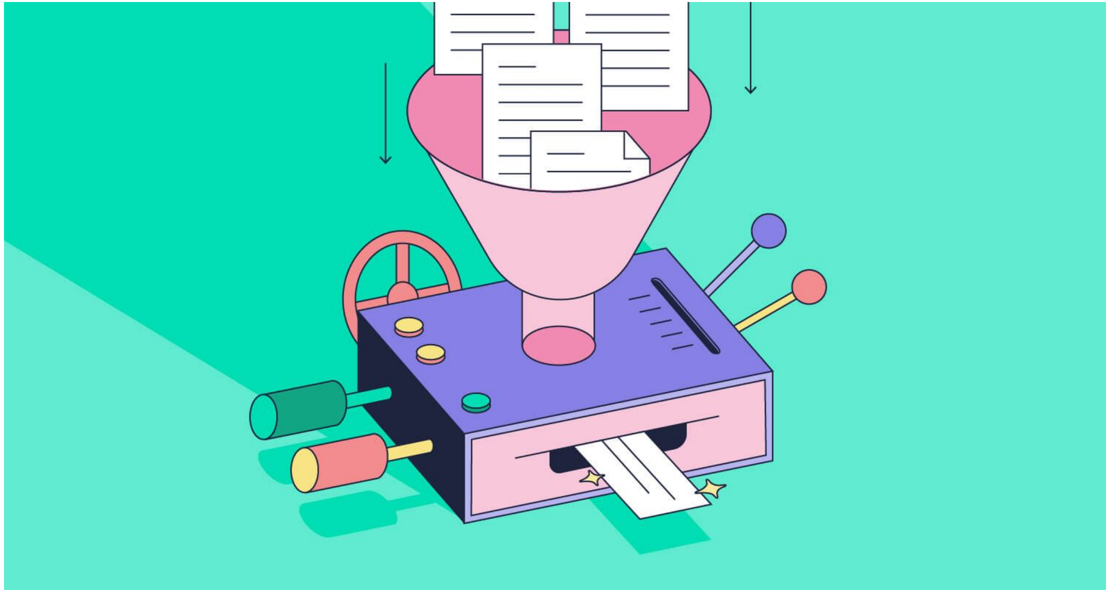
Final Presentation:
Visual Recognition

Group 1:
Carles Pregonas
Marc Pérez
Pau Vallespí
Carlos Boned

Index

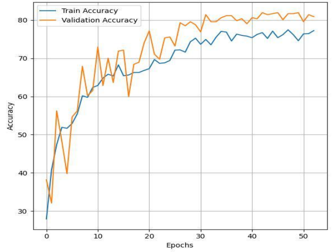
1. Week’s summary
   a. Week 1: Image Classification using PyTorch
   b. Week 2: Object Detection, Recognition and Segmentation
   c. Week 3: Image Classification
   d. Week 4: Cross-modal Retrieval
   e. Week 5: Diffusion models
2. Multimodal Human Analysis

WEEK'S SUMMARY



Week 1: Image Classification using PyTorch

Aspect	PYTORCH	TensorFlow
Developed by	Facebook (smaller community)	Google (larger community)
Graph Computation	Dynamic computation graph (changes on the fly)	Static computation graph
Test Accuracy	82.03%	80.93%
Parameter initialization	glorot/xavier_uniform	kaiming_uniform



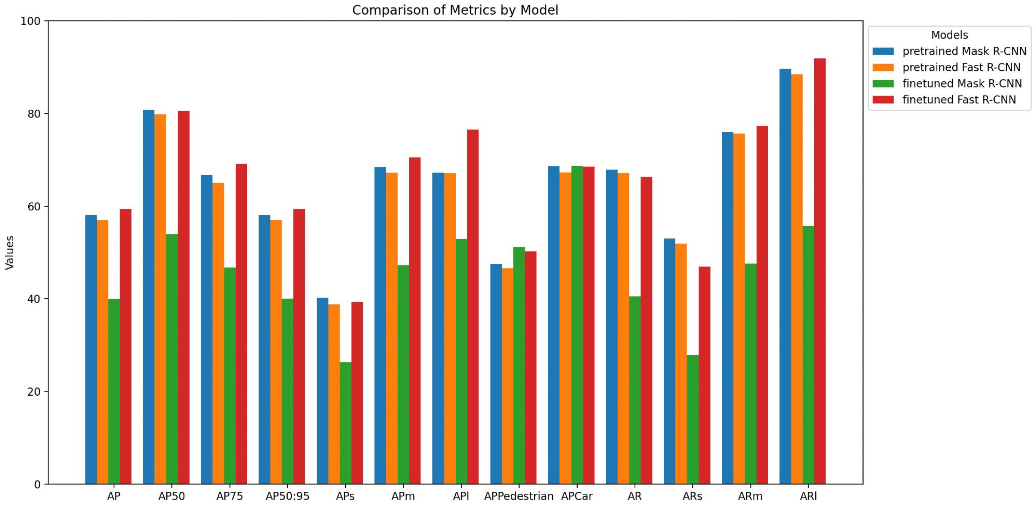
(1)

(2)

- Throughout the epochs, the updates in accuracy and loss appear to be more stable in PyTorch(1) compared to Keras(2).
- In our small study we have seen some **details about the decision making** of the model which in the future can be corrected, for example by forcing more general features boundaries.
- Additionally, we've demonstrated that certain **errors** may not be resolved due to data and its **labeling**.

