Nicolò Bertozzi, Francesco Bianco Morghet
Machine Learning and Deep Learning
Master Degree in Data Science Engineering
Politecnico di Torino

Aid of WFCNet for FPAR Task

Project Description

2^{nd} Semester | 13 July 2020

Table of Contents

Introduction

Dataset

Related Works

Proposed Methods

Results

Overview

Introduction

Dataset

Related Works

Proposed Methods

Results

Introduction 1/4

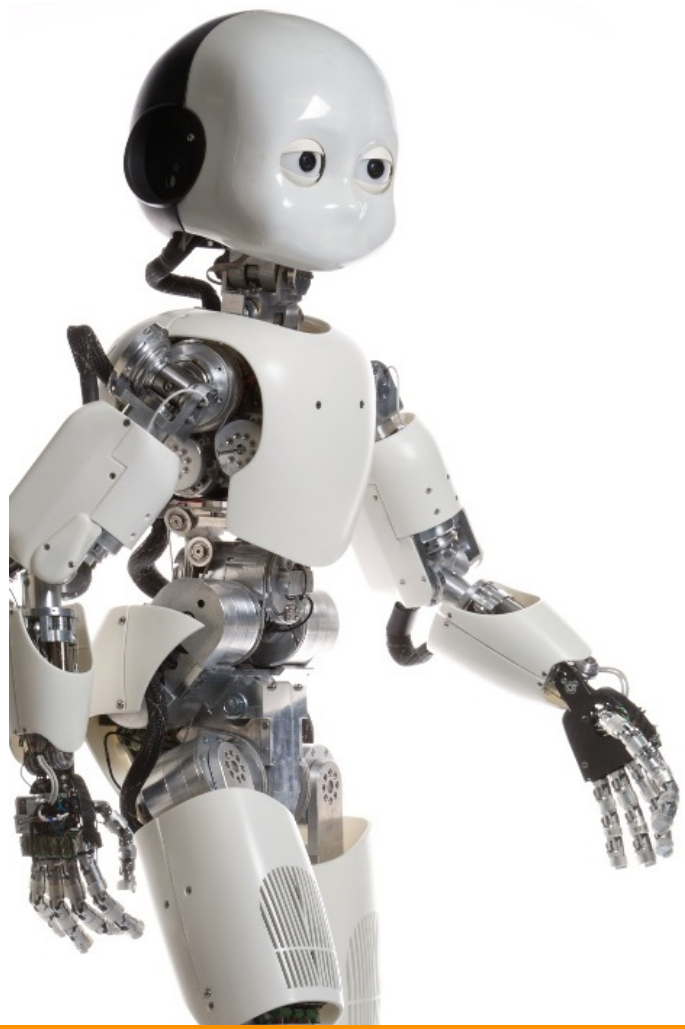
Goal:

- Record videos with the same cameraman's point of view;
- Recognize the actions performed by the subject;

Introduction 2/4

Interested Areas:

- Android intelligence;
- Autonomous driving;
- Surveillance;
- Loyalizing users' experience;



Introduction 3/4

Issues:

- Small datasets;
- Presence of **parts of the cameraman's body** in the video;
- The action **must be represented** by a *verb + noun*;

Introduction 4/4

Solutions:

- Sales of **wearable devices**;
- Incrementing chance of **having at hand a camera**;
- Incrementing number of **images taken every day** [1];
- Deeper neural networks;

Overview

Introduction

Dataset

Related Works

Proposed Methods

Results

Dataset

Dataset

Overview

Introduction

Dataset

Related Works

Proposed Methods

Results

Two Stream Approach 1/2

Main characteristics:

- Two **CNNs**: one to **extract features** from RGB images and one to **extract features** from flow images;
- **ConvLSTM** to take into account the **temporal dependencies**;
- Linear **classifier** to **join the networks**;

Two Stream Approach 2/2

Issue:

- The correlation and the **mutual influence** between motion and appearance information is **not taken into account**;

Solution:

- Implementing a single network accompanied by a **motion segmentation task**;

