

TECNOLOGIAS MULTIMÉDIA, VISÃO POR COMPUTADOR E INTELIGÊNCIA ARTIFICIAL

Desafios e Perspectivas de Investigação

Paula Viana, Polytechnic of Porto, School of Engineering & INESC TEC

AI Master Class, 2021 Abril 19

CAN YOU IMAGINE A WORLD WITHOUT MULTIMEDIA CONTENT?

AN IMAGE IS WORTH A THOUSAND WORDS

BUT ONLY IF YOU CAN FIND IT

MOTTO

Data, data, data. The capacity of producing information creates tremendous opportunities but introduces also new challenges: ***how can humans process so much information?***

When dealing with **multimedia content**, the concept of **big data is even more relevant** than in other domains. How can I find in my archive, containing **thousands of photos**, the ones that are really relevant for **producing the video clip** that I need to publish in the social networks to announce the fashion show that I intend to promote? And how can I **re-purpose content**, decreasing production costs, while **still creating intelligent context-aware and appealing media**?

How can I find, in all the **thousands of hours of broadcasted TV programs**, the exact instant (video timecode) where the President appears? Or, how can I go directly to the goals of Cristiano Ronaldo in the 90 min video file?

Computer Vision, Artificial Intelligence and Multimedia Technologies can outperform Humans in such tasks.

In this talk, I will present ongoing work and results that try to cope with some common problems in the area of multimedia content annotation and management.







Paula Viana



FOTOINMOTION

*Repurposing and enriching images for immersive
video storytelling*



- H2020-ICT-Research and Innovation Action
- 8 Partners
 - academia, users, content owners, software integrators

WHAT ARE WE TRYING TO ACHIEVE?



- Turn **static images** - photographs - **into** rich, appealing and engaging multimedia **stories**
- Make photos searchable

HOW?

- By developing tools that can identify the most relevant aspects of the photo so that this information can be used to enable automatic creation mechanisms

WHAT INFORMATION CAN WE PROVIDE?

- Contextual Information collected together with the photo
 - Location, sound classification, activity, voice recognition, weather, ...
- Object Identification
- People Recognition
- Perceptually important areas



FOTO
IN
MOTION

HOW?

- Computer Vision
- Machine Learning
- Crowdsourcing

OVERALL APPROACH



- Analysis focused on use cases
 - Fashion: Genre, Clothing (*Jacket, Bottoms, Tops, ...*), Accessories (*Necklace, Purse and Bags, ...*)
 - Photojournalism: Protest rallies, Flags, Personalities
 - Festivals: Personalities, Logos
- Using different **Machine Learning** models optimised for each case
- Optimising **Datasets** for training
- Fusing information to filter erroneous detection or to adapt to the scenario

MACHINE LEARNING

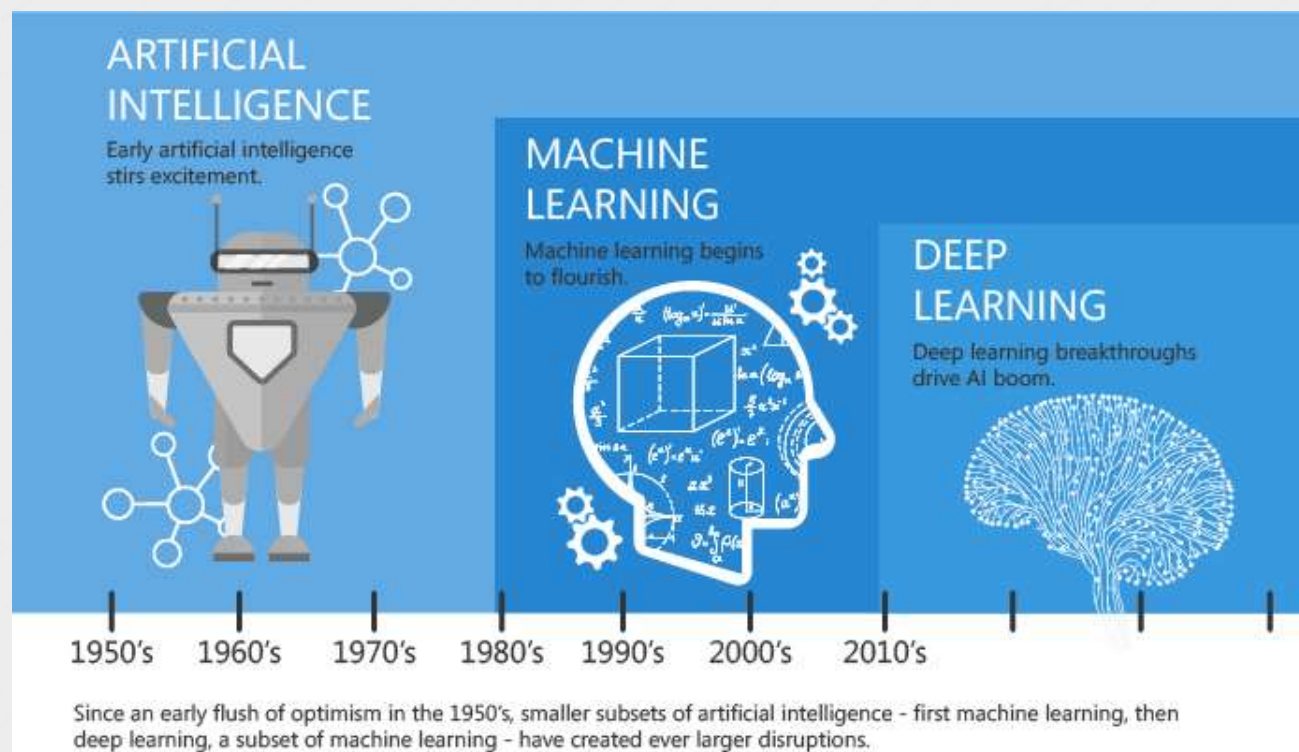


Image: Linked In | Machine Learning vs Deep learning

MACHINE LEARNING AIM

- Enable the computers to learn from data without being explicitly programmed

HOW

- Building algorithms that can receive input data and use statistical analysis to predict an output
- Outputs are updated when new data becomes available

(MAIN) CLASSES OF MACHINE LEARNING

- Supervised Learning
 - Requires labelled data to be input (ground-truth)
- Unsupervised Learning
 - No need for labelled data