Week 2: Object Detection, Recognition and Segmentation

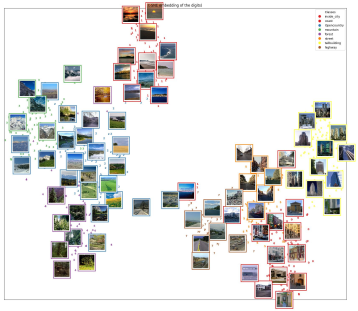
- The **detection threshold directly influences the accuracy** of the model in accepting misclassification errors or excluding correct detections of poorly visible elements.
- Cars and persons may be more representative on traffic environments, so **pre-trained models fail more when detecting other classes** on KITTI-MOTS data.
- After **fine-tuning** the pre-trained models, results only **improved for Faster R-CNN**, probably due to its minor complexity.
- **Small objects are hard to detect** due to its low spatial representation and quality resolution.
- **Context has a highly influence** in the confidence of detecting an object.



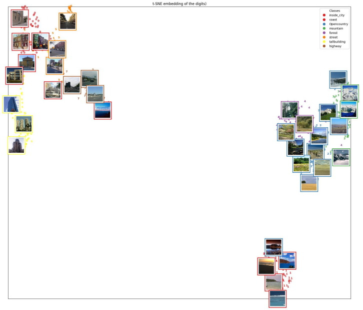
Week 3: Image Classification

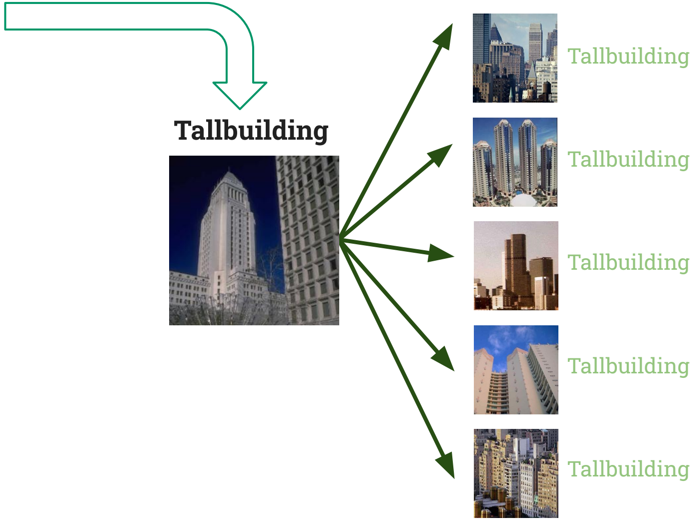
- **Resnet50** is a **good tool to perform image retrieval** when removing its last layer, since the features obtained are very helpful to get similar images from a query. Its feature space is **linearly separable**.
- <u>Metric Learning:</u>
  - **Siamese networks**, results have **improved** thanks to the use of the **shared weights** for training with a **contrastive loss**, what makes the model more **robustly trained**.
  - **Triplete networks**, we haven't been able to get better results, despite showing distinct class clusters in the learned metric space. Error analysis revealed persistent **misclassifications**, with classes **overlaps or semantic ambiguities** (need further refinement).
- The used **COCO dataset is <u>not</u> optimal for an image retrieval** task as it is intended for object detection.
- For COCO retrieval the KNN **has learnt to group parent labels** ("animals", "food", "sports", "vehicles", …) **and not individual objects**.

TSNE



UMAP

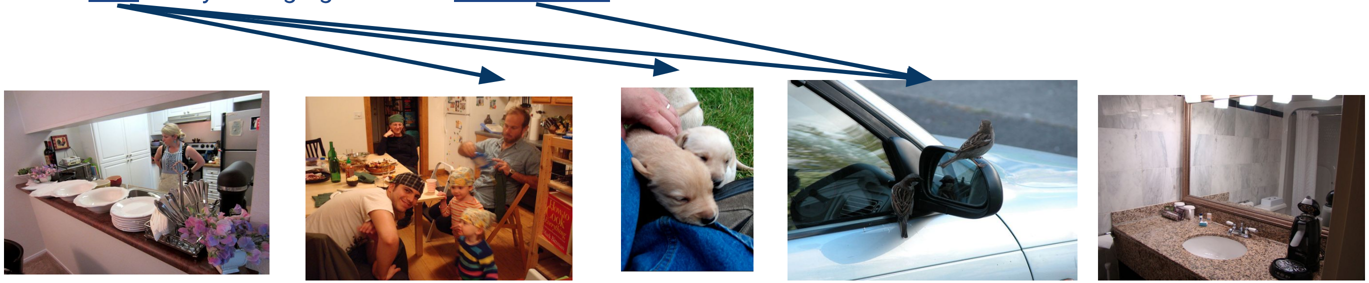




Week 4: Cross-modal Retrieval

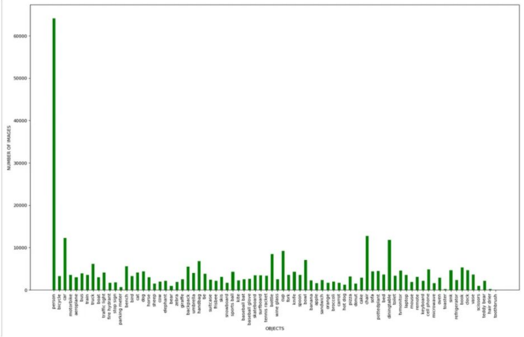
- We applied two strategies for selecting the Triplets:
  - **Online**: All possible triplets combinations are generated in each minibatch.
  - **Offline**: Selecting a negative image with a low cosine similarity between anchor-negatives captions.
- Due to the limited available resources we have decided to use the **25% of the training dataset**, which makes it **unbalanced** given that there are no labels on the dataset.
- On **Image-to-Text** task we have found different cases where:
  - The retrieved caption **is exactly the same.**
  - The retrieved caption **means the same but it is semantically different.**
  - The retrieved caption **does not means the same but is somehow related to the validation data.**
  - The retrieved caption **has no coherence with the input data.**
- On **Text-to-Image** task we have obtained also some incoherent results, but others that make sense, with **images containing an element from a word or set of words** in the caption query.
- We have seen a **better qualitative results when working with BERT** than with FastText

Two husky's hanging out of the car windows.



