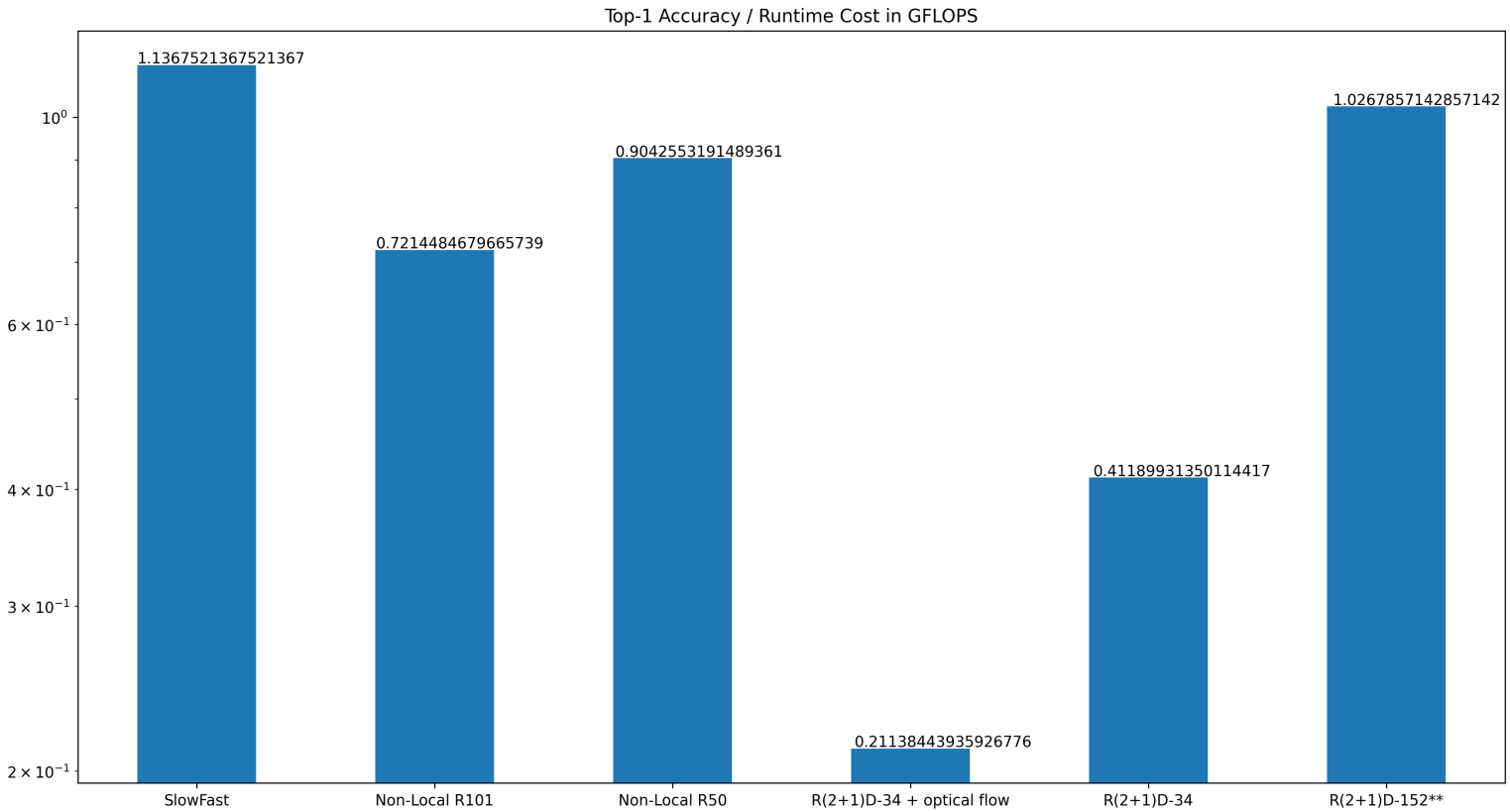
- Important measure to determine feasibility for deployment on time critical systems
- Evaluation of runtime by regarding inference cost per spacetime view
- Spacetime view: temporal clip with spatial crop
- For SlowFast: spatial size of 256px×256px for 10 temporal clips with 3 spatial crops each $\rightarrow$ 30 views
- Cost per view in GFLOPS for SlowFast: 234 GFLOPS $\rightarrow$ in total $234\text{GFLOPS} \times 30\text{views} = 7020\text{GFLOPS}$