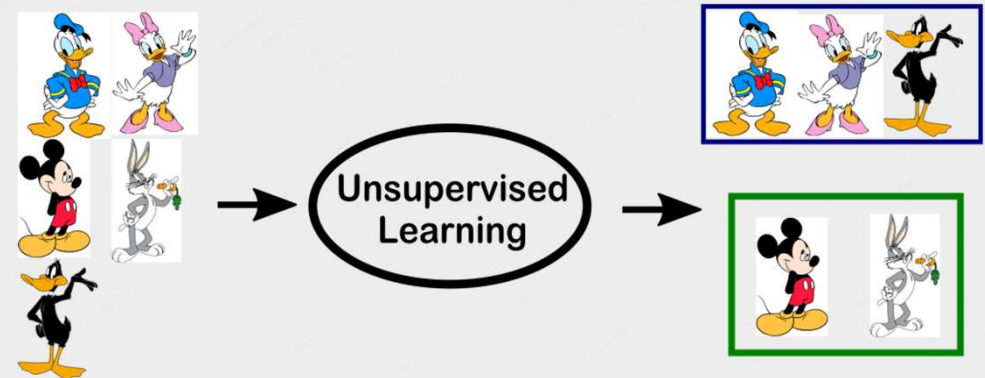
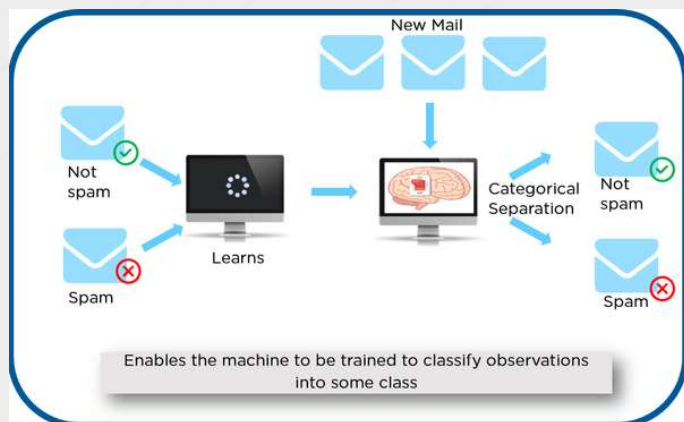


Image: Ayush Pant, Machine Learning for Beginners

SUPERVISED LEARNING

→ Classification

- given an input, classify (output) in one of the available classes (categories): “man”, “woman”, “kid”

→ Regression

- map input to a continuous output (the output variable is a real value): weather forecast, estimate life expectancy, etc.

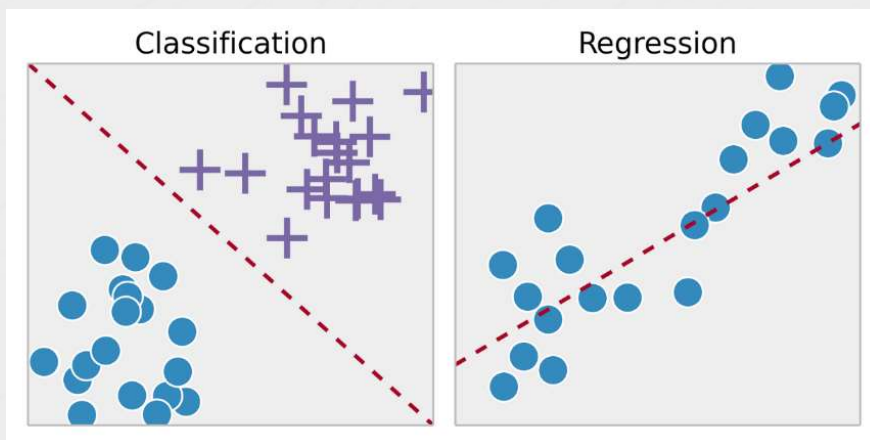
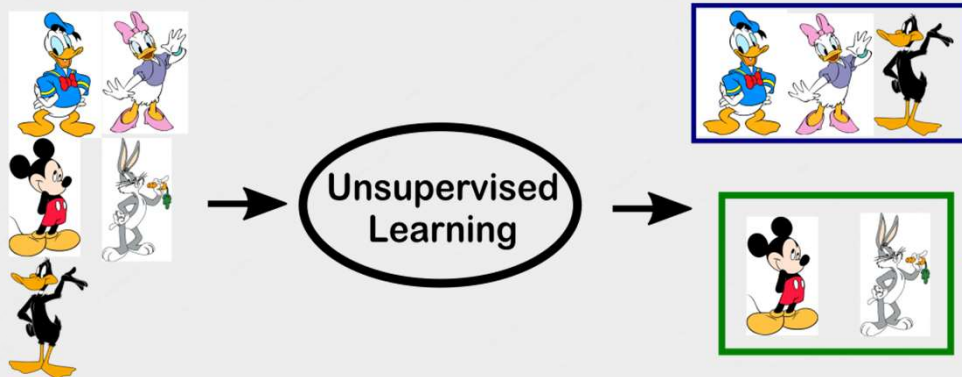


Image: Devin Soni, Supervised vs. Unsupervised Learning

UNSUPERVISED LEARNING

→ Clustering

- try to find similarities/patterns within the data to group entities: “ducks” and “not ducks”



→ Dimensionality Reduction

- try to learn relationships between individual features, and represent data in a more compact way (Representation Learning)

SUPERVISED LEARNING: CHALLENGES

- Datasets for training (ground-truth)
 - Costy (manually annotating/labelling)

- Complexity of the model
 - Computational costs
 - Overfitting vs learning the full structure of the data
 - Small amount of data -> simple model
 - Large amount of data -> more complex models

OVERFITTING

- Responds very well to the training data
 - *If we input an image used in the training phase, it will get the solution right*
- Does not generalize to other (slightly different) data
 - *The model is not really learning the actual structure in the data that leads to a given output*
 - *It will fail if we input a somehow different image*

LEARNING ALWAYS FROM SCRATCH?

- Humans Learning
 - don't learn everything from the "beginning"
 - use previous knowledge to learn new thinks

- Machine Learning
 - use this same paradigm to reduce the resources needed for implementing a new task
 - utilize knowledge acquired for one task to solve related ones

Transfer Learning

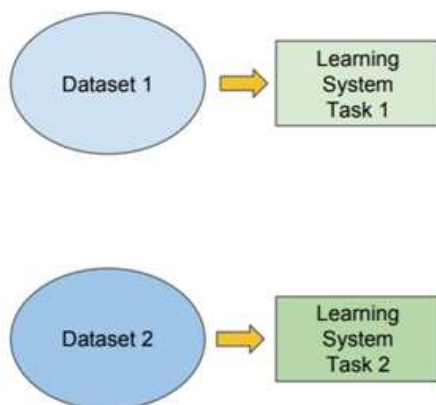
TRANSFER LEARNING

Traditional ML

vs

Transfer Learning

- Isolated, single task learning:
 - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks



- Learning of a new tasks relies on the previous learned tasks:
 - Learning process can be faster, more accurate and/or need less training data