Week 5: Diffusion models

- **Detecting less represented objects in our training subset.** We have analyzed our subset data distribution extracting the nouns and verbs from the captions. Then, we have checked for **unbalanced objects compared to COCO's distribution** and **generate captions using these objects' nouns with their most common verbs**, to get **a similar data distribution than COCO dataset**.



- To generate new images we have tested different models, choosing "**stable-diffusion-xl-base-1.0**". New generated images have been appended to the training set we already had, and we trained again the network.

Positive: "Amidst the rubble, a lone brick stands as a reminder of the building that once stood here."

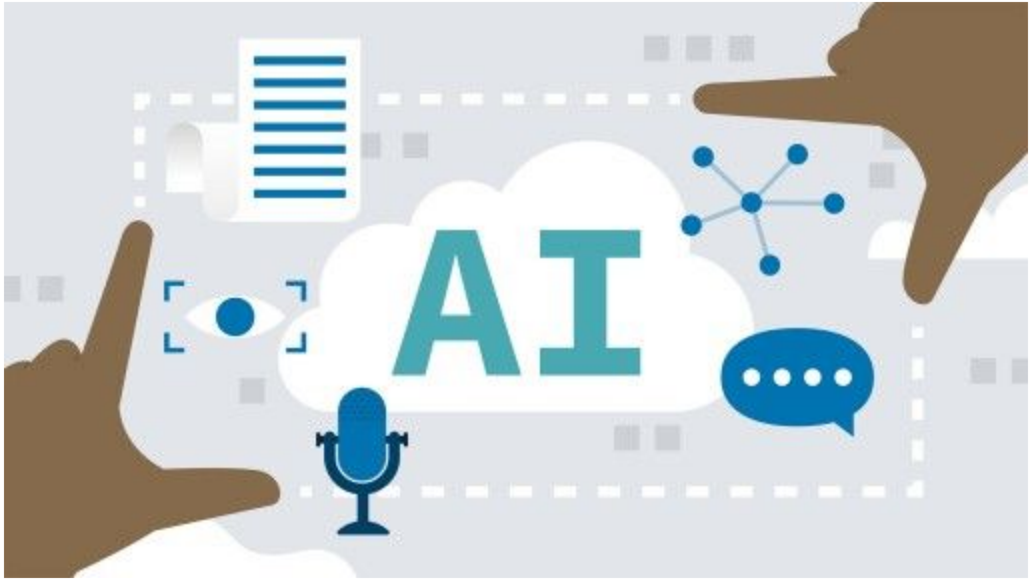
Negative: "A dog is sitting patiently beside its owner, waiting for a command."

- **Results have not been much more satisfactory than last week.** However, with our further analysis, we are closer to solve this task.

'An airplane is doing tricks and emitting smoke.'

'A lot of spectators watch a motorcade on Washington D.C.'

