Deep Learning for Universal Emotion Recognition in Still images

Juan Luis Rosa Ramos

April 2018

Advisors: Dr. Sergio Escalera and Dr. Andrés Cencerrado
Master in Artificial Intelligence 2016 - UPC, UB, URV

Motivation: face expressions are universal

Human emotions produce physiological changes: heart rate, breathing rate, perspiration, hormone levels or facial expressions

[1] Even in blind individuals

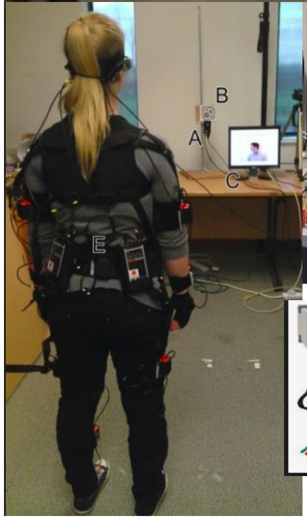
[2] Even in nonhuman primates

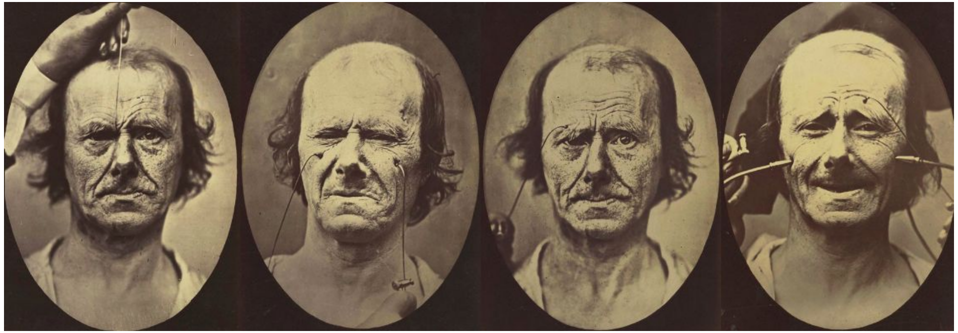
1 - *Pamela M Cole, Sarah E Martin, and Tracy A Dennis. Emotion regulation as a scientific construct: Methodological challenges and directions for child development research. Child development, 75(2):317–333, 2004.*
2- *Amy S Pollick and Frans BM De Waal. Ape gestures and language evolution. Proceedings of the National Academy of Sciences, 104(19):8184–8189, 2007.*

Motivation: apply new technology

Using computer vision and machine learning methods

Less intrusive methods, better understanding





New startup related to driving assistance

Driver's state by Face Analysis

Driver's drowsiness

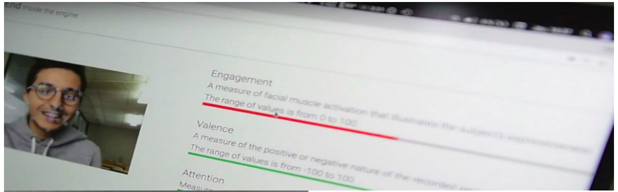
Driver's distraction

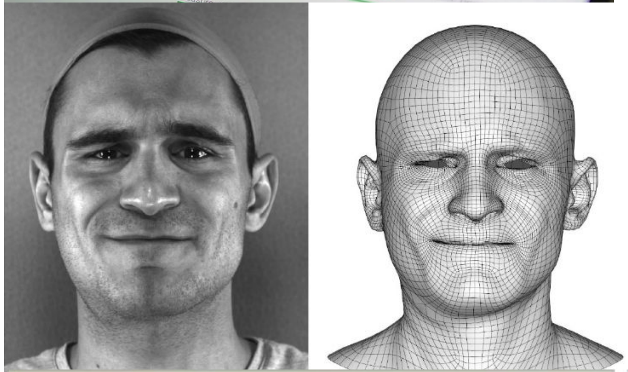
Some applications

Helping doctors: detection of pain[1]

Product management: frustration, engagement [2]

Affective computer, HCI, Robotics [3]





1 - Roy, Sourav Dey, et al. "An approach for automatic pain detection through facial expression." Procedia Computer Science 84 (2016):
2 - *The software, called Nestor, will be used two online classes at the ESG business school from Paris*
3 - *Gunes, Hatice. "Automatic, dimensional and continuous emotion recognition." (2010)*
4- *Face2face: Real-time face capture and reenactment of rgb videos. In Computer Vision and Pattern Recognition (CVPR), 2016*

Objectives

Understand human emotions through automated face analysis

Present a software methodology for solving this classification problem

Provide recommendations learned by implementing it

Keywords: Facial Expression Recognition; Convolutional Neural Networks; Overfitting; Face analysis dataset

Presentation structure

Presentation will follows methodology pipeline

- Understanding the problem
- Building a dataset and preprocessing data
- Fine-tuning and testing different CNN models
- Tricking the model searching for overfitting
- Inference recommendations

