EMOTIC

Context Based Emotion Recognition using CNN

Simone Mattia 901716 s.mattia2@campus.unimib.it
Daniel Montalbano 897383 d.montalbano1@campus.unimib.it

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

About us

SIMONE MATTIA

- Education
  - Cyber Security @unimi
  - Data science @unimib
- Work
  - Security Engineer @Cybergon

DANIEL MONTALBANO

- Education
  - Cyber Security @unimi
  - Data science @unimib
- Work
  - Frontend Engineer @NotJustAnalytics

Emotion Recognition

WHAT IS

Emotion recognition is the process of identifying and categorizing human emotions based on cues like facial expressions, vocal intonations, body language, and context.

TASKS

- Facial Expression Recognition
- Multimodal Emotion Recognition
- Context-Based Emotion Recognition

APPLICATIONS

In human-computer interaction, virtual reality, healthcare, market research, and sentiment analysis, enabling better communication and understanding of human emotions in various domains.

EMOTIC dataset[1]

Database of images with people in real environments, annotated with their apparent emotions.

SOME NUMBERS

- 23,571 images
- 34,320 people annotated
- 66% male - 34% female
- 83% adults - 17% others
- 25% of people have their face partially occluded or very low resolution

ANNOTATIONS

The images are annotated with bounding box and an extended list of 26 emotion categories combined with the three common continuous dimensions Valence, Arousal and Dominance.

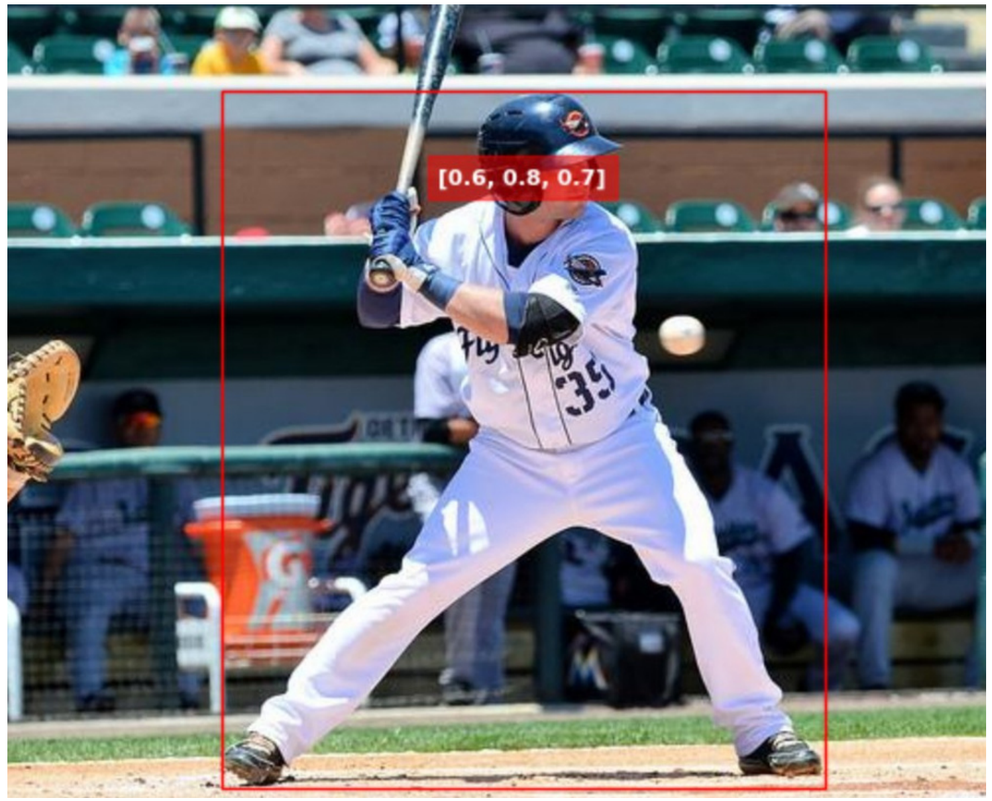
OUR GOAL

We want to process only the images with one person, find the relative bounding box and predict its three continuous dimensions: Valence, Arousal and Dominance.