MULTIMODAL HUMAN ANALYSIS



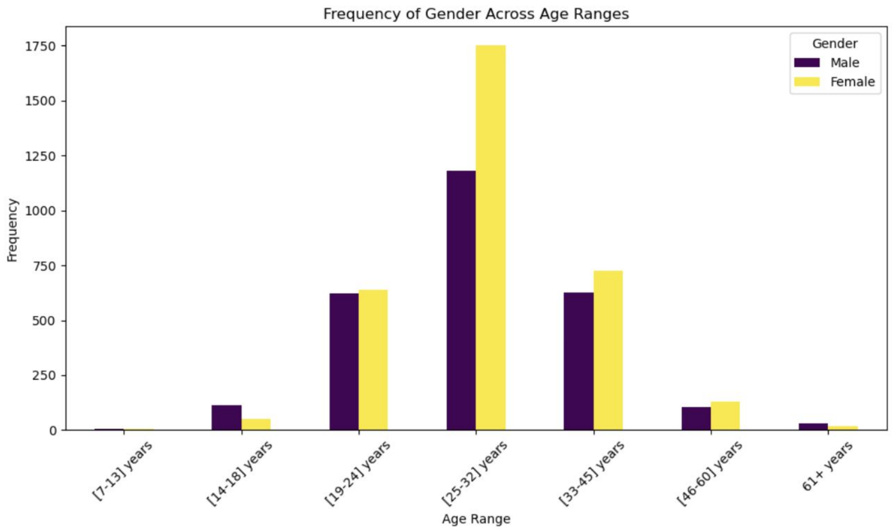
Introduction

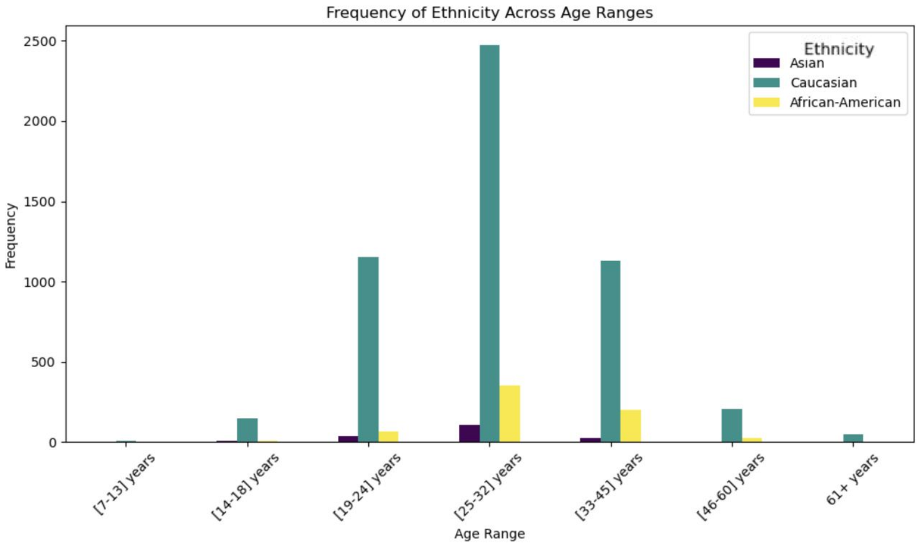
- During this week's project, we worked on **age classification** with images only and also using audios and text.
- We performed a **detailed dataset analysis** to see if data was unbalanced and we performed some strategies to fix the issues with the dataset.
- Developed an **image-only classifier** tailored to extract and utilize visual age indicators effectively. We trained it using three different strategies.
- Incorporated **acoustic** and **text** data representations, aiming to capture diverse age-related features from multiple data sources.
- Created of a **multimodal classifier** that combines data from visual, acoustic, and textual modalities.
- Rigorous **testing strategies** for both single-modality and multimodal classifiers to ensure comprehensive performance evaluation.

Dataset Study

Data distribution

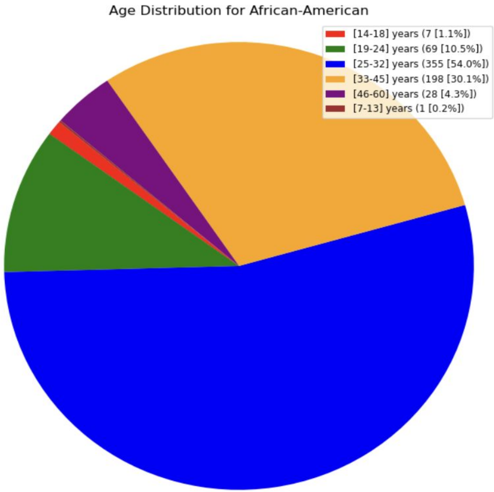
- Data is **heavily unbalanced**.
- Majority of samples belong to **Caucasian individuals** aged between **19 and 45 years**, with a higher representation of males compared to females.
- Absence of samples from **Asian and African-American** individuals across both young and old age groups.