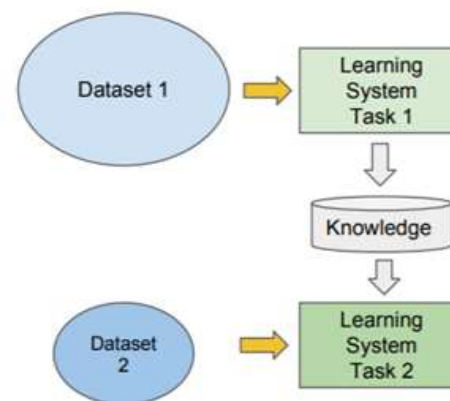


Image: Dipanjan (DJ) Sarkar | Hands-on Guide to TL

DEEP LEARNING: PARADIGM SHIFT

Traditional Pattern Recognition: Fixed/Handcrafted Feature Extractor



Feature
Extractor

Trainable
Classifier

Automatically
learn
hierarchical
features
directly from
data

Deep Learning: Representations are hierarchical and trained



Low-Level
Features

Mid-Level
Features

High-Level
Features

Trainable
Classifier

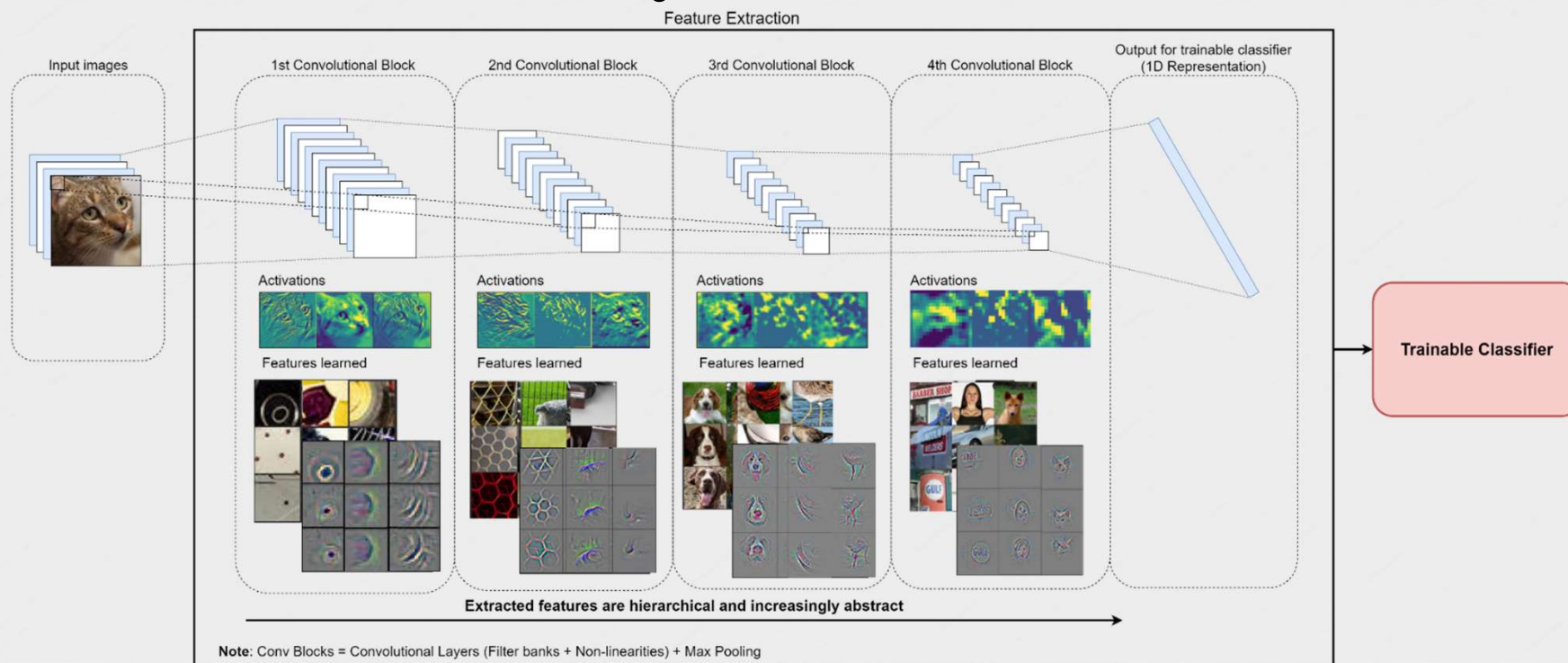
Internal Hierarchical Representation

DEEP LEARNING: FEATURE EXTRACTION

Low-level features
e.g edges

Mid-level features
e.g object parts, combined
edges

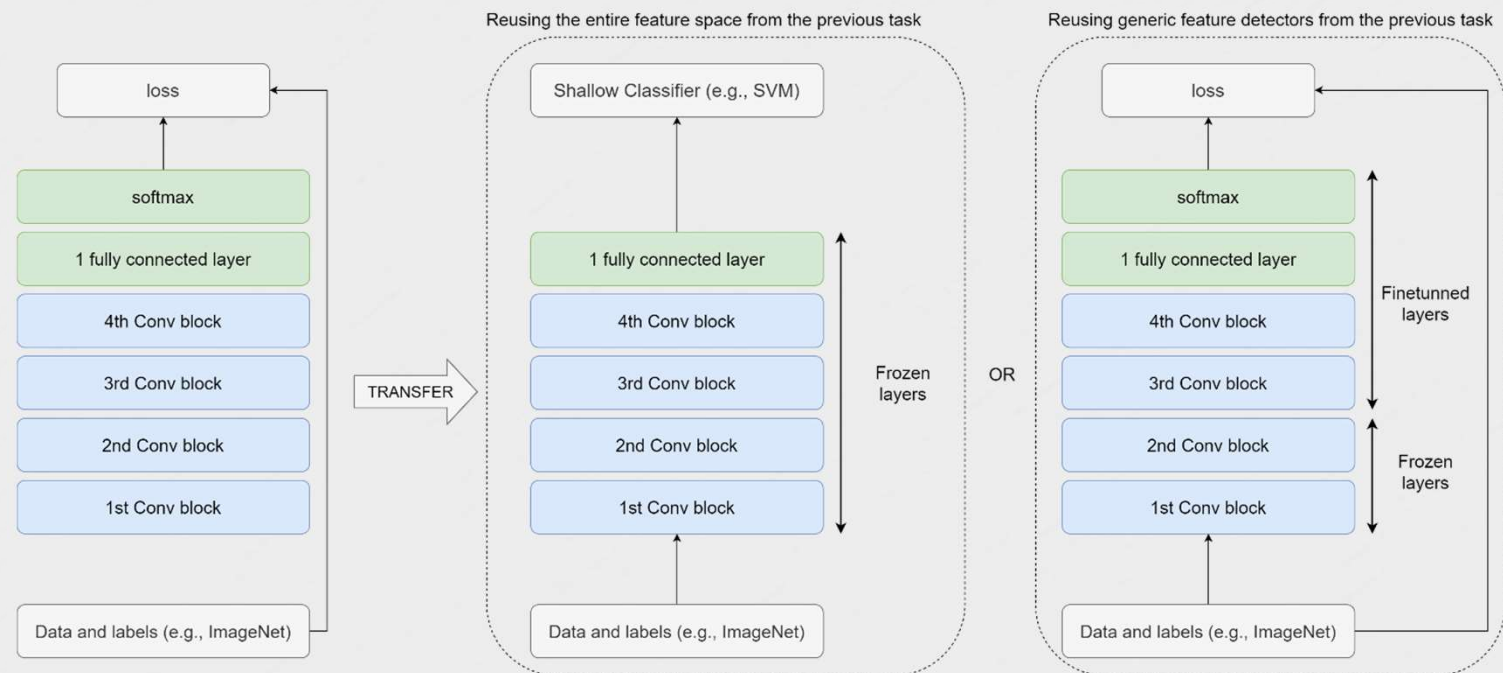
High-level features
e.g object models



TRANSFER LEARNING IN DEEP LEARNING

→ What is considered knowledge from previous tasks?

Features that can be used as a baseline to build representations for both tasks (generic feature detectors) – Simple features such as Sobel Filters which are extracted in the first layers of CNNs



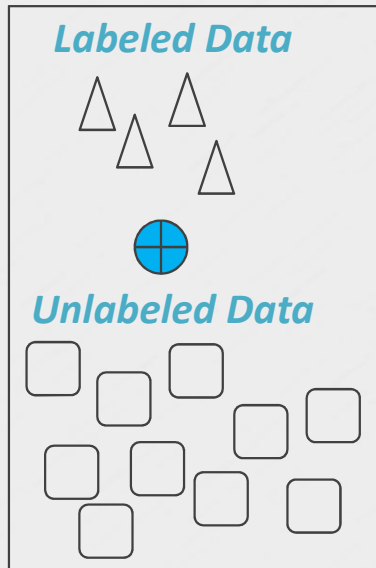
Note: Frozen layers - not updated during backpropagation (optimization); Finetuned layers - Updated during backpropagation (optimization).

Image: Dipanjan (DJ) Sarkar | Hands-on Guide to TL

SUPERVISED + UNSUPERVISED LEARNING

COMBINING THE BEST FROM TWO WORLDS

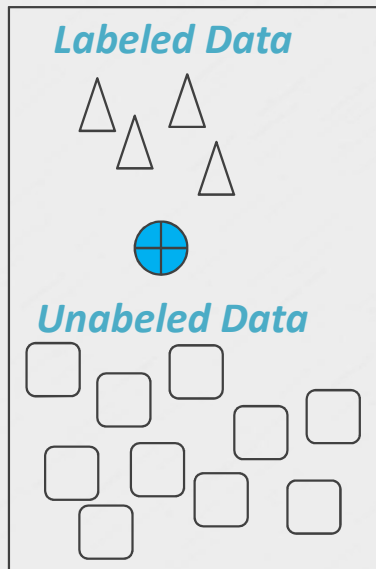
HOW?



- Use small amount of labeled data + a large amount of unlabeled data
 - Avoids the challenges of finding a large amount of labeled data

SEMI-SUPERVISED LEARNING

SEMI- SUPERVISED LEARNING



- Use the labeled data to (partially) train a model
- Use this (partially) trained model and label the unlabeled data, creating 'pseudo-labelled' data.
- Combine labelled and pseudo-labelled datasets are combined, creating a unique algorithm that combines both the descriptive and predictive aspects of supervised and unsupervised learning.

SOME APPROACHES



- Pick some pre-existing models that do not cope with all the requirements
- Use Transfer Learning for new classes of objects
 - Without requiring large datasets