Runtime Cost



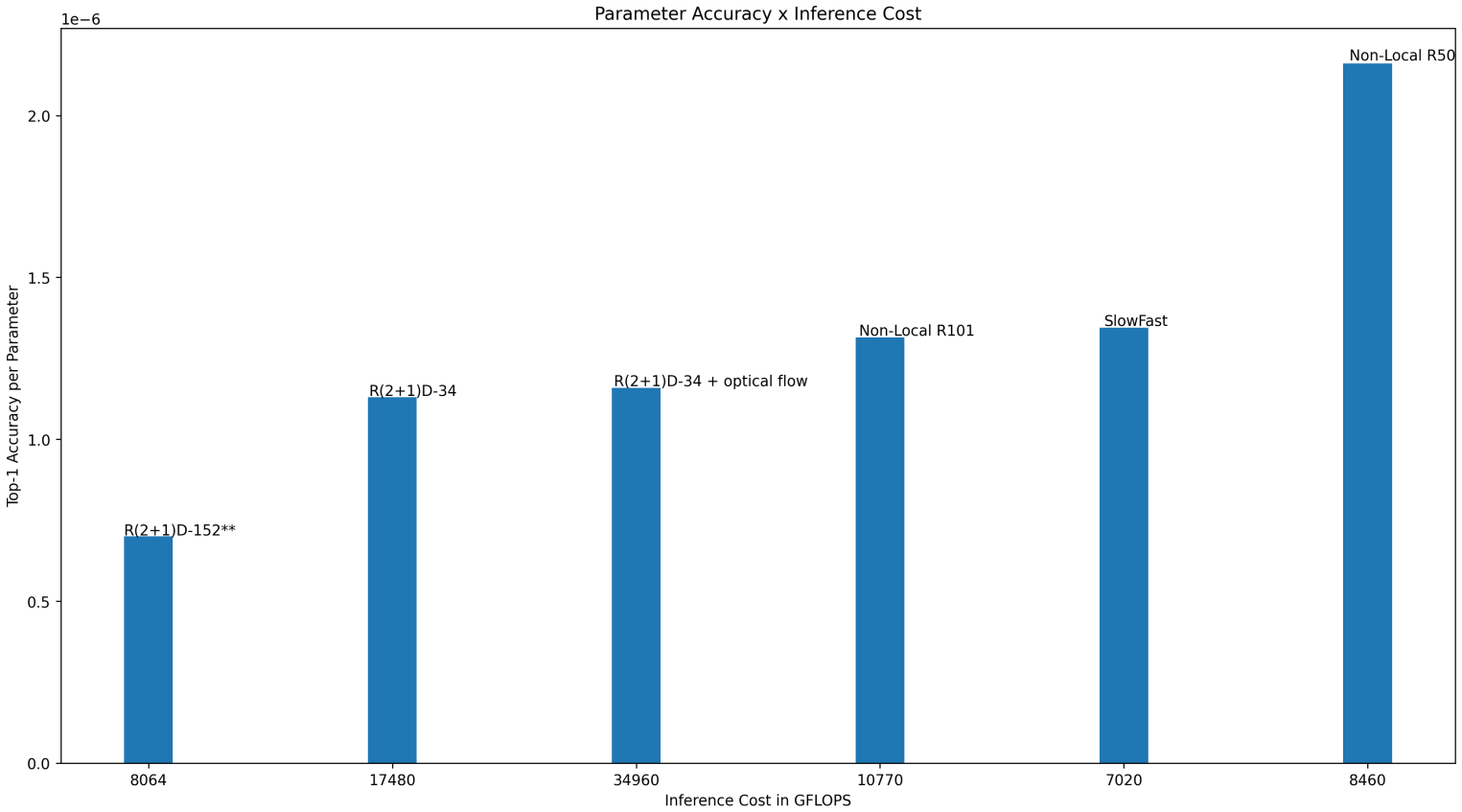
Accuracy and Throughput

- Accuracy per floating point operation is key metric for evaluation of practicability of given models