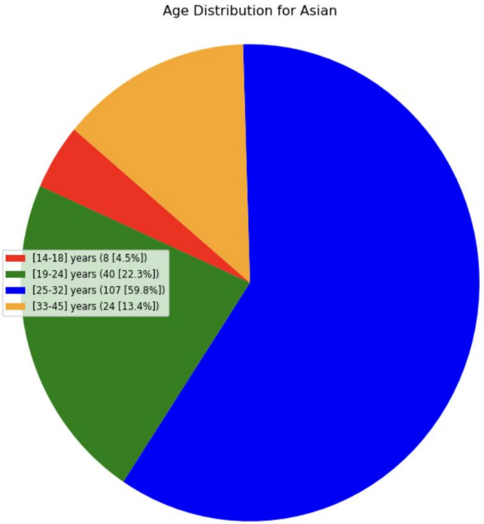


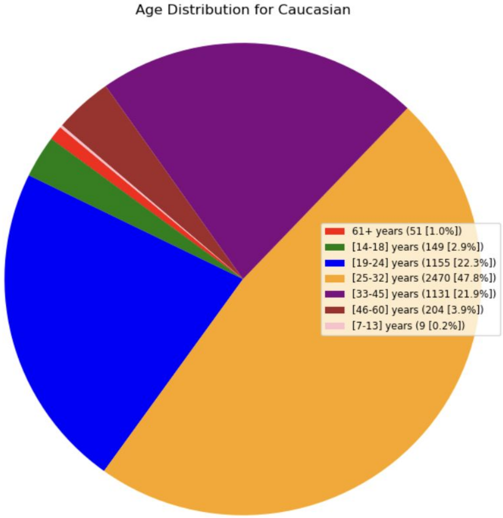


Dataset Study

Age distribution by Ethnicity



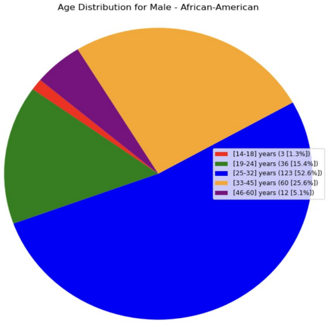


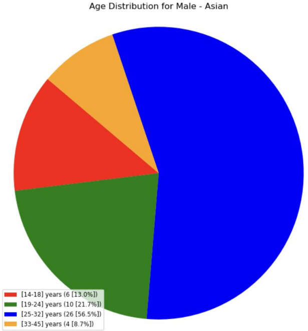


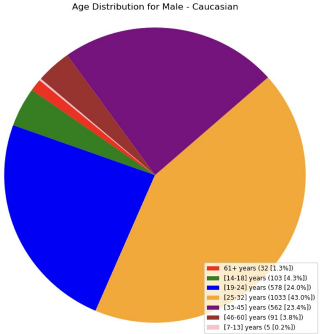
Dataset Study

Age distribution by Gender / Ethnicity

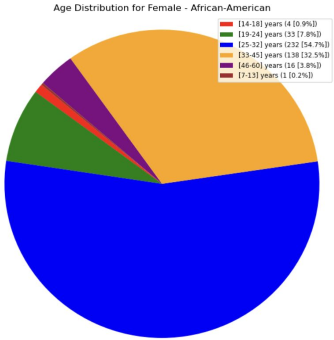
Female

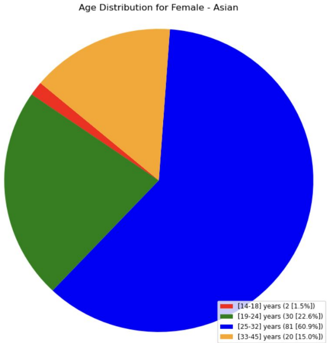


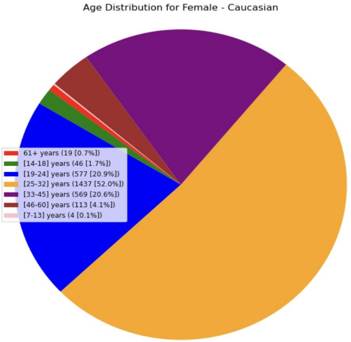




Male





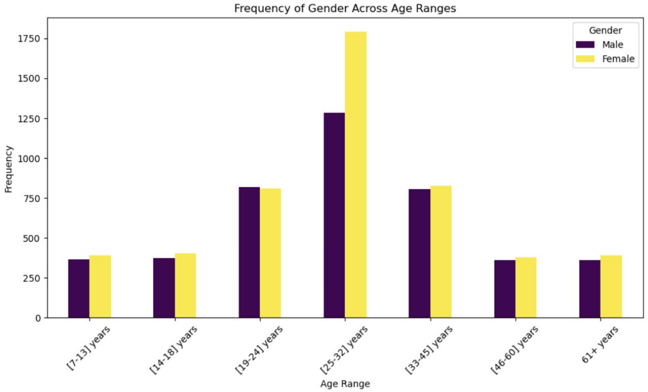


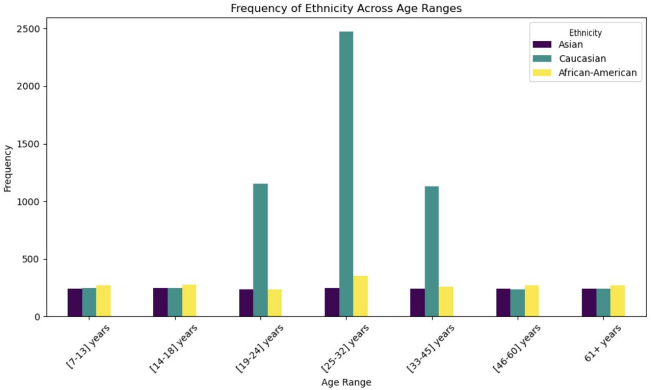
Dataset Study

Data augmentation: *Generation & Downsampling*



- We used *[stable diffusion XL]()* in order to generate images.
- **YOLOv5** as a face detector to crop face from images but **manual quality control** mechanisms and **further refinement** to address remaining outliers.





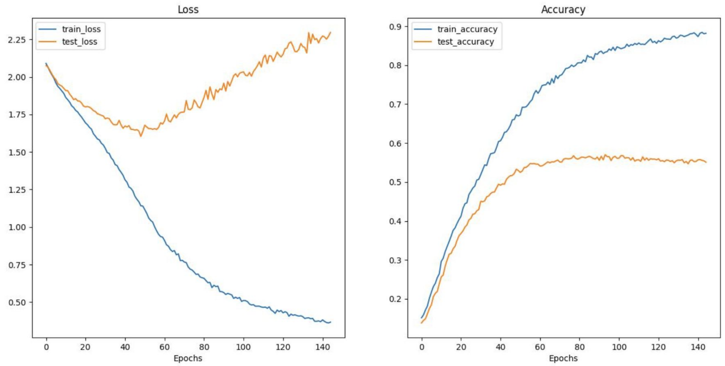
"22 years old african-american female"

"40 years old asian male"

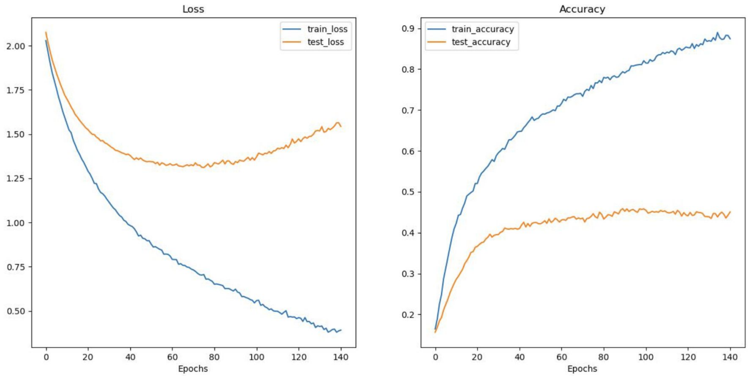
"65 years old asian female"

Image-Only Classifier

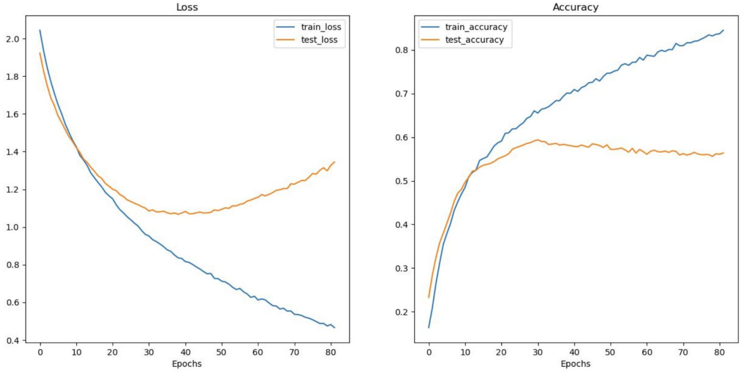
Comparison between classifiers



Baseline model



Using downsampled data



Using augmented data

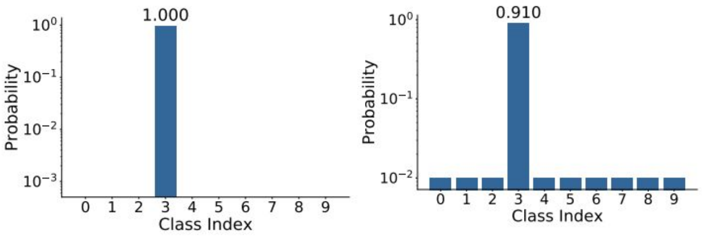
Training Strategy

- For training we treated each sample **independently**, without grouping equal user ID frames. The main reason was to keep with the complete training data set even if it could lead to **overfitting**, because of almost equal images.