(1) Emotic dataset

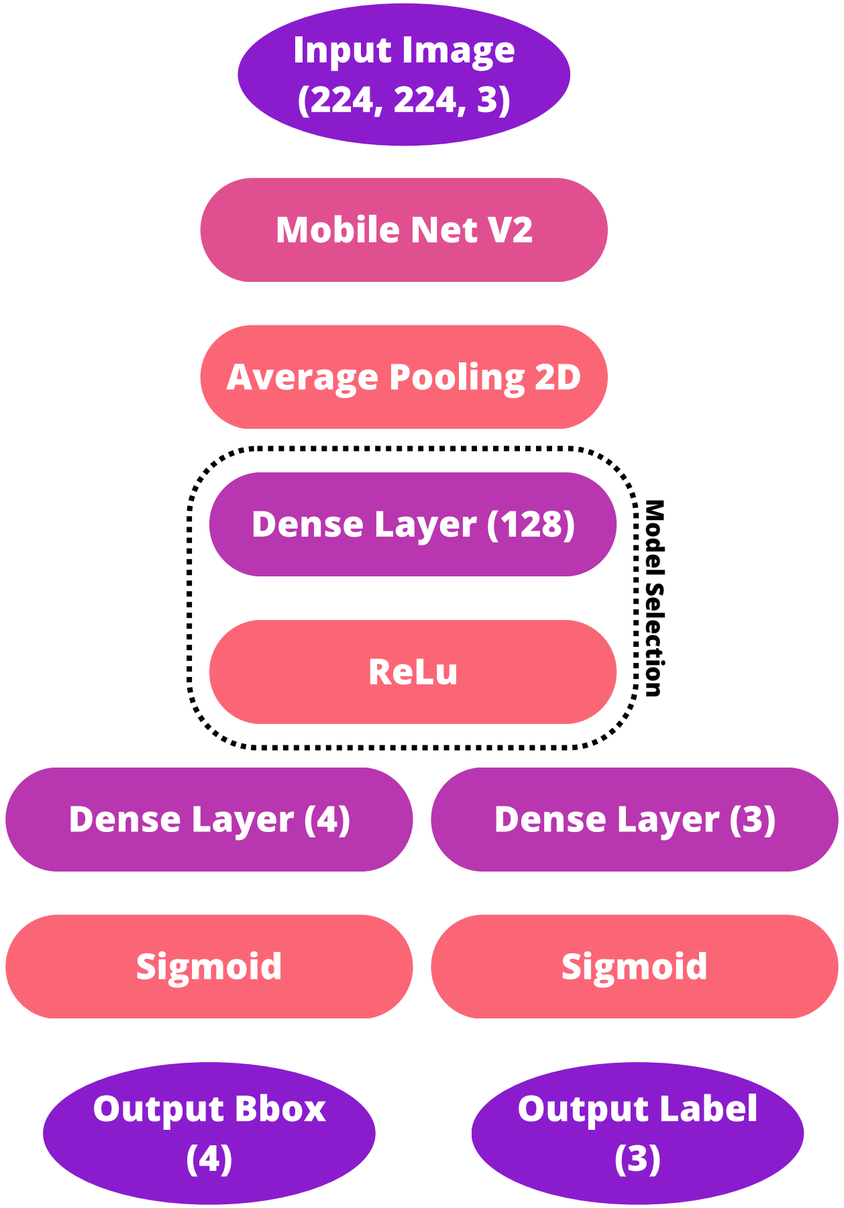
Pre Processing

- **Parsing from Matlab format**
- **Data Normalization**
  - Images
  - Box - Label
- **Sampling**
  - Performance
  - Simplification



Our Solution

- **One Input:** image
- **Transfer Learning of Mobile Net V2**
- **Additional Dense Layer**
- **Two outputs**
  - **Bbox**: bounding box
  - **Lable**: continuous dimensions



Model Selection

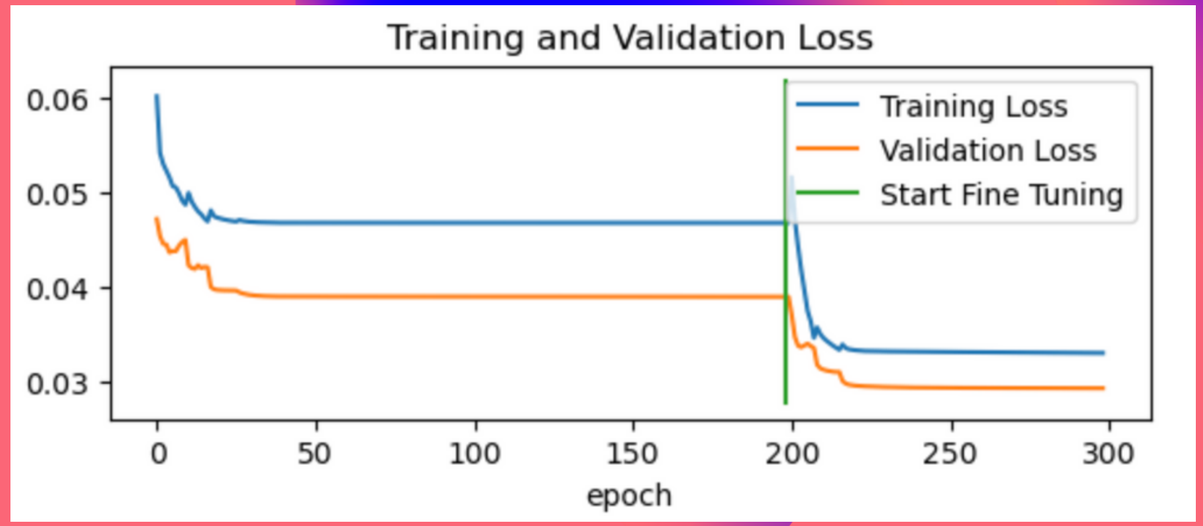
Training Setup:

- **Base model**: Mobile Net V2
- **Dataset size**:
  - Train: 5000
  - Validation: 500
- **Epochs**:
  - 10 Freeze (LR 0.001)
  - 5 Fine (LR 0.0001)
- **Callbacks**:
  - Early Stop: patience 5
  - Reduce LR: patience 2
- **Loss Function**: MSE
- **Fixed Seed**

Layers			Loss		
Conv (256) + Batch Norm. + ReLu	Dropout (0.3)	Dense (128) + ReLu	Val Loss	Val Loss Box	Val Loss Label
			0,0497	0,0267	0,0231
		X	0,0477	0,0261	0,0216
	X		0,0502	0,0265	0,0237
	X	X	0,0523	0,0307	0,0216
X			0,0529	0,0265	0,0264
X		X	0,0612	0,0339	0,0273
X	X		0,055	0,0283	0,0267
X	X	X	0,0545	0,0293	0,0252

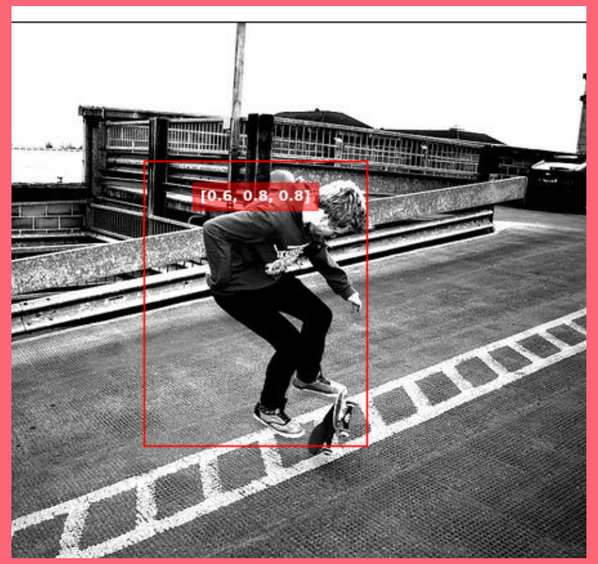
Training

- **Dataset size:**
  - Train: 12632
  - Validation: 1300
- **Epochs:**
  - 200 Freeze (Starting LR 0.001)
  - 100 Fine Tuning (Starting LR 0.0001)
- **Callbacks:**
  - Early Stop: patience 10
  - Model Checkpoint
  - Reduce LR: patience 5
- **Loss Function:** MSE



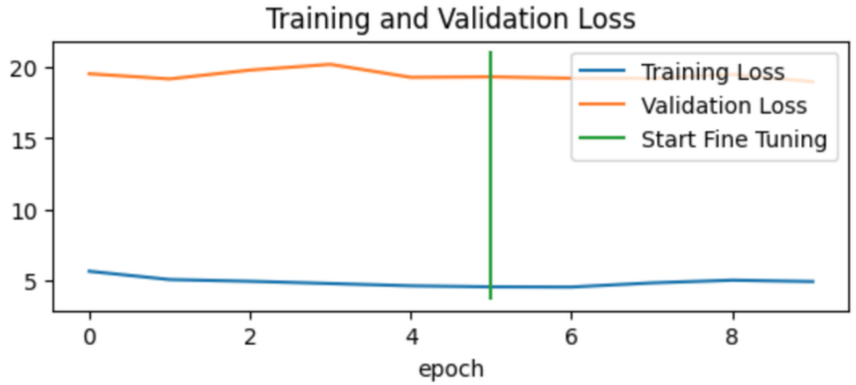
Evaluation

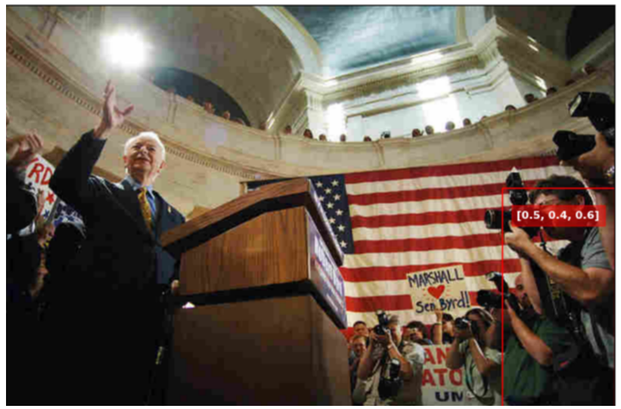
- MSE on Test Set: 0.043
  - Bbox: 0.020
  - Label: 0.023



Problems

- **Data quality of annotations**
  - Missing box and lables
  - Low accuracy
- **String Lables**
  - Over fitting
  - Non convergent training
- **Detecting people**
  - Partially occluded
  - At the edges of the image





Future Developments

MORE COMPLEX NETWORK

- RCNN / SSD
  - Detect more people in one image
  - Detect occluded people
  - Handle images with no people
- Data augmentation

IMPROVE DATA QUALITY

- More accuracy in annotations
- Data integration

STRING LABELS

- Return also string label as network output
  - Categorical labels

Thank you for your time

Simone Mattia
Daniel Montalbano