 - Without requiring large computational resources
- Use a semi-supervised + power of the crowd
 - Label datasets
 - Improve the quality of the datasets

AND WHAT ABOUT USING THE POWER OF THE CROWD TO HELP?



This can be used for two purposes


1. Use crowdsourcing approaches to reduce the AI mistakes
 - In a 1st step let the computers do their job
 - In a 2nd step let the humans correct the computers
2. Use AI to reduce content annotation time by providing a first clue
 - ML approaches require a lot of data for training
 - This requires a tremendous human effort to create the ground-truth
 - Then, let AI work even if not in good shape and then correct AI

HOW DO WE COMPARE?

Google Vision API

Clarifai



Labels	Web	Properties	Safe Search
			
		Necklace	85%
		Jewellery	83%
		Fashion Accessory	81%
		Neck	60%
		Magenta	53%
		Glasses	52%

General	VIEW DOCS
PREDICTED CONCEPT	PROBABILITY
portrait	0.982
fashion	0.974
people	0.966
necklace	0.955
woman	0.953
jewelry	0.928
isolated	0.919
man	0.912
sunglasses	0.904
eyeglasses	0.902
smile	0.869
glamour	0.855

Not region based

Not the required classes

Require upload (security!!)

Only objects

THREE COMPUTER VISION TASKS

→ Image Classification

- Input: an image
- Output: class labels



CAT

→ Object Localization

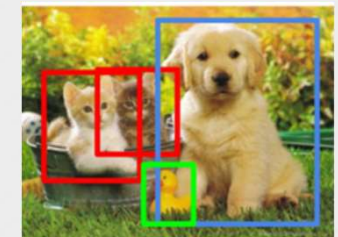
- Input: an image
- Output: one or more bounding boxes



→ Object Recognition

- Input: an image
- Output: one or more bounding boxes and a class label for each

DOG CAT DUCK



FEATURE EXTRACTION

- Image Classification
 - Generate image features of the full image

- Object Recognition
 - Generate image features on a more fine-grained, granular, regional level of the image

R-CNN: REGION BASED CONVOLUTION NEURAL NETWORKS

- 1st Step: Region Proposal
 - Generate category independent region proposals (candidate bounding boxes)
- 2nd Step: Feature Extractor
 - Extract feature from each candidate region, e.g. using a deep convolutional neural network
- 3rd Step: Classifier
 - Classify features as one of the known class (e.g. SVM classifier)