Dataset

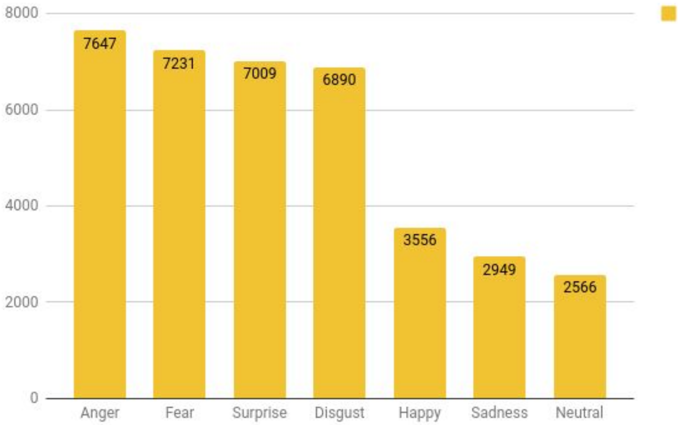
Step 1 choosing a classification model.

Psychologist Paul Ekman[1] 6 basic emotions + neutral

1 - Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. Emotion in the human face: Guidelines for research and an integration of findings. Elsevier, 2013.

Step 2 merging datasets

Class	Total images	JAFFE	CK+	Kdef	Radboud	Internet	Pain expr	Oulu-CASIA	Dartmouth
Anger	6647	30	546	111	201	375	46	5178	160
Fear	6231	32	293	111	201	47	273	5114	160
Surprise	6509	30	730	112	201	173	92	5011	160
Disgust	4385	29	479	111	201	112	48	3245	160
Happy	3556	31	714	111	201	355	46	1938	160
Sadness	2949	31	297	112	201	180	0	1968	160
Neutral	2566	30	1633	111	201	383	48	0	160
SUM	32843	213	4692	779	1407	1625	553	22454	1120



Unbalanced dataset

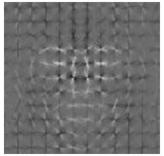
Step 3 Preprocessing data

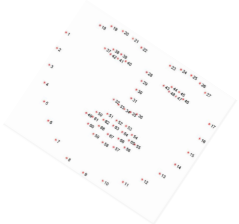
Facial recognition algorithms benefits of face alignment [1]

Face detection: HOG detector and SVM classifier applied in sliding windows [2]

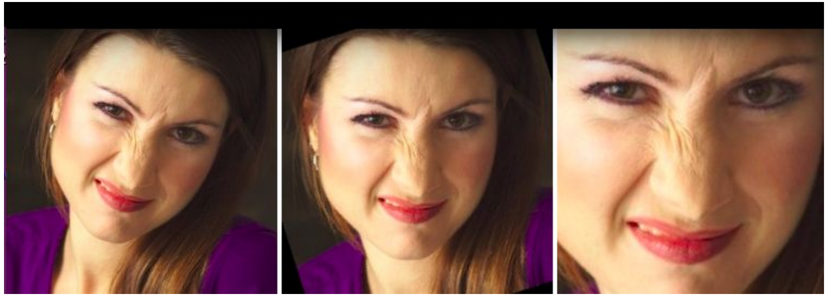
Face alignment: landmarks detection from a DLIB shape predictor with an Ensemble of Regression Trees[3]

Warp transformations: rotation and cropping





Keep a 25% of the bounding box image



1 -Gary B Huang, Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report

2 -Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005..

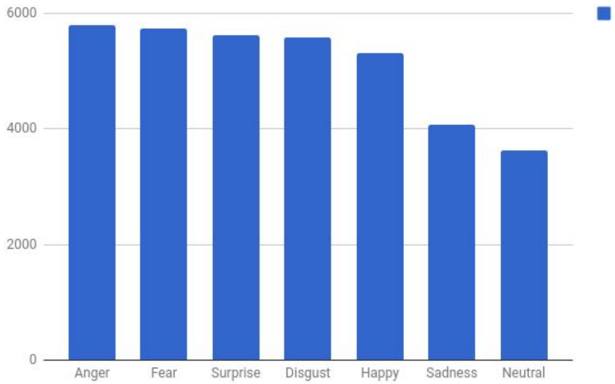
3 -One Millisecond Face Alignment with an Ensemble of Regression Trees by # Vahid Kazemi and Josephine Sullivan, CVPR 2014

Step 4 Data augmentation

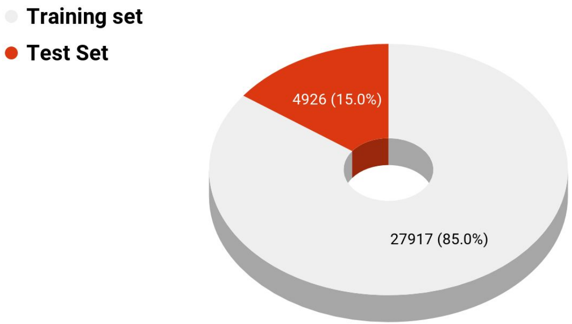
7562 new generated images

Affine transformations: Flip horizontally, rotate the image x degrees to the left/right, apply blur and a small quantity of noise