Parameter Utilization

- Accuracy per parameter is used to measure efficiency ofthe networks representative capabilities



Results of the Practical Analysis

- SlowFast has lowest runtime cost with state-of-the-art accuracy
- Memory cost is highly implementation specific → could be reduced for critical applications
- SlowFast has highest accuracy and throughput
- SlowFast has high representative capabilities even with large number of parameters

→ Results show potential of two-streamed approaches and illustrate superior motion-modeling capabilities of SlowFast

Methodical Analysis

What are the processes that extract features?

Deep Hierarchical Processing

- Visual processing system has deep hierarchical structure with more than 10 processing stages
- Visual processing takes up more than 50% of cortical area
- Previous video recognition architectures exclusively used models with ≥ 50 layers
- SlowFast uses two 101 layer ResNets as backbone, thus is very deep architecture
- ResStages in SlowFast are similar to processing stages in mammalian visual processing
- Division into spatial and temporal streams introduces additional hierarchical stages

2-Streamed Analysis

- Early activations of both pathways very similar
- Low level feature extraction happens in first ResStages $\rightarrow$ demonstrated by high spatial resolution of activations
- Two pathways only differ in temporal resolution
- Slow pathway has high channel capacity but low temporal resolution, thus focus on structure/form clues forces
- Fast pathway has low channel capacity but high temporal resolution, thus focus on temporal feature aggregation
- With increasing depth, difference of the two pathways activations increases $\rightarrow$ specialization clearly visible

$\rightarrow$ Findings validate design decision of two-streamed approach

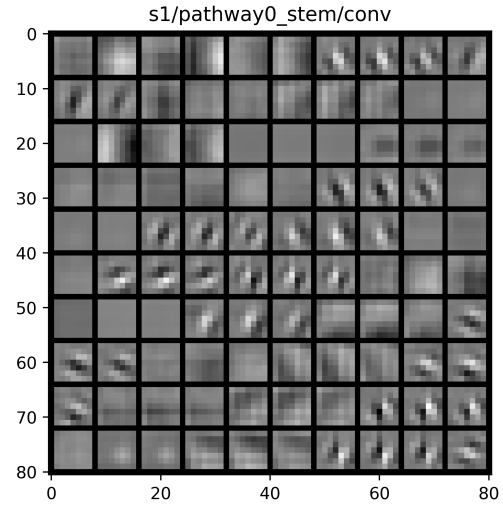
2-Streamed Analysis

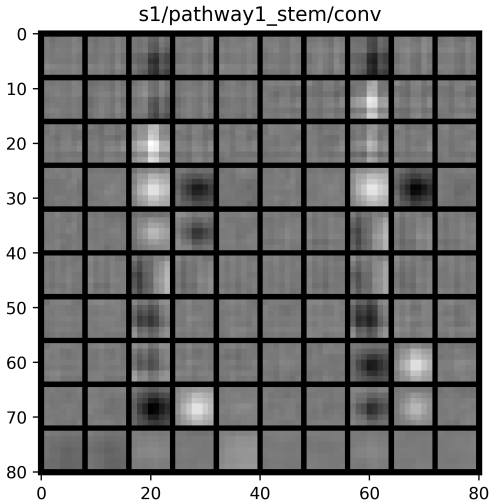
Demo

High Functional Specialization

- Filters in both pathways are highly specialized due to great variety of clues in training set
- Filter complexity rises with network depth
- Early filters are relatively simple, in late stages filters only respond very selectively
- High difference in filters in slow and fast pathways throughout all stages