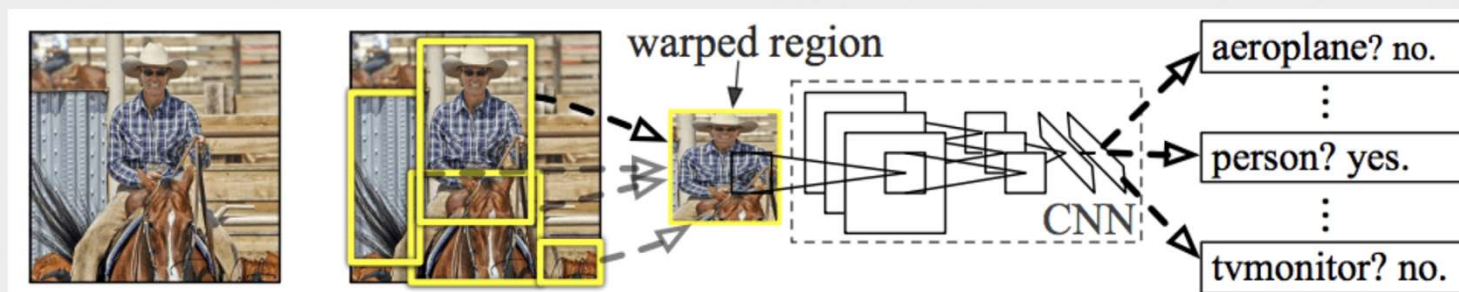


Image: A Gentle Introduction to Object Recognition With Deep Learning

R-CNN AND FAST R-CNN

→ R-CNN drawbacks

- (Low) Speed of training and prediction
 - Training a CNN on a lot of region proposals per image
 - Make predictions using a CNN on a lot of region proposals

→ Fast R-CNN and Faster R-CNN

- 1st Step: Region Proposal Network.
 - CNN for proposing regions and the type of object to consider in the region.
- 2nd Step: Fast R-CNN
 - CNN for extracting features from the proposed regions and outputting the bounding box and class labels.

METHODOLOGY



- SoA NN Architectures
 - Faster R-CNN Resnet 101
 - Faster R-CNN Inception-Resnet v2
 - YOLO V2
- Relevant datasets
 - COCO (Common objects in Context)
 - Open Images
 - LFW (Labeled faces in the wild)
 - Pascal VOC
 - Imagenet

METHODOLOGY



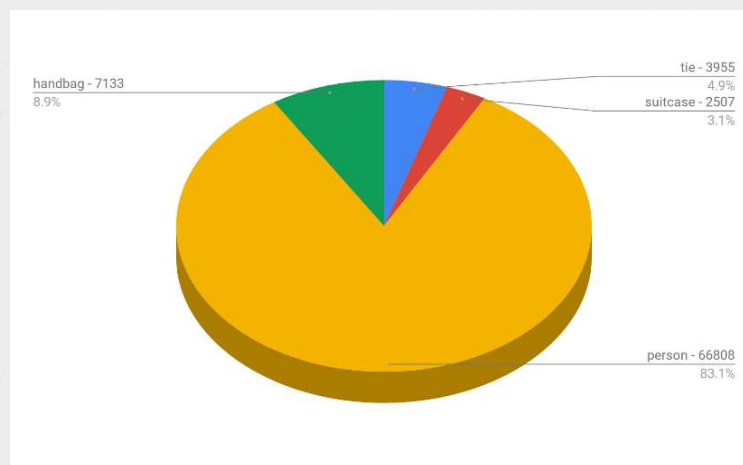
- Baseline
 - Faster R-CNN Resnet 101 w/ COCO
 - Faster R-CNN Inception-Resnet v2 w/ Open Images
- Application of Transfer Learning to optimise the classifier for
 - existing classes
 - new use-case oriented classes