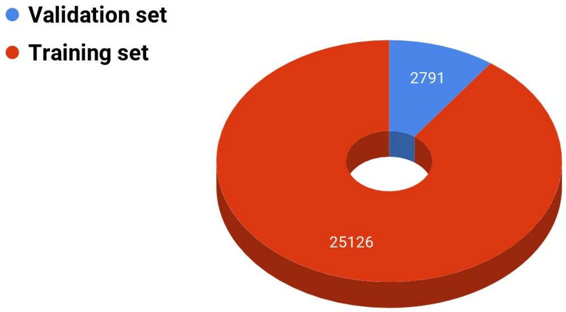
Balanced dataset of 35741 images



Step 5 final number and sets distribution

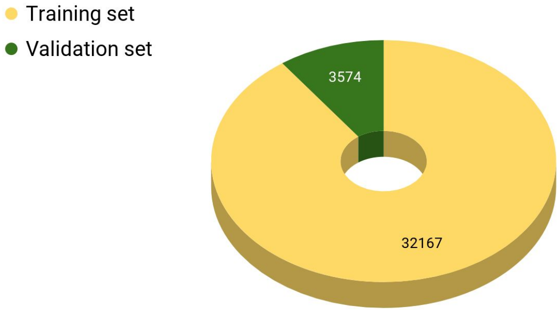


32843 Images
Isolated Test Set of 4926 images



Original dataset
27917 images

10% validation



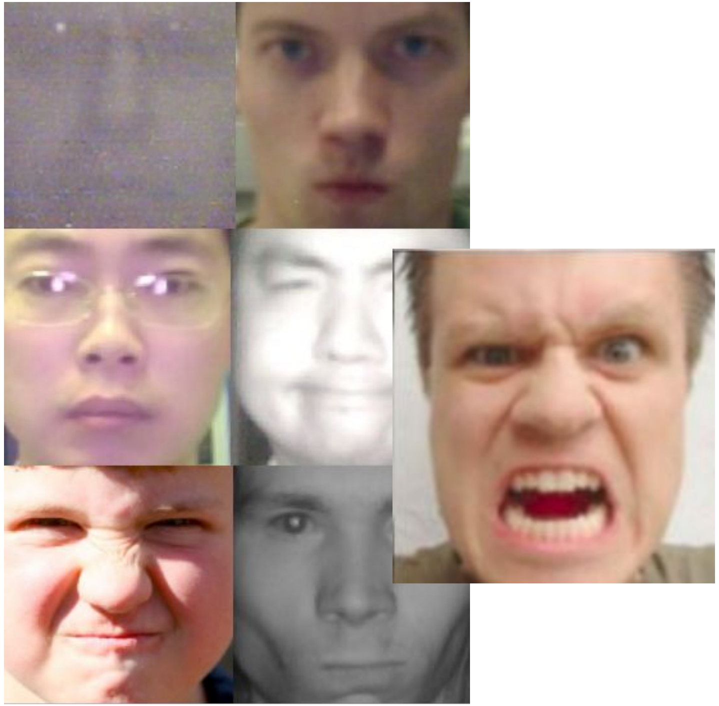
Augmented dataset
35741 images

10% validation

Final dataset doubts

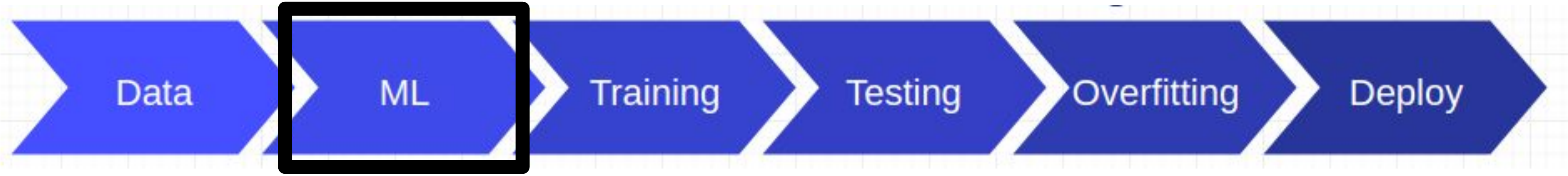
Will my model work with intensities representation?

How accuracy will be affected by occlusions and illumination changes?

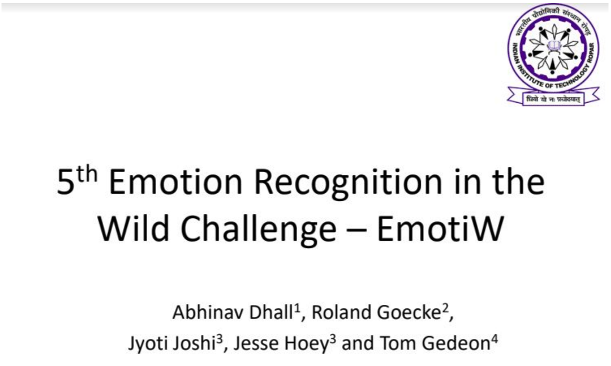


Intravariety in Anger class

Model



1. Choosing an architecture



5th Emotion Recognition in the Wild Challenge – EmotiW

Abhinav Dhall[1], Roland Goecke[2], Jyoti Joshi[3], Jesse Hoey[3] and Tom Gedeon[4]

Based on the EmotiW Challenge conclusions:

- **Deep learning** based methods outperform traditional vision, machine learning methods.
- Transfer learning shown useful in several task.