High Functional Specialization





Results from Methodical Analysis

- Splitting of motion and form showed significant improvements
- Varying channel capacity and temporal resolution lead to strong functional specialization
- Functional specialization further enforced by increasingly more complex filters
- Modeling capabilities of SlowFast stem from architectural decisions
- Inspiration from mammalian visual processing deemed highly valuable
- Similarities between biological and artificial neural networks are product of learning not programming

Concluding Remarks

Concluding Remarks

- SlowFast has highest from-scratch accuracy
- Practical analysis showed strength of SlowFast in breadth of different practical metrics
- Methodical analysis showed, two-streamed approach is very promising design
- Powerful spatio-temporal modeling capabilities are reason for good performance on tasks of action recognition and detection
- Showed that SlowFast is capable for real world applications

The End

Thank you for listening!
I'll be happy to answer your questions!

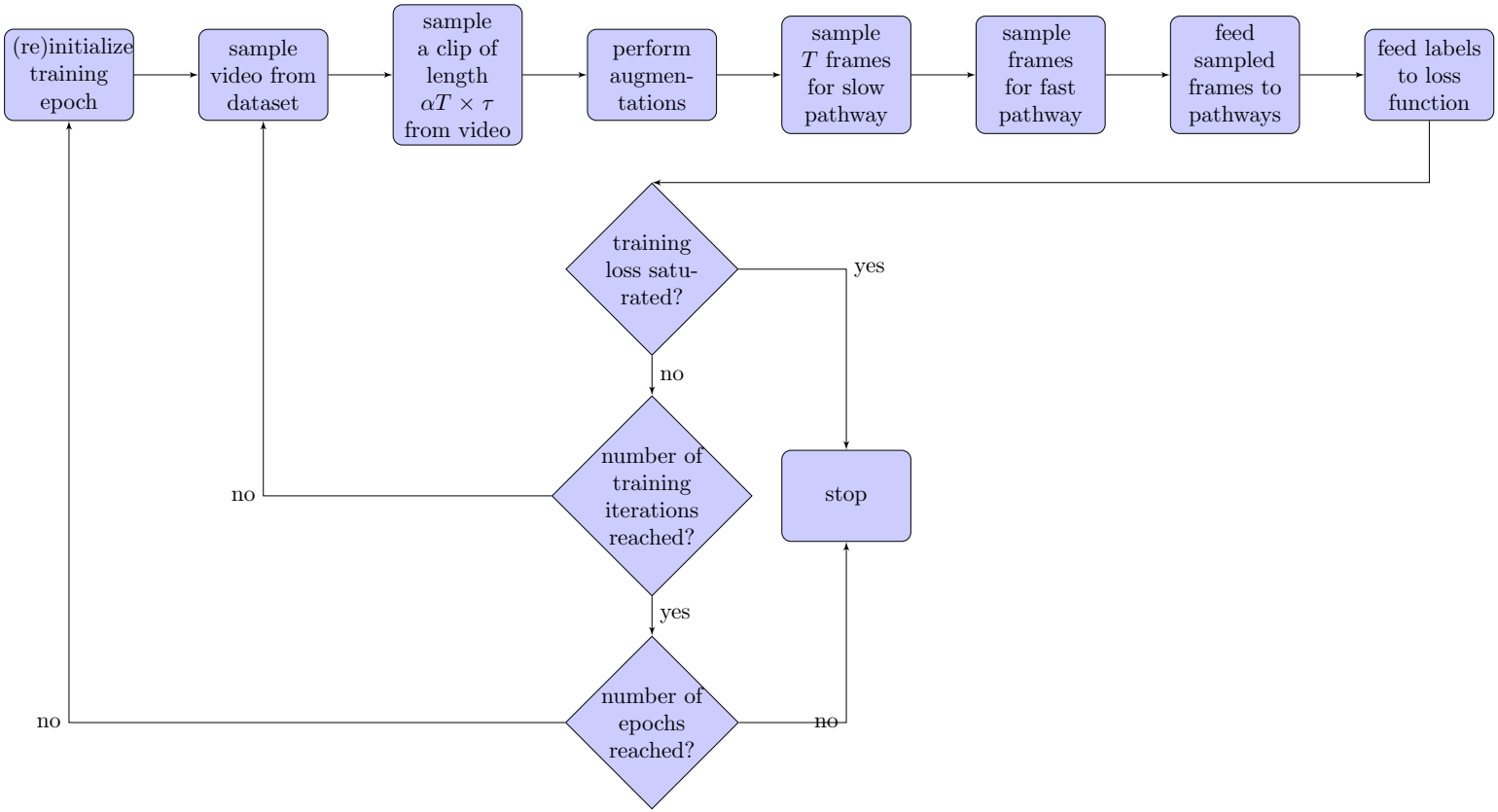
Bibliography I


Christoph Feichtenhofer et al. *SlowFast Networks for Video Recognition*. 2018. arXiv: 1812.03982 [cs.CV].

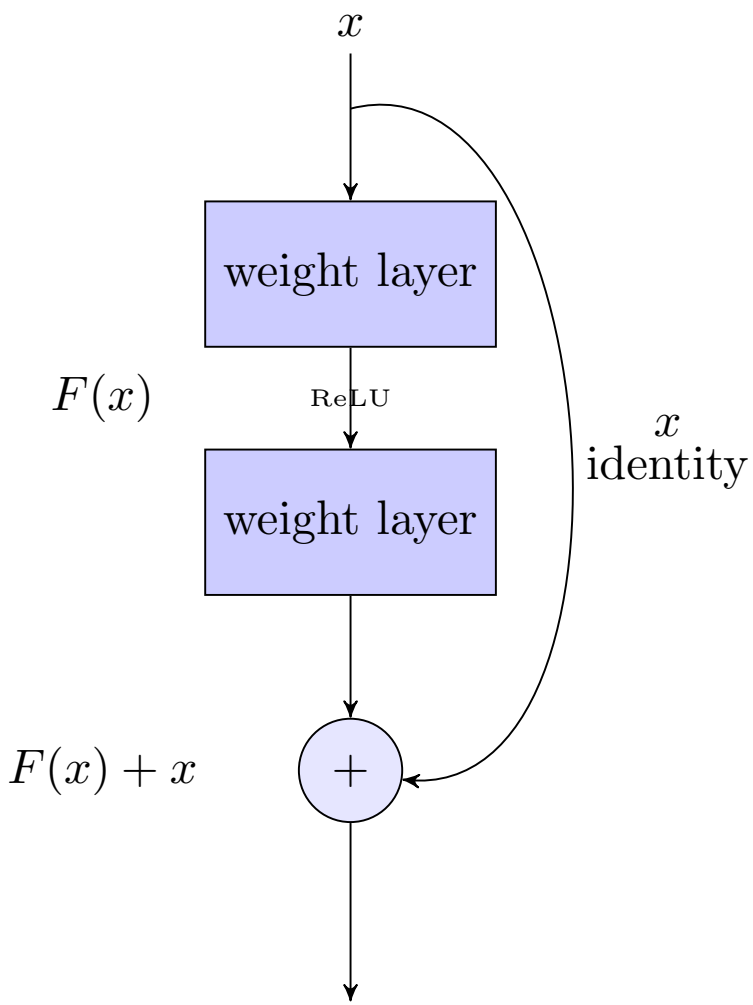

EDAG Engineering GmbH. *EDAG CityBot – An autonomous transport and working vehicle for the smart city of tomorrow*. 2019. URL: https://www.edag-citybot.de/file/edag-whitepaper-01-2019_en.pdf.


Prof. David Heeger. *Perception Lecture Notes: LGN and V1*. Accessed July 15, 2020. URL: https://www.cns.nyu.edu/~david/courses/perception/lecturenotes/V1/lgn-V1.html.