BASELINE DATASETS CHARACTERIZATION

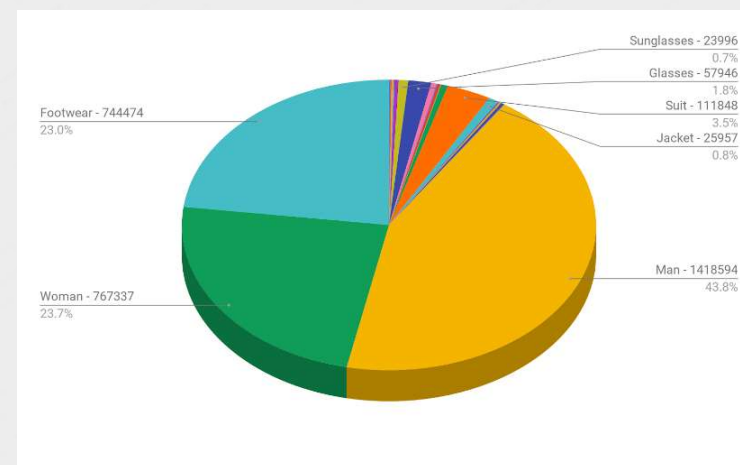
FASHION USECASE



COCO dataset



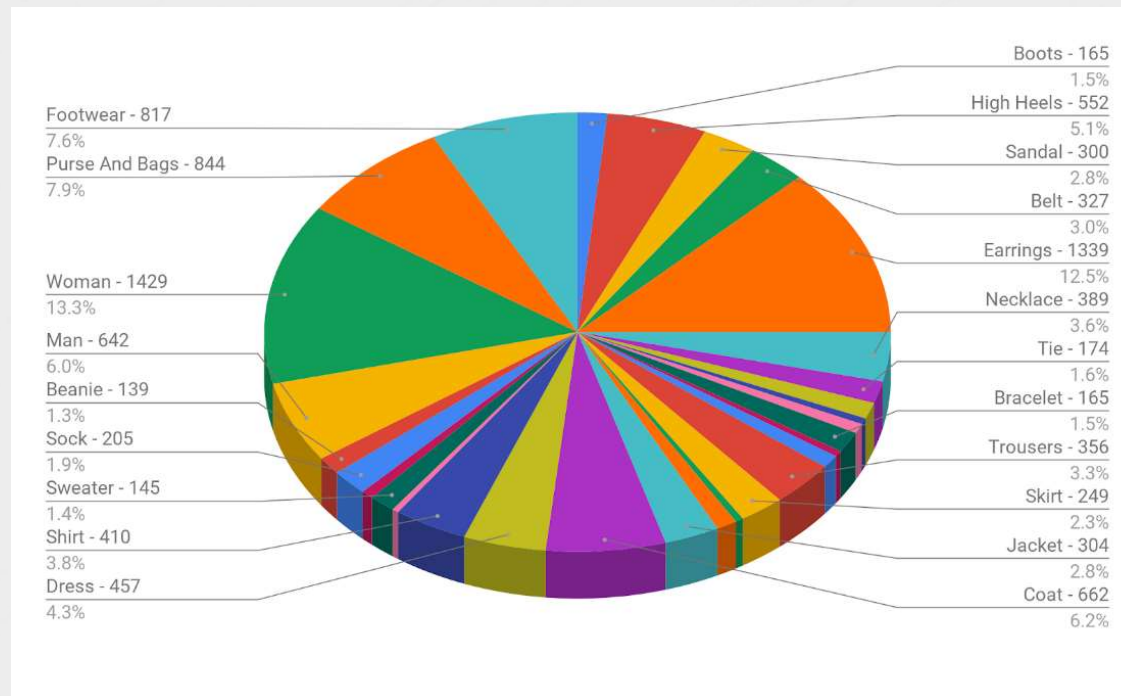
Open Images dataset



FIM DATASET CHARACTERIZATION

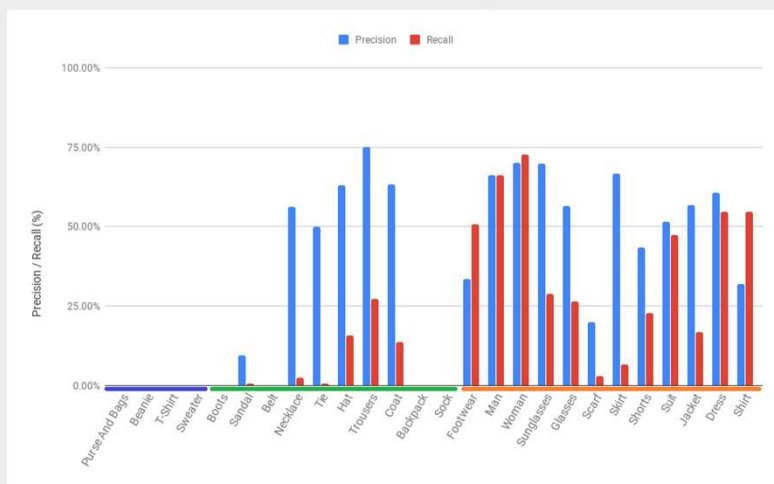
FASHION USECASE

FiM dataset

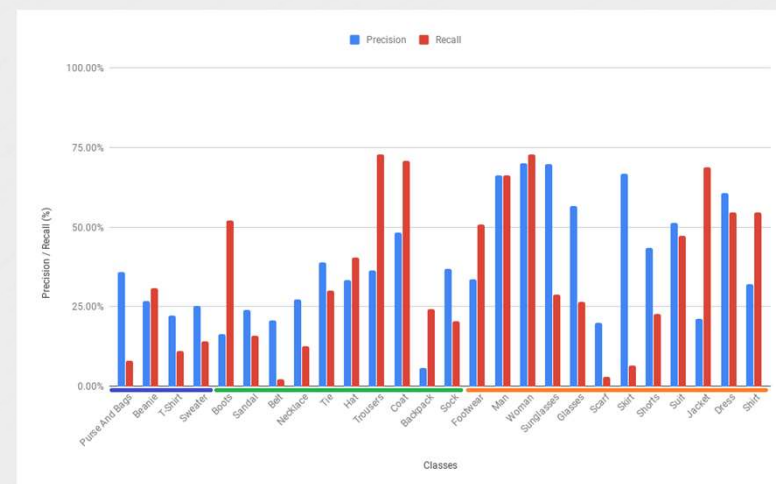


FASHION USECASE RESULTS

FOTO
IN
MOTION



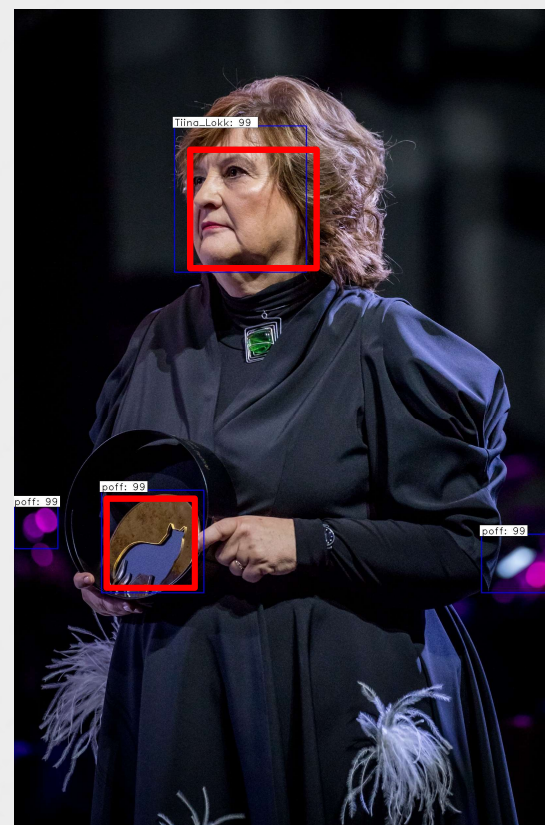
Before



After

SOME (VISUAL) RESULTS

FOTO
IN
MOTION





Paula Viana

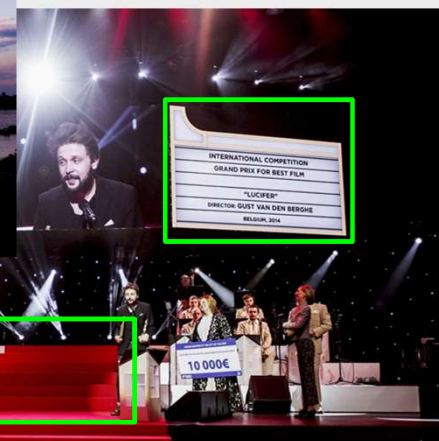
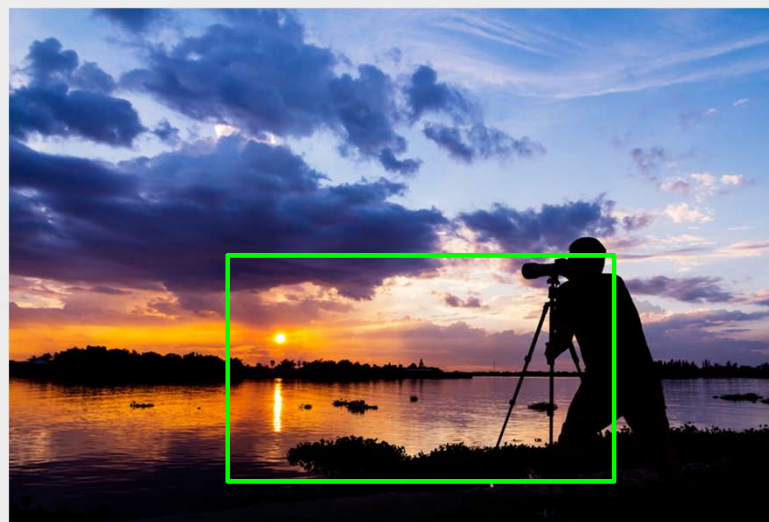


Paula Viana

AND IF THE SYSTEM FAILS?