I chose three CNN architectures:

- **Alexnet** [1] for it’s simplicity - 8 layers
- **VGG16** for it’s generalization abilities - 16 layers
- **Resnet 101** [2]for it’s complexity - 101 layers

1 - *Farfade, Sachin Sudhakar, Mohammad J. Saberian, and Li-Jia Li. "Multi-view face detection using deep convolutional neural networks." Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 2015.*

2- *Masi, Iacopo, et al. "Do we really need to collect millions of faces for effective face recognition?." European Conference on Computer Vision. Springer, Cham, 2016.*

2.Training resources

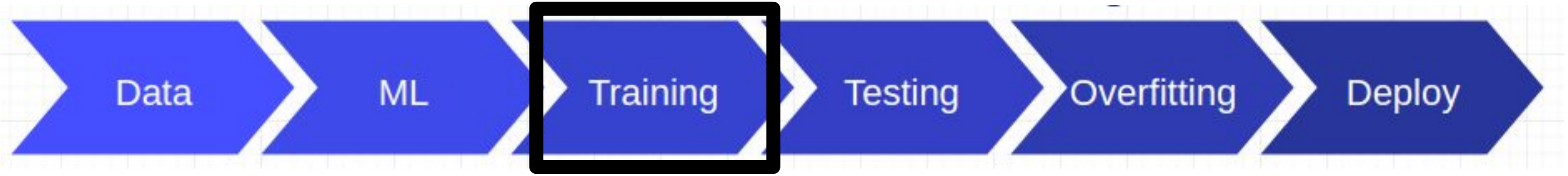
HARDWARE		
Cloud	Virtual machines	2x GPUs
Google Cloud Platform	ubuntu	
amazon web services	n1-standard-4 p2.xlarge --------------- 4 vCPUs, 61 GB	**Pascal P100** 1.6x more double-precision flops 4.7 teraflops **Kepler K80** 2.91 teraflops

SOFTWARE		
Containers	DL	GPUs
docker	Caffe	NVIDIA CUDA, cuDNN

DIGITS	Digits docker version	CuDNN	CUDA	Caffe
6.1.0	18.02	7.11	9.0.176	0.16

Table 8.3: Software used for training

Training time



6 experiments

Same software, same parameters and 2 data sets: Original/Augmented

Same hardware Nvidia Tesla p100 for 30 epochs

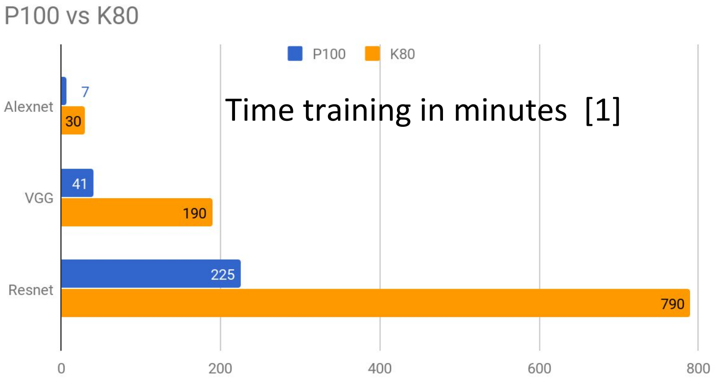
Data **normalization**: mean subtraction

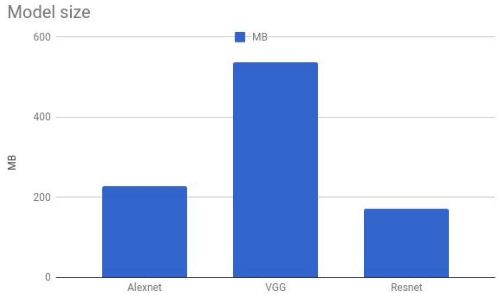
Same measurement: Softmax with Loss + Accuracy

Network	Dataset	Time training	Iterations for 30 epochs	Learned parameters	Solver
Alexnet	Original	7 mins	6570	56,896,903	1681 img/sec
VGG16	Original	41 mins	27960	134,289,223	356.6 img/sec
Resnet	Original	3h 45 mins	41940	42,619,847	68.68 img/sec
Alexnet	Augmented	11 mins	8400	56,896,903	1681 img/sec
VGG16	Augmented	1 h 32 mins	33552	134,289,223	356.6 img/sec
Resnet	Augmented	4 h 45 mins	53640	42,619,847	68.68 img/sec