Motion Segmentation Task 1/2

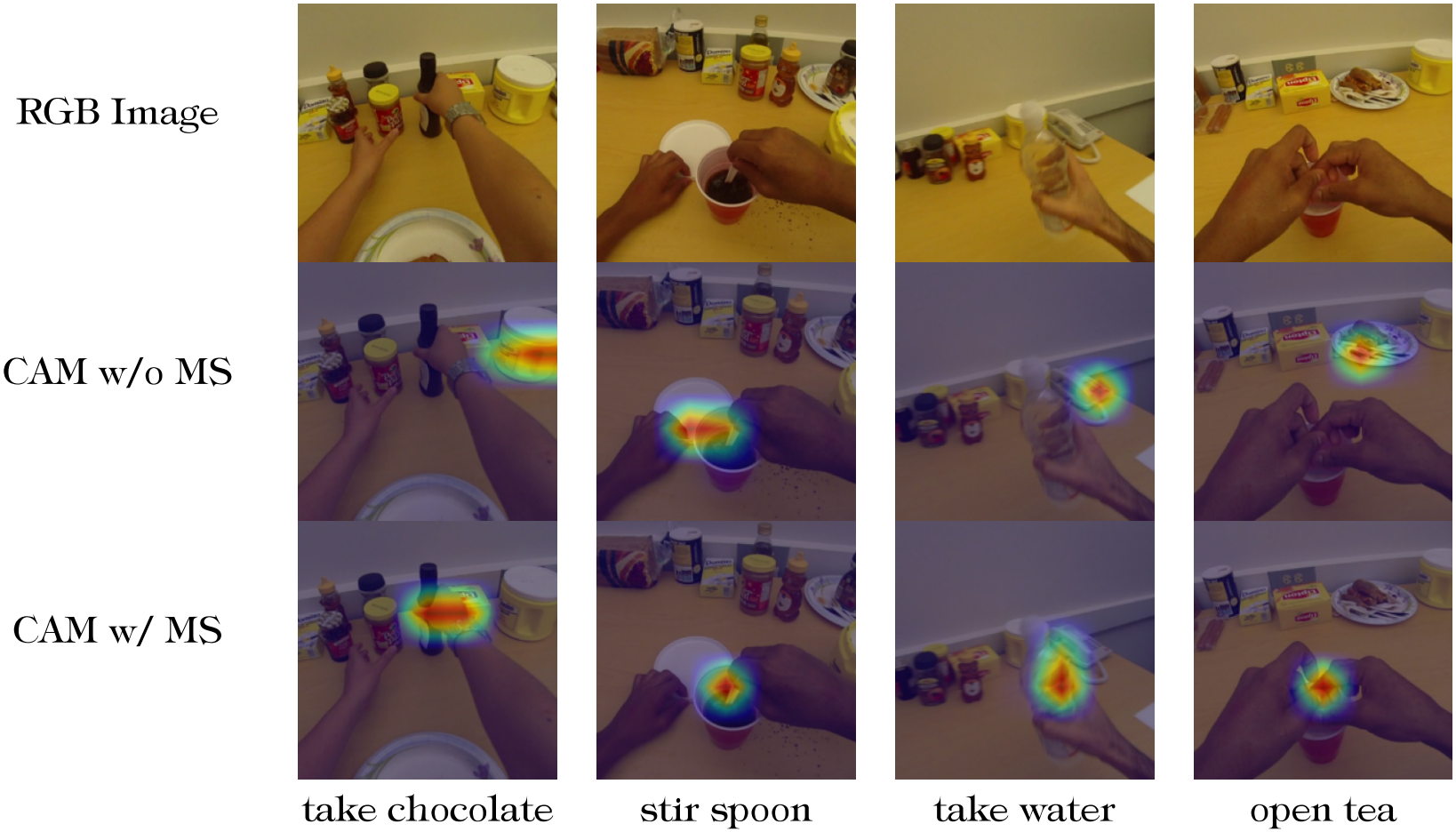
Features:

- Each **feature map** is forwarded to an **auxiliary branch** with a convolutional and a FC layer;
- **IDT** as ground truth: image which indicates if a **pixel is moving or not**, net to the camera motion;
- **Pixel-per-pixel loss** between the predicted motion map and the IDT;

Motion Segmentation Task 2/2

Formule

CAMs Visualizations



Attention Mechanism 1/2

Features:

- **Focusing** the recognition on the **most important parts** of the video;
- **Discarding** the regions with **low importance**;
- The temporal flow information, i.e **the motion**, **is not included** in the mechanism;

