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

Overview

Introduction 1/4

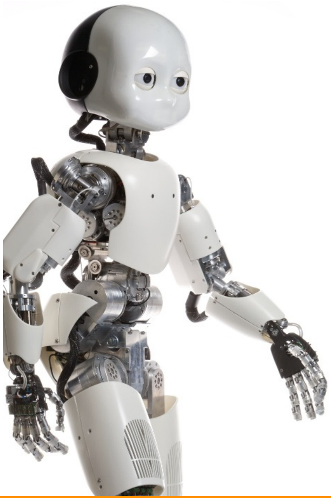
Goal:

- 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:

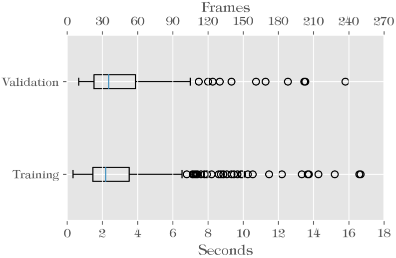
- Sales of **wearable devices**;
- Incrementing chance of **having at hand a camera**;
- Incrementing number of **images taken every day** [?];
- Deeper neural networks;

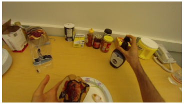
Overview

Dataset

GTEA-61:

- First person point of view;
- 61 classes of *verb+noun*;
- 4 subjects;
- Most clips 1.5s to 4s long.



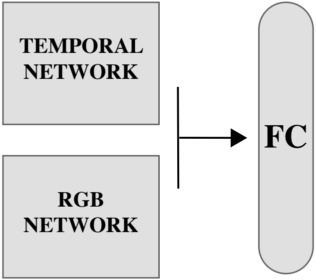


Overview

Two Stream Approach 1/2

Main characteristics:

- **Two networks** with separate **CNNs**: one to **extract features** from RGB images and one to **extract features** from warped optical flow frames;
- **ConvLSTM** in the RGB network to take into account the **temporal dependencies**;
- Linear **classifier** to **join** the two networks.



Two Stream Approach 2/2

Issue:

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

Possible solution:

- Implementing a single network with an auxiliary **self supervised 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

- Both tasks are jointly trained by minimising single loss:

$$\mathcal{L} = \mathcal{L}_{main} + \alpha \mathcal{L}_{ms} \quad , \quad \alpha \in \Re$$

- Loss of classification task:

$$\mathcal{L}_{main} = -\sum_{i}^{N} y_i \log(p(x_i)) \qquad y_i = [0 \dots 1 \dots 0] \quad p(x_i) = \frac{1}{\sum_j e^{f_j}} [e^{f_0} \dots e^{f_c}]$$

x_i sample, y_i label, f_j output neuron for class j

- Loss of motion segmentation task:

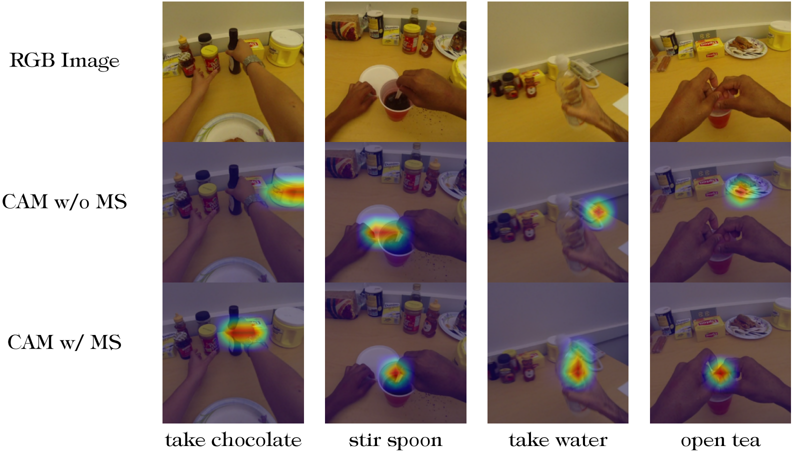
$$\text{CLF: } \mathcal{L}_{ms} = -\sum_{i}^{N}\sum_{t}^{T}\sum_{s}^{S} m_{i,t,s} \log(l_{i,t,s}) \qquad m_{i,t,s} \in \{[1,0],[0,1]\} \; m'_{i,t,s} \in \{0,1\}$$

$$l_{i,t,s} = \left[\frac{e^{f_{s_0}(i,t)}}{\sum_j e^{f_{s_j}(i,t)}}, \frac{e^{f_{s_1}(i,t)}}{\sum_j e^{f_{s_j}(i,t)}} \right]$$

$$\text{REG: } \mathcal{L}_{ms} = -\sum_{i}^{N}\sum_{t}^{T}\sum_{s}^{S} (m'_{i,t,s} - l'_{i,t,s})^2 \qquad l'_{i,t,s} = .5 \tanh(f_{s_0}(i,t) + f_{s_1}(i,t)) + 1$$

i sample, t time-step, s spatial location, m mmap, l predicted mmap,
f_{s_j} output neuron for spatial location s and pixel value $j = 0|1$

CAMs Visualizations



Attention Mechanism 1/2

Features:

- **Focusing** the recognition on the **most important parts** of the video;
- **Discarding** the regions with **low importance**;

Attention Mechanism 2/2

Attention mechanism in Ego-RNN:

1. Find best neuron of hidden fc layer: $\text{argmax}_c(\sum_l(\text{avgpool}_i(f_i(l)) \cdot w_l^c) + b^c)$
2. Compute CAM for all spatial locations: $\text{CAM}_c(i) = \sum_l w_l^c f_l(i)$
3. Compute features with spatial attention:
   $f_{SA} = CAM' \odot f$
   $CAM'(i) = \frac{e^{CAM(i)}}{\sum_i e^{CAM(i)}}$

Proposed simpler attention mechanism:

1. Compute AM for all spatial locations: $\text{AM}(i) = \sum_l w_l f_l(i)$
2. Compute features with spatial attention:
   $f_{SA} = AM' \odot f$
   $AM'(i) = \frac{e^{AM(i)}}{\sum_i e^{AM(i)}}$

i spatial location index, l output feature map index, c hidden neuron index
$f_l(i)$ backbone l-th output feature map at spatial location i,
w_l^c l-th weight of c-th neuron, b^c bias of c-th neuron,
w_l l-th weight of linear classifier

Overview

WFCNet 1/2

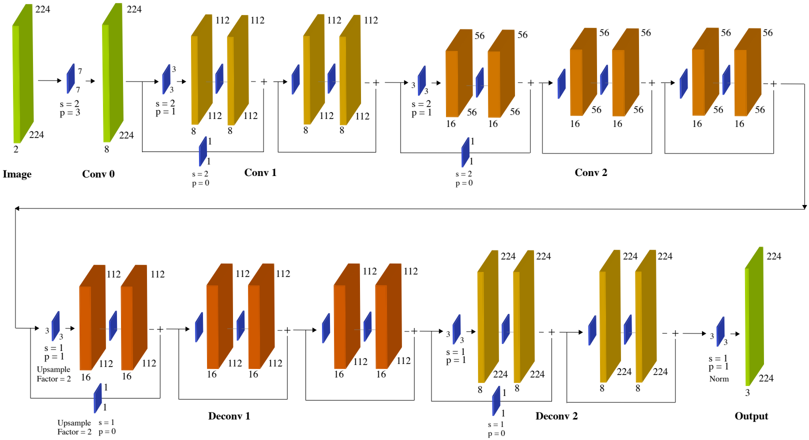


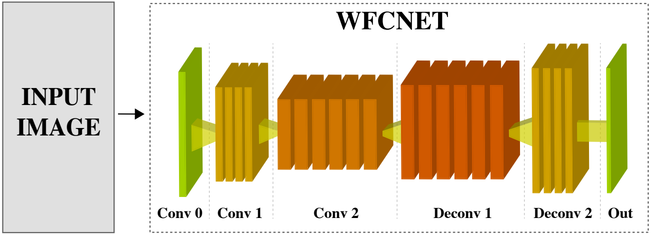
Image
Conv 0
s = 2
p = 3
Conv 1
s = 2
p = 1
s = 2
p = 0
Conv 2
Upsample
Factor = 2
s = 1
p = 1
s = 1
p = 0
Deconv 1
Deconv 2
Norm
Output

WFCNet 2/2

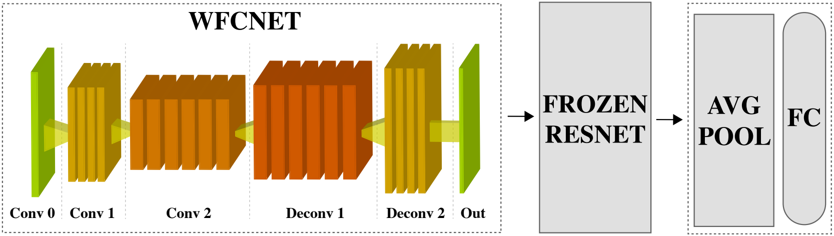
Network's composition:

- **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

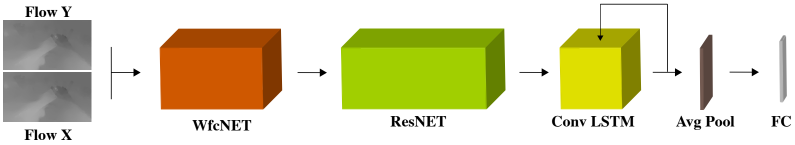


Training WFCNet 2/2



Colorized Blocks

Single Input 1/2

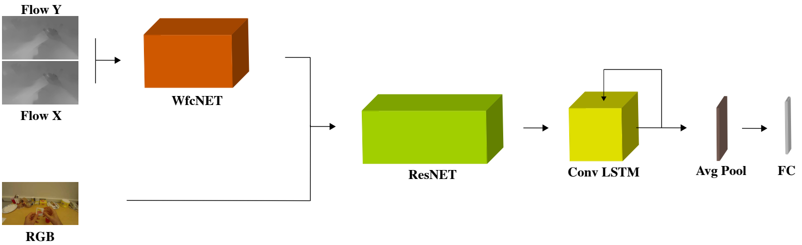


Single Input 2/2

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 Input 1/2



Two Input 2/2

Limitations of single input:

- **The warp flow alone is not sufficient** to achieve high results;
- **The warp flow** is not generated with special measurement equipment, so it **represent a simple domain projection**;
- The **appearance** is **discarded**;

Solution:

- **Analyze both** warp flow and RGB frames;

Overview

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

Aid of WFCNet for FPAR Task | MLDL
13 July 2020