Results

ALEXNET	Validation Acc	Test Acc
Original	95.24	95.48
Augmented	96.025	95.91
RESNET		
Original	97.211	**97.17**
Augmented	97.315	96.88
VGG		
Original	91.73	93.97
Augmented	93.29	95.82

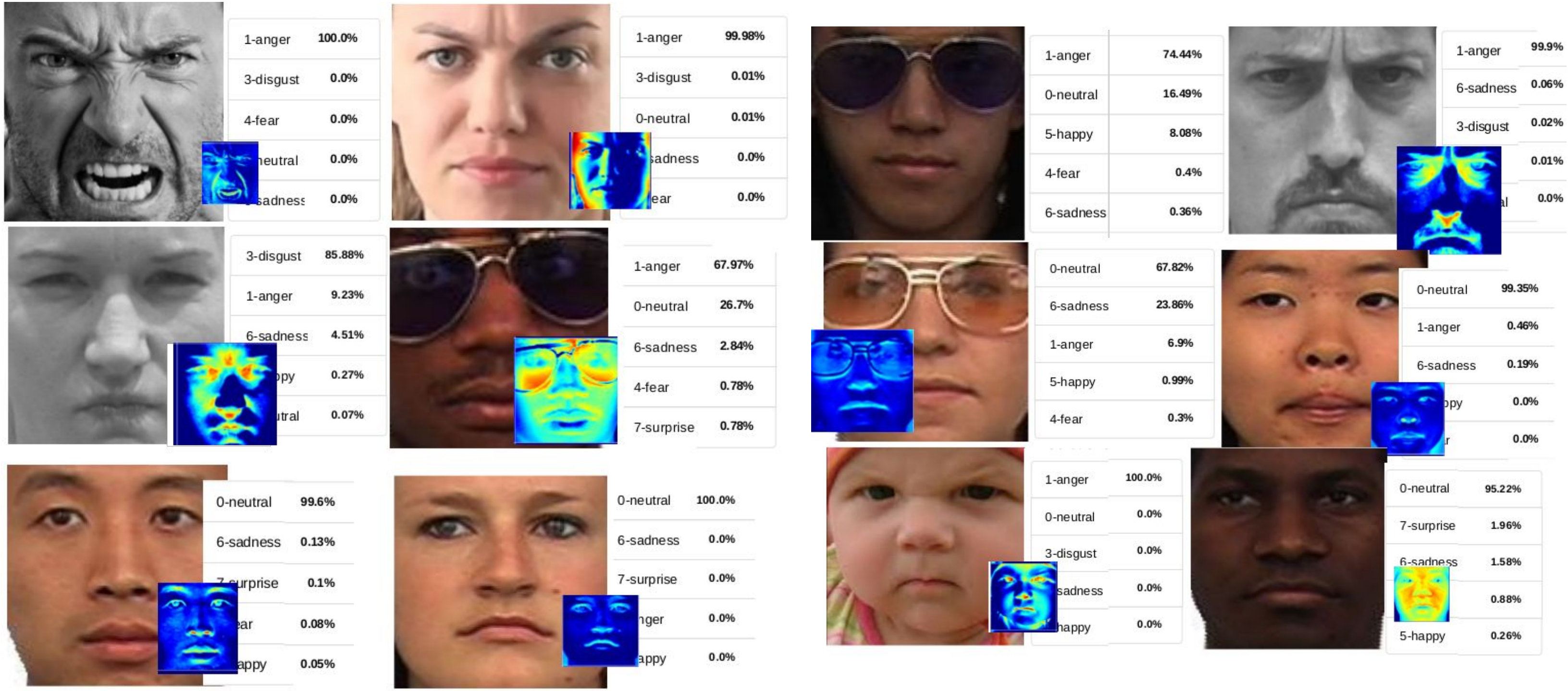




Model Size:
Alexnet: 228MB
VGG16: 537MB
Resnet 101: 171MB

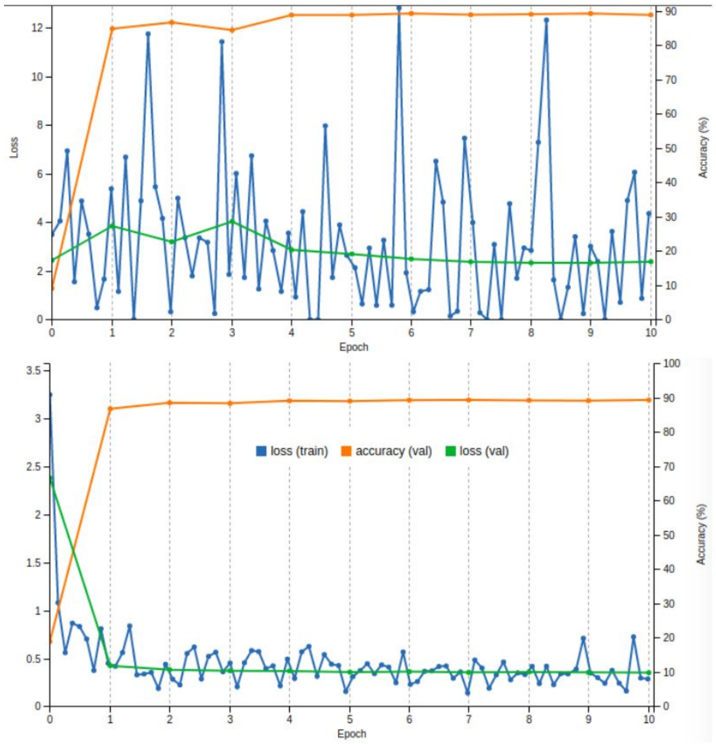
1 - April 2018, cost/hour of a Tesla k80 0,45$ and a Tesla P100 1,60€