- To add extra regularization we applied **label smoothing** introducing noise to the labels.

$$y_{\text{smoothed}} = (1 - \alpha) \times y + \frac{\alpha}{K}$$

- LS can be beneficial when dealing with **imbalanced datasets** or datasets with **class distribution biases**.



(a) Hard Label (b) LS

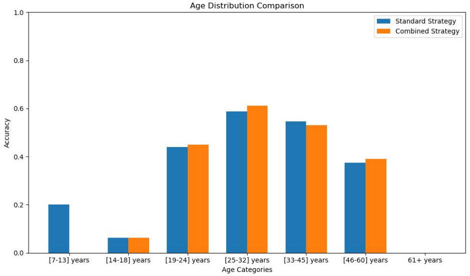
Test Strategy

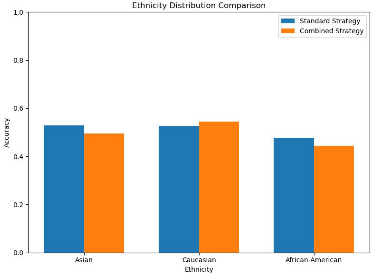
When testing our trained model we have followed two strategies:

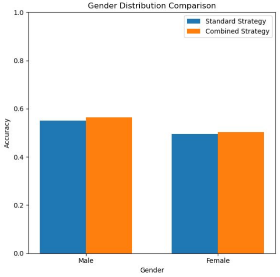
- **Standard Strategy:** Different samples from the same person **may have different** age predictions.
- **Combined Strategy:** Different samples from the same person **must have equal** age predictions.