- *No known objects*
- *New application scenarios*

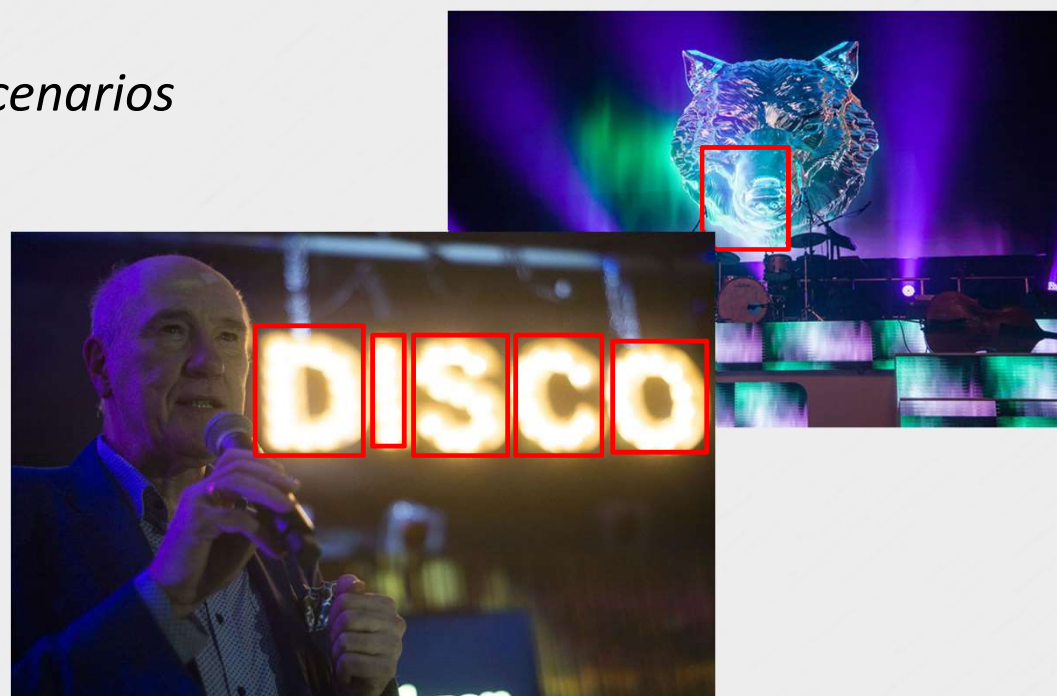
FOTO
IN
MOTION



SALIENCY MAPS CAN HELP

AND IF THE SYSTEM FAILS?

- *No known objects*
- *New application scenarios*



SPOTLIGHTS CAN BE RELEVANT



AWARD

EUROPEAN COMMISSION INNOVATION RADAR

USE OF AUTOMATED ANNOTATION IN IMAGES

CHIC

Cooperative View On Internet and Content



COOPERATIVE VIEW ON INTERNET AND CONTENT

- P2020 “Projecto Mobilizador”
- 28 Partners
 - academia, users, content owners, software integrators

CHIC: *A.3. Ecosystem of production and distribution of television, centered on the consumer, in a cloud environment with decentralized contributors*



COOPERATIVE VIEW ON INTERNET AND CONTENT

Main Partner: Porto Canal

Main Objective: exploit new approaches for timed annotation of content

WHAT ARE WE TRYING TO ACHIEVE?

- Make TV archives searchable
- Enable re-purposing of content

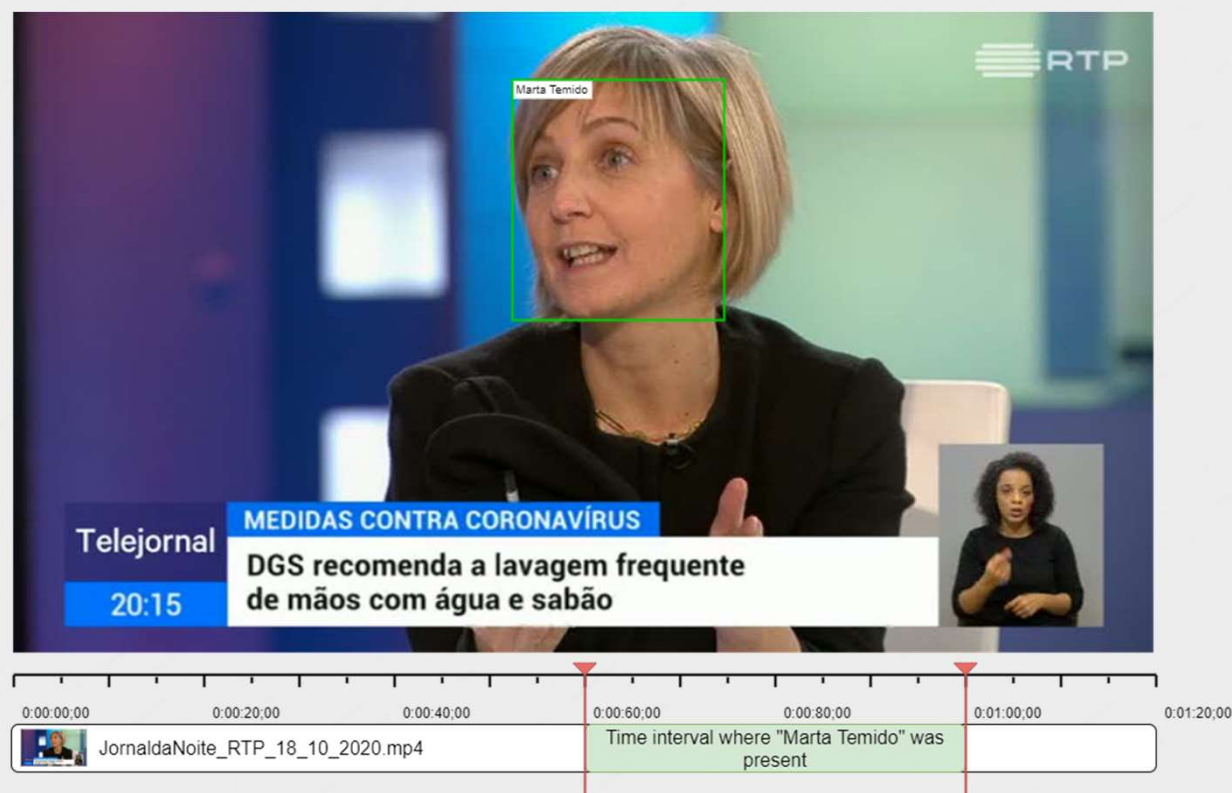


HOW?

COOPERATIVE VIEW ON INTERNET AND CONTENT

- Develop tools that enable timecoded video annotation
- 1st step: personality detection
- Make the process easy to adapt
 - Computational cost
 - Data

WHAT WE AIM



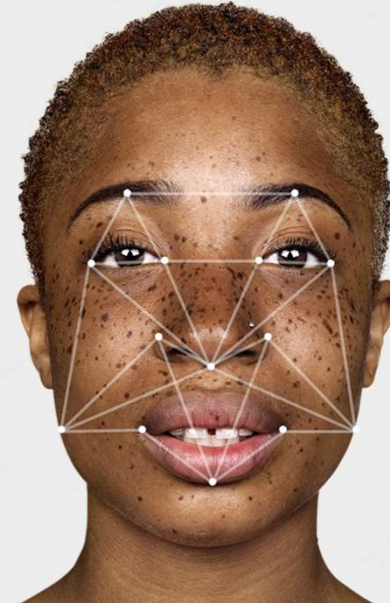


OVERALL APPROACH

- Enhance the quality of the (training) dataset
 - Choose the best images (increase diversity)

HOW?