Examples, doubts solved



Fine-tuning details

- Learn last Conv + Fully connected ones
- **Initial LR: 0.001**
- LR policy: stepsize: 33%
- Learning algorithm: **SGD**



Loss curve stability differences between different LR (0,01 vs 0.001 in the bottom)

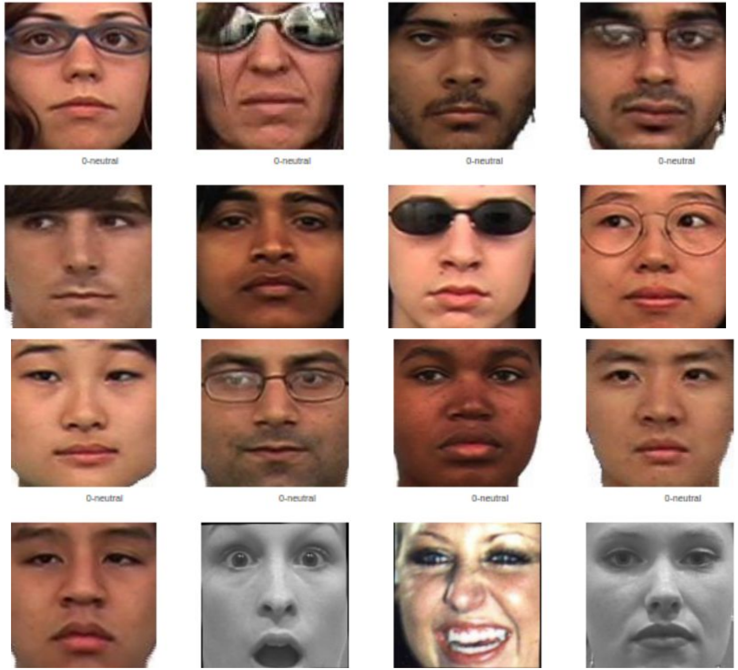
Machine Learning doubts: overfitting?

Capacity of a model to generalize outside of the training set

- Dropout between the FC layers
- Data augmentation
- Transfer learning
- Isolated Test Set

New experiments

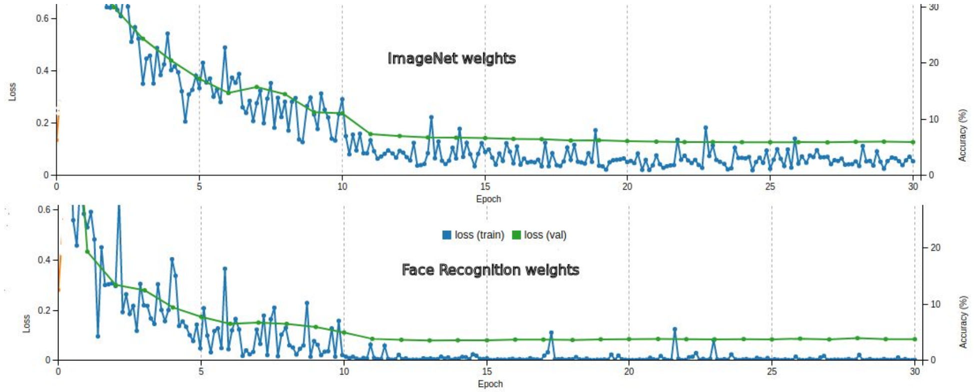
- New overfitting test set



Dealing with Overfitting

Networks with ImageNet weights can't generalize

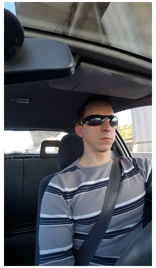
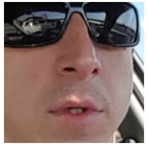
Weights	Validation	Test set	Overfitting set
ALEXNET			
Face	0.95	0.96	0.79
ImageNet	0.96	0.96	**0.4**
VGG			
Face	0.93	0.96	0.8
ImageNet	0.92	0.95	**0.5**
RESNET			
Face	0.97	0.97	0.82
ImageNet	0.96	0.97	**0.65**