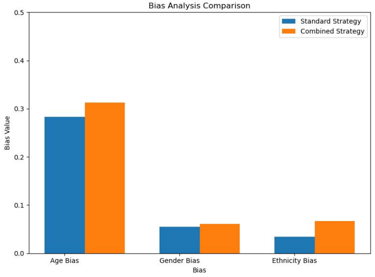
Test Evaluation

Baseline
52,10% / 53,16%

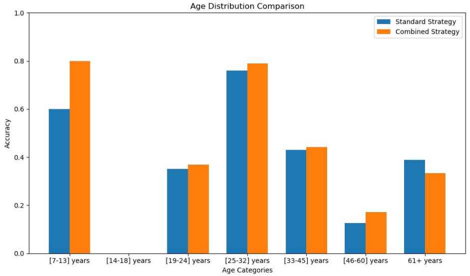


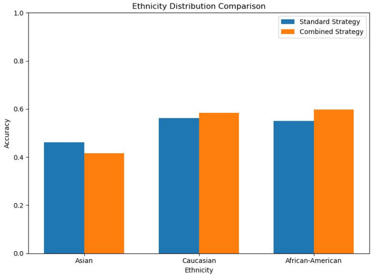


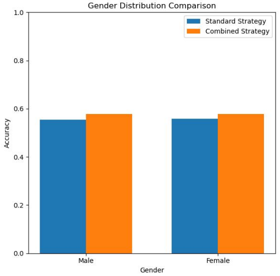


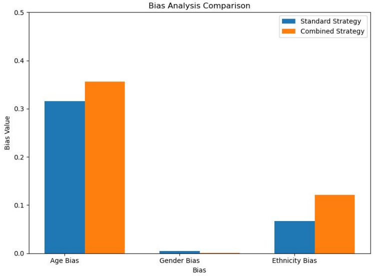


Data Augment.
55,59% / 57,82%

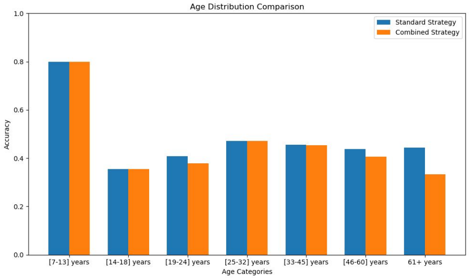


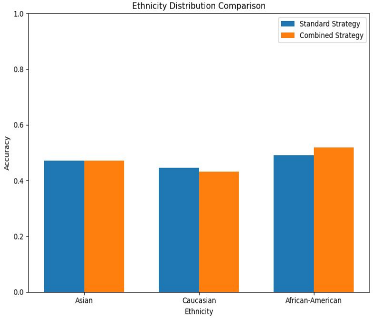


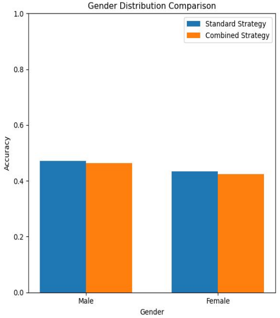


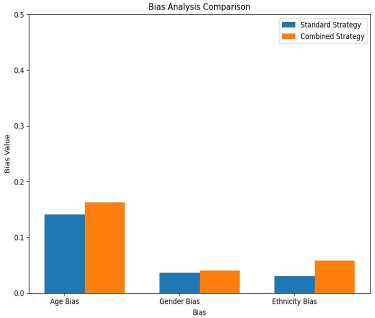


Downsampling
45,11% / 44,25%



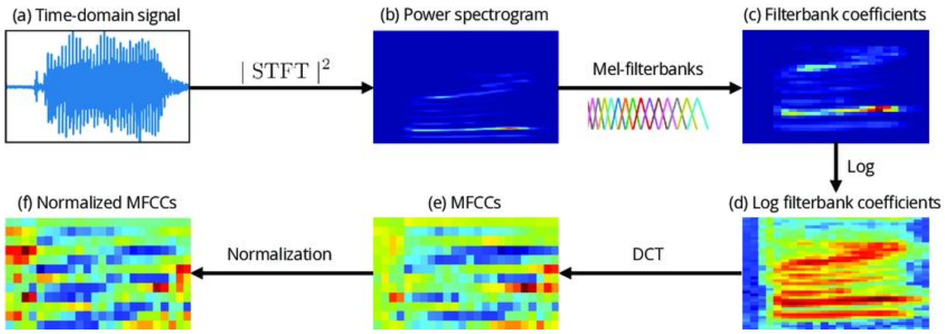




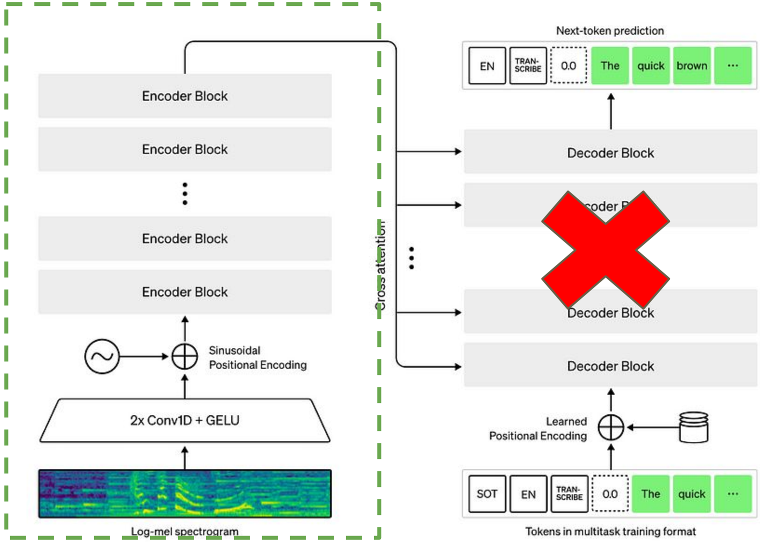


Acoustic Data Representation

- We employed the Librosa library to preprocess the audio file and compute Mel-frequency cepstral coefficients (**MFCCs**).



- Using a pre-trained model called **Whisper**, which is specifically designed for audio feature extraction. It loads the audio file,



Text Data Representation

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

We have used the **BERT** "AutoModel" from Hugging Face's Transformers.

We have selected **BERT** because its embeddings capture rich contextual information, enabling better understanding of text semantics. Moreover, we got a better performance on previous tasks than with other text embeddings methods.

Text Data Representation

Handling inconsistency

However, the transcriptions are pretty inconsistent. We tried to handle this extracting the transcriptions with the whisper model from hugging face.

GroundTruth:

I'm thinking with how much time and energy I'm spending on doing my [inaudible 00:00:04] play, I'm probably going to end up really liking to do [inaudible 00:00:11], which is from Assassin's Creed, because I really ...

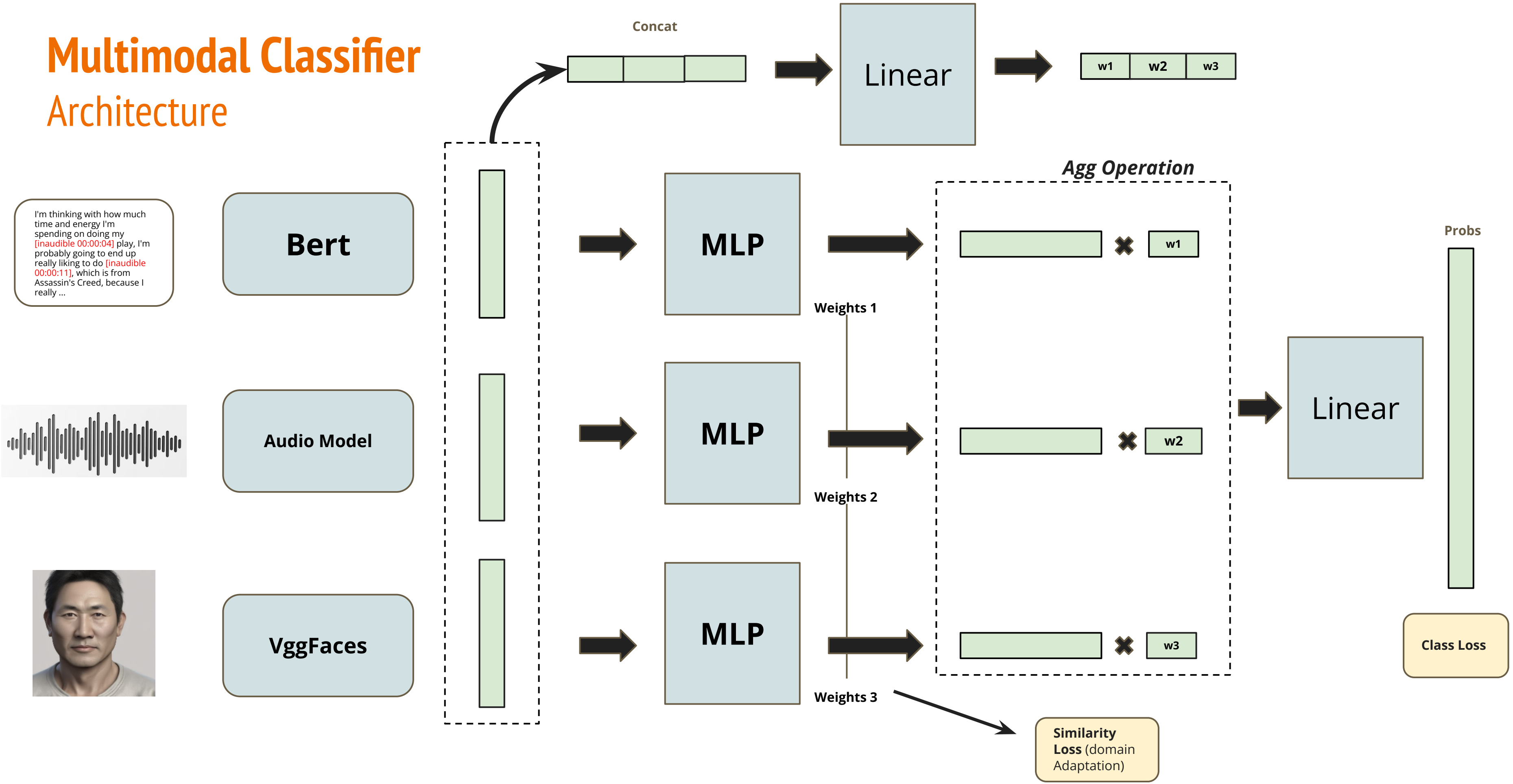
Whisper transcription:

I'm thinking with how much time and energy I'm spending on doing my EV cosplay, I'm probably going to end up really liking to do Wii for EV, which is from Assassin's Creed because I really like...

Not to much sense

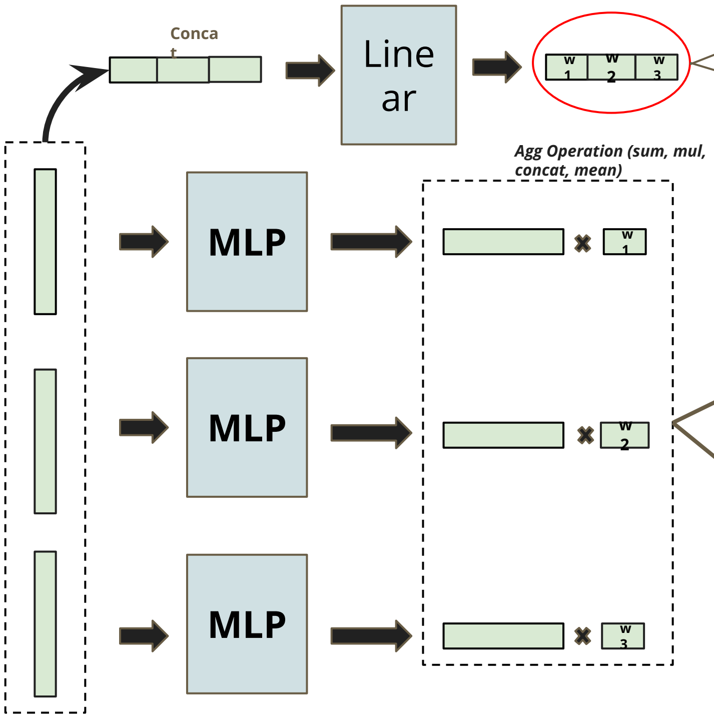
As the transcription from whisper also shows to have some inconsistencies, We end up removing the "inaudible" parts and extract the embeddings from the other part of the sentence

Multimodal Classifier
Architecture



Multimodal Classifier

Extra Ablation



Common importance order

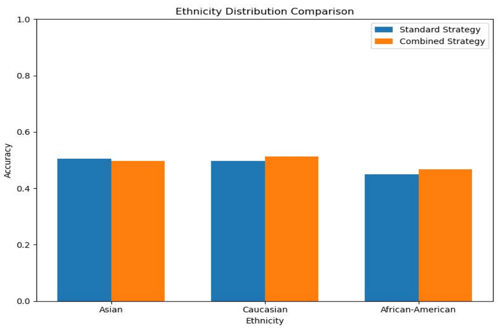
W1	***W2***	***W3***
Image *Domain*	***Audio*** *Domain*	***Text*** *Domain*
1º	2º	3º

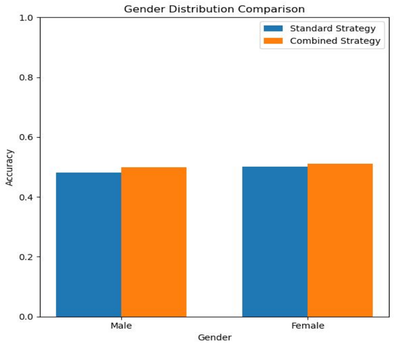
We tried different things:

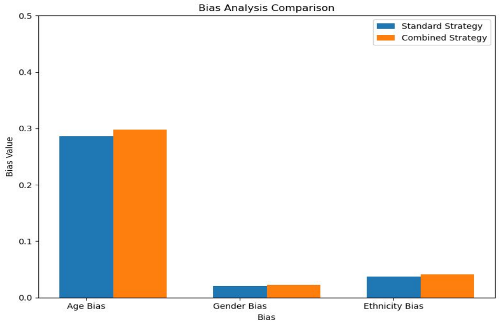
- Aligning only the image domain with the audio and not with the text. (Equal Bad results)
- Predicting separately each domain and averaging the output. (couldn't finish the training)

Test Evaluation (II)