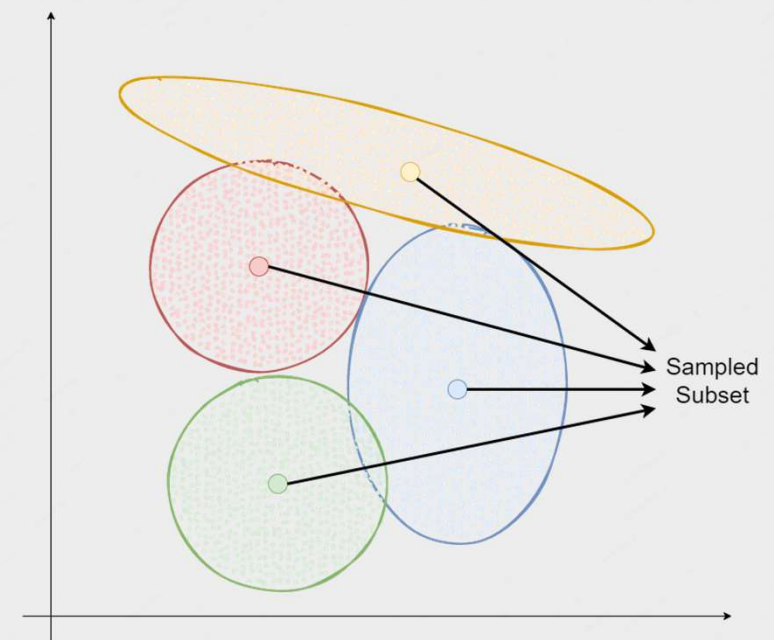
- A lot of facial metrics proposed in literature
 - Make the problem easier by selecting the best metrics (dimensionality reduction)
 - Make sure the approach is universal (genre, age, ethnicity, ...)
- Use some post-processing to enhance the results
 - One image - > One person
 - We are dealing with video content -> Tracking



HOW?

- Collect facial metrics for each image in the training set
- Group them by similarity
- Build a reduced dataset with the least number of images that have the most amount of information

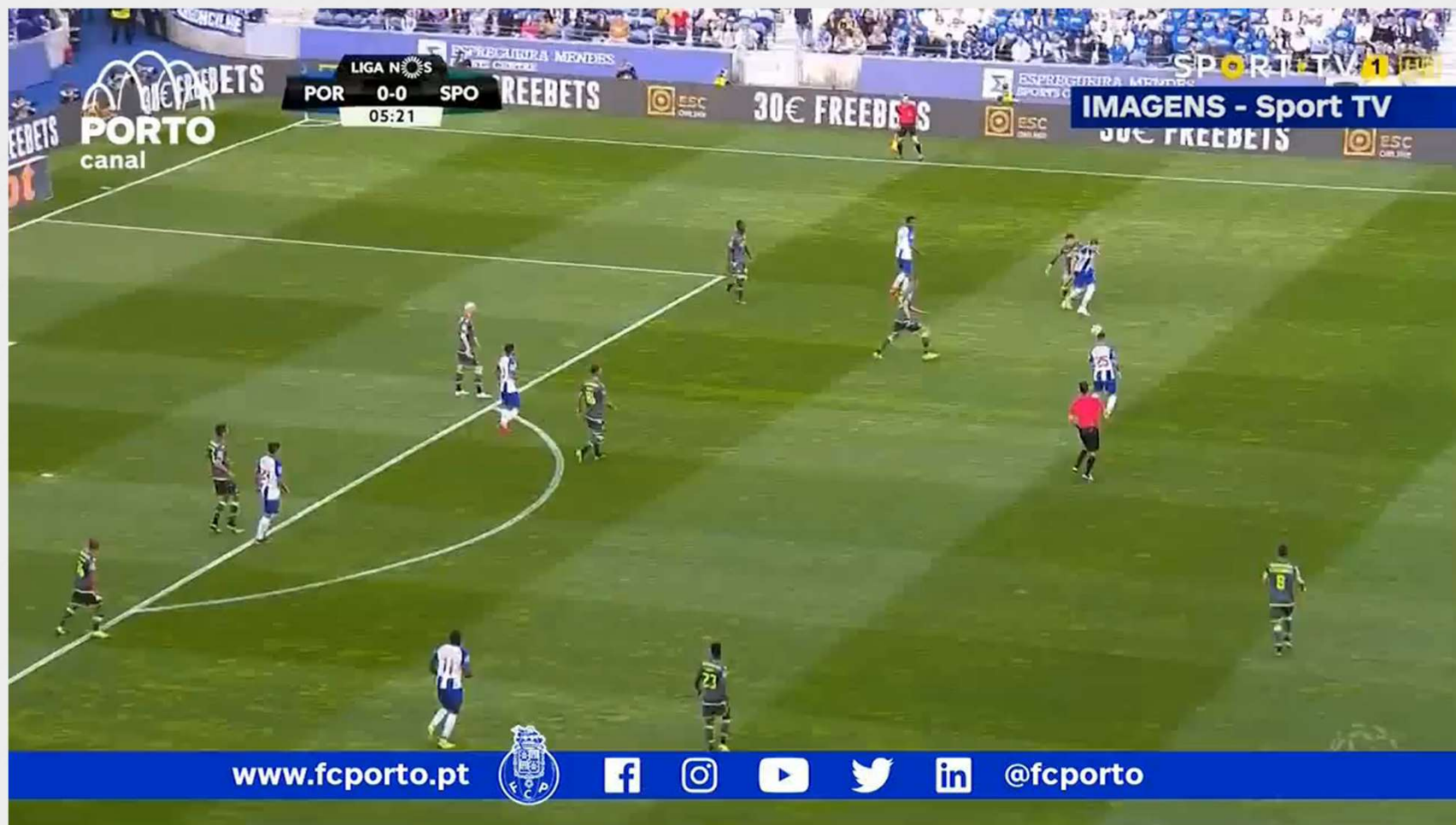
From each cluster select only the image that is closer to the cluster centroid



WHAT ABOUT OTHER APPLICATION AREAS WHERE CONTENT ANNOTATION MAY HELP HUMANS?

*SPORTS, ADVERTISING, SOCIAL
STUDIES, CINEMA... AND SO ON*

- Event Detection
- Advertising Impact
- Person Impact
- Emotion Detection
- Content Relations



Thanks!

For being patient...

Paula Viana
P. Porto & INESC TEC
pmv@isep.ipp.pt