Training time

Deployment and inference

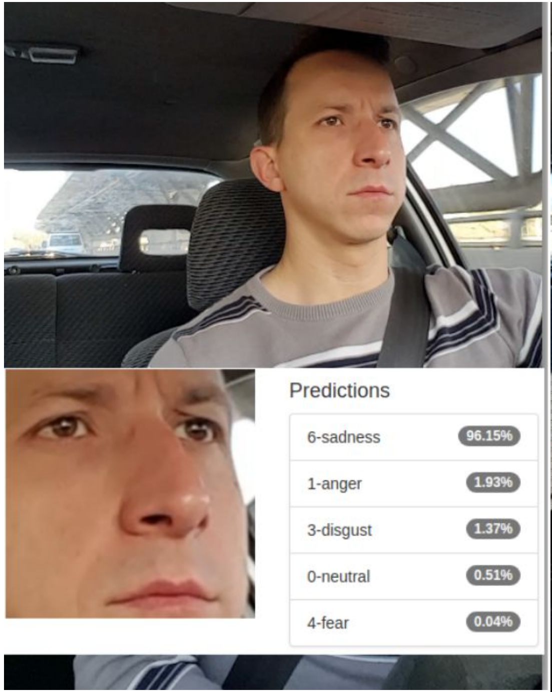
RESTFul Web Service

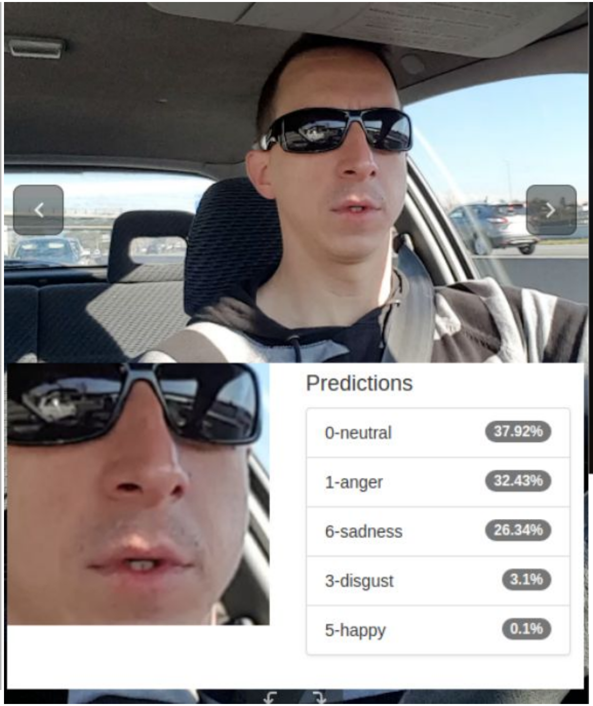
Image	Preprocessing	Input image	Resnet model	Classification
	python dlib OpenCV		Caffe	Predictions 0-neutral 37.92% 1-anger 32.43% 6-sadness 26.34% 3-disgust 3.1% 5-happy 0.1%

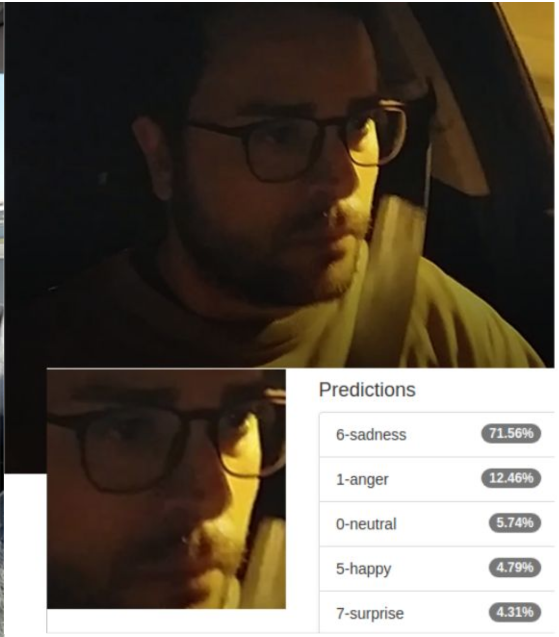
Inference of **742 images** comparing Nvidia P100 GPU and 4vCPUs

Network	GPU P100	CPU
VGG	0m 14s	30m 54s
Alexnet	0m 7s	5m 49s
Resnet	1m 14s	3h 10m 5s