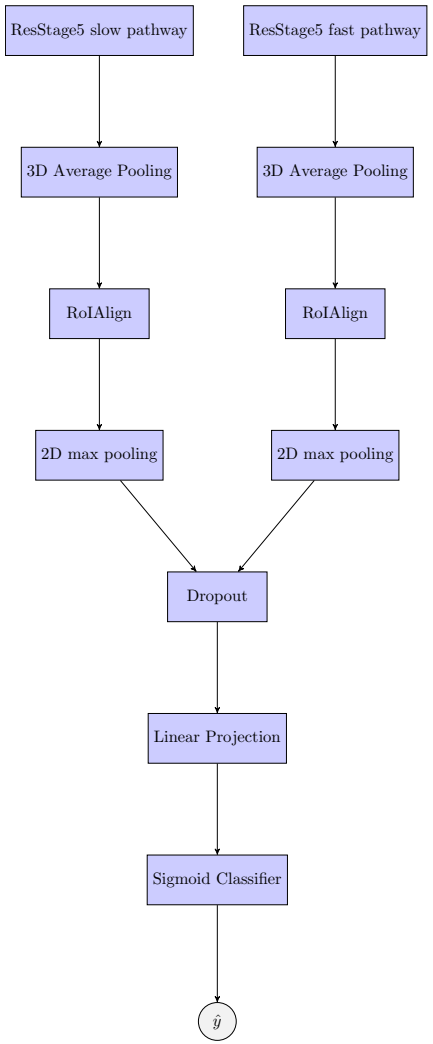
Training Procedure



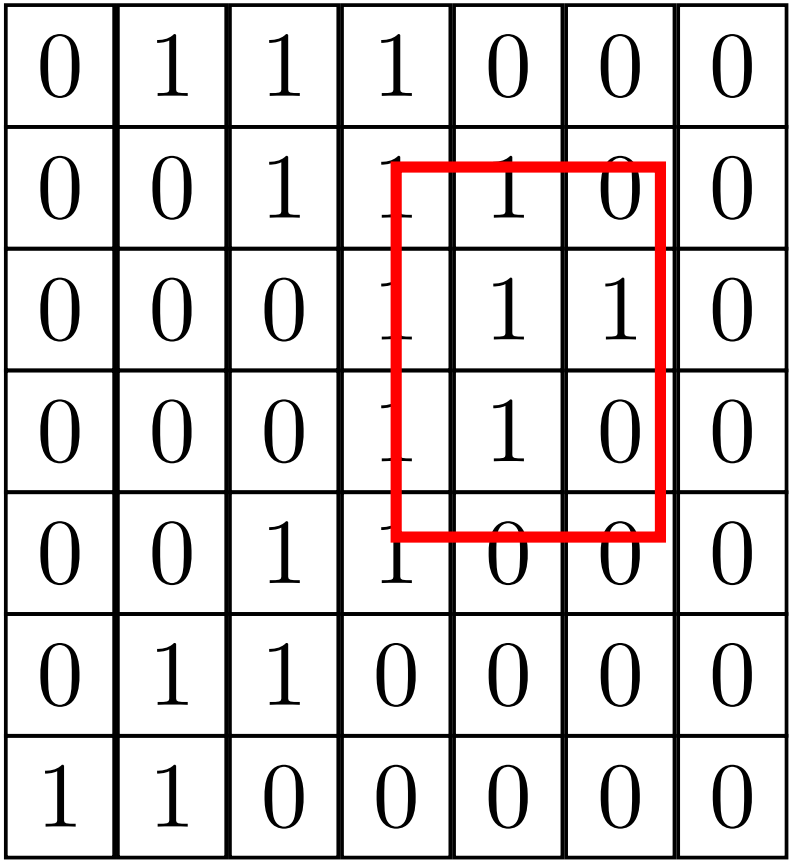
ResNet Building Blocks



RoI-Align


ResStage5 slow pathway
ResStage5 fast pathway
3D Average Pooling
3D Average Pooling
RoIAlign
RoIAlign
2D max pooling
2D max pooling
Dropout
Linear Projection
Sigmoid Classifier
$\hat{y}$

RoI-Align


0 1 1 1 0 0 0
0 0 1 1 1 0 0
0 0 0 1 1 1 0
0 0 0 1 1 0 0
0 0 1 1 0 0 0
0 1 1 0 0 0 0
1 1 0 0 0 0 0
I