Attention Mechanism 2/2

Formule

Overview

Introduction

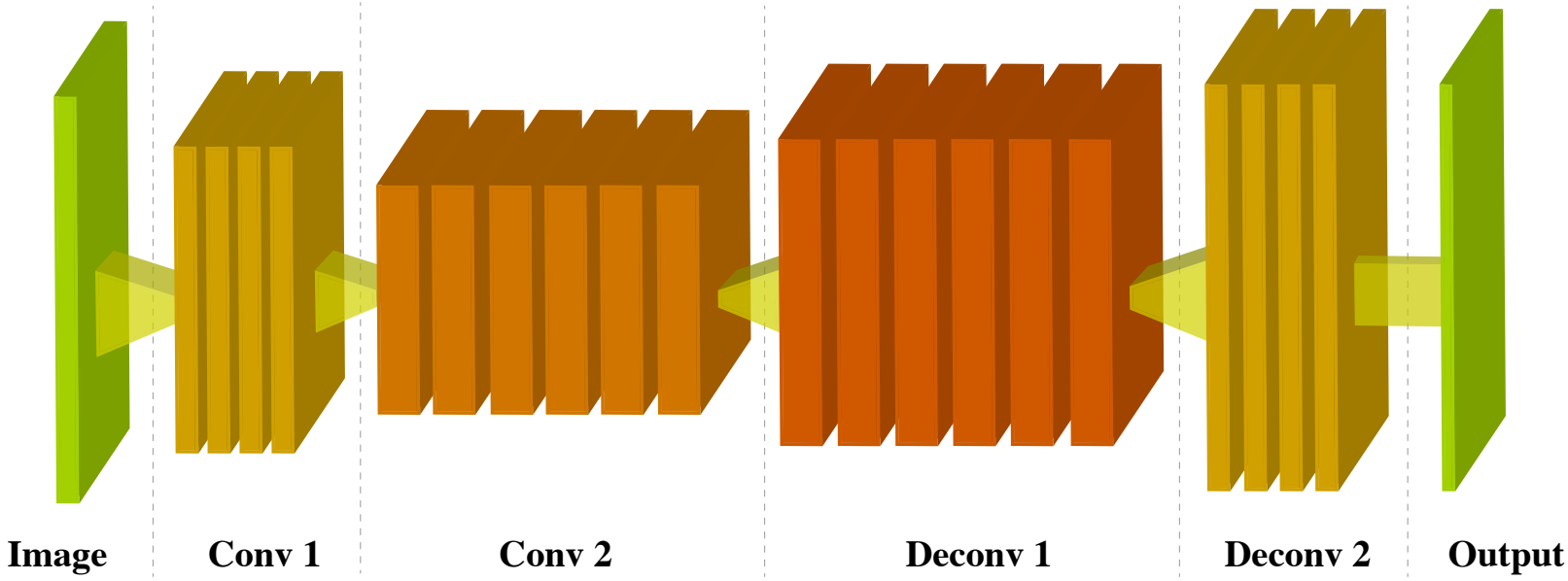
Dataset

Related Works

Proposed Methods

Results

WFCNet 1/2

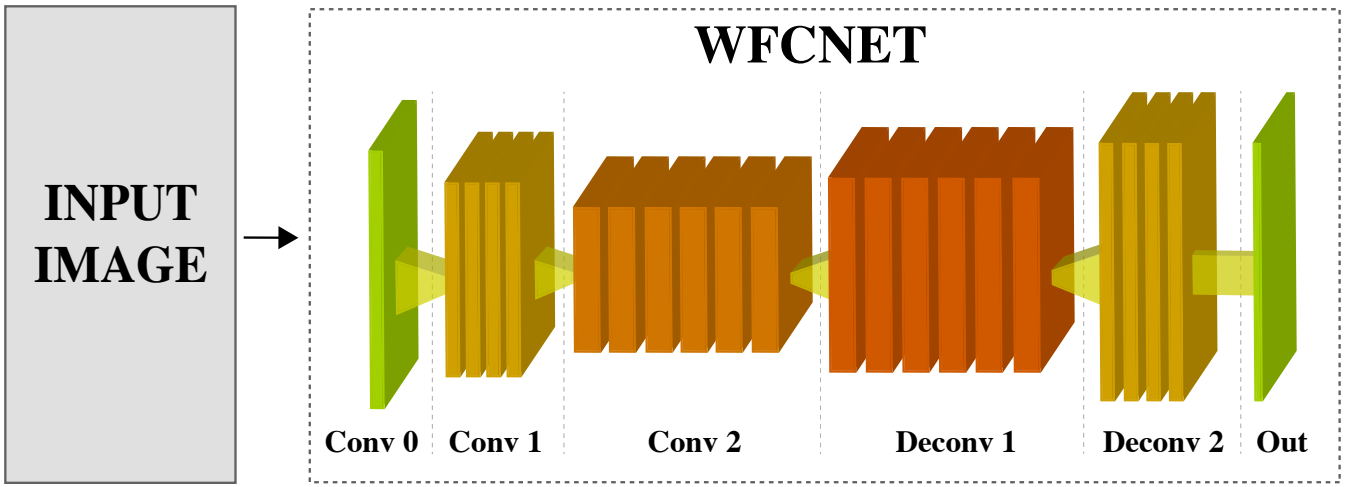


WFCNet 2/2

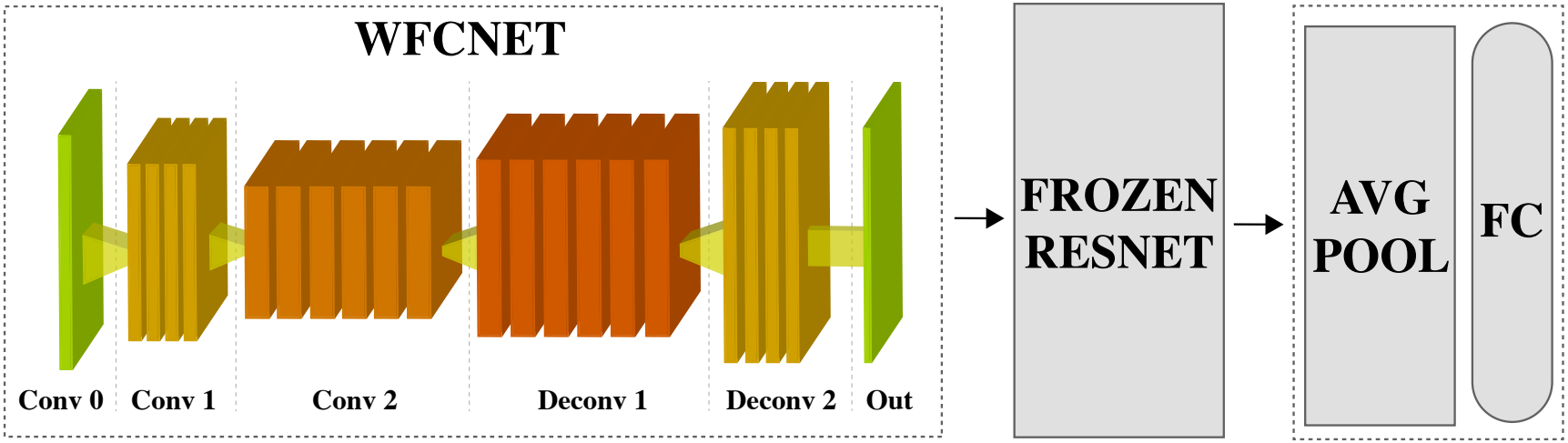
Network's composition:

- **Macro blocks** execute directly the **downsampling** or the **upsampling**;
- **Downsampling as convolutional filters** which maintain the pros of residual blocks;
- **Upsampling as neighbour resize** which performs better than transpose convolution;
- Finally the **activation function** with sigmoid and the **normalisation** with mean and std of ImageNet are applied;