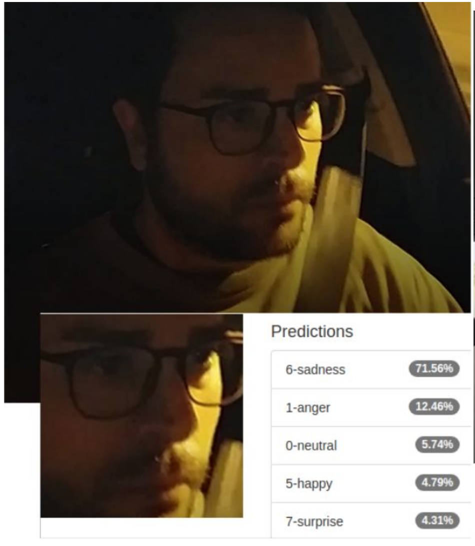
Testing in drivers

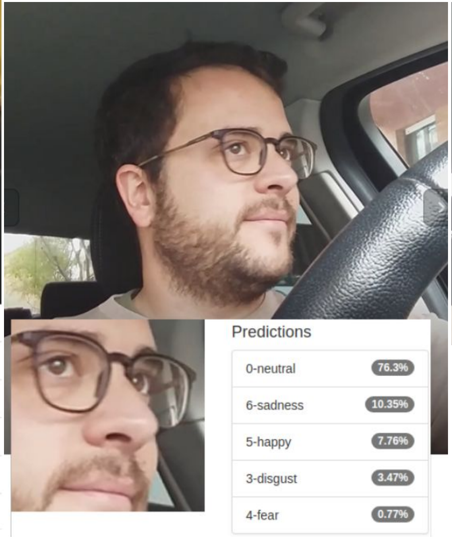


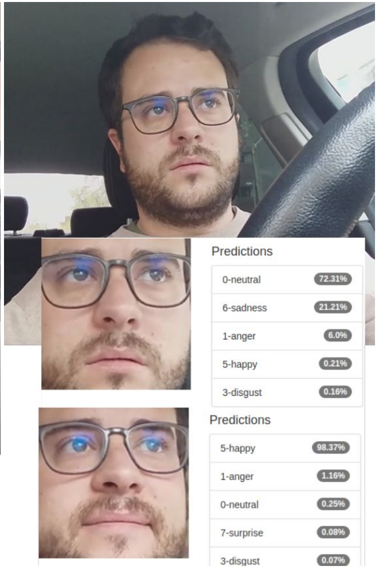


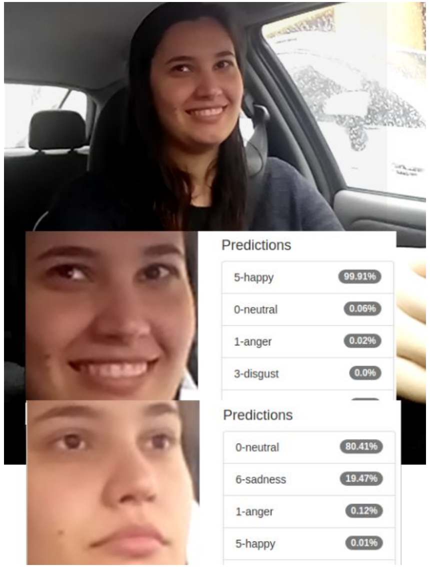


Testing in drivers









Work achievements

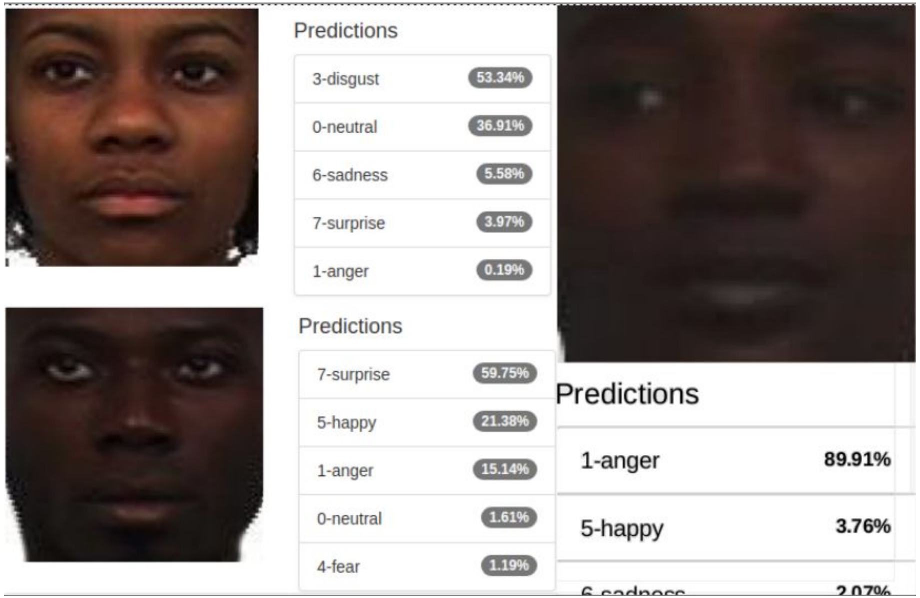
- Dataset for face analysis.
- Face alignment pipeline.
- Experimentation with different models.
- Provide recommendations and time references.

Lessons learned

- Data collection takes time
- Data preprocessing
- Cheap cloud resources
- Dockers

Biased Artificial Intelligence

In software architecture, why not to have a ethical by design principle?



Future work

- Improve detection of extreme poses (spontaneous and naturalistic ones).
- Improve changes in illumination and shadows.
- Work with temporal information detection and transitions

Thank you.

Juan Luis Rosa Ramos