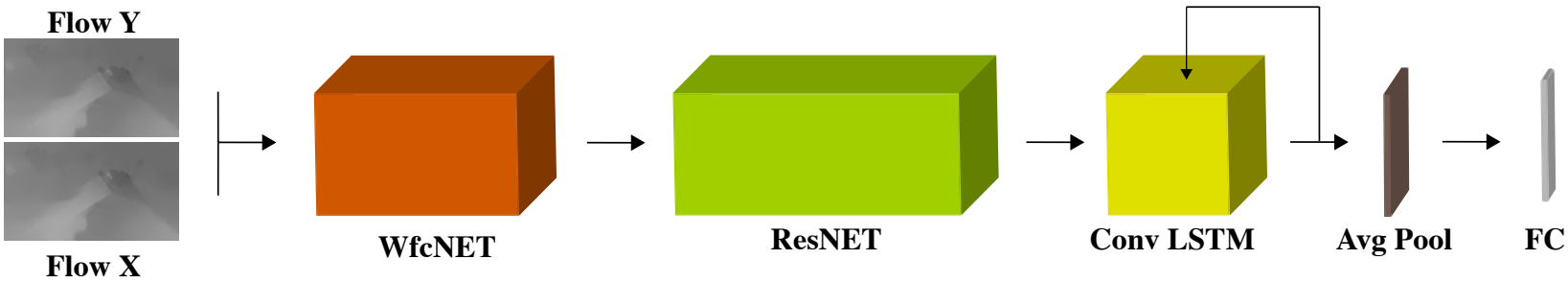
Training WFCNet 1/2



Training WFCNet 2/2



Single Stream 1/2

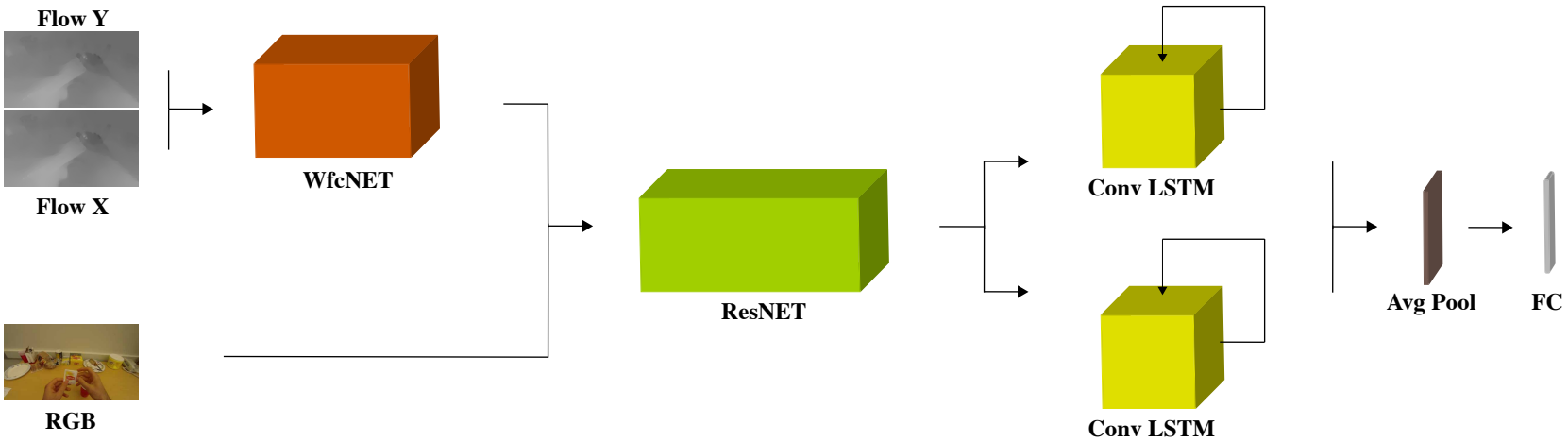


Single Stream 2/2

Basic features:

- A **WFCNet** to infer **RGB warp flow** frames;
- A **ResNet**, with an attention mechanism implemented, **trained on ImageNet**;
- A **ConvLSTM** to encode the **temporal correlations** between the spatial maps;
- An **Average Pooling** layer and a **FC** layer;
- Due to the kind of problem, i.e classification, a **Cross Entropy Loss is used**;

Two Stream 1/2



Two Stream 2/2

Overview

Introduction

Dataset

Related Works

Proposed Methods

Results

References

Caroline Cakebread
"People will take 1.2 trillion digital photos this year — thanks to smartphones"
Businessinsider.com, 1 September 2017
Available at:
https://www.businessinsider.com/12-trillion-photos-to-be-taken-in-2017-thanks-to-smartphones-chart-2017-8?IR=T

The End

Thank you for your attention!

Nicolò Bertozzi
Francesco Bianco Morghet

FPAR Project | MLDL
10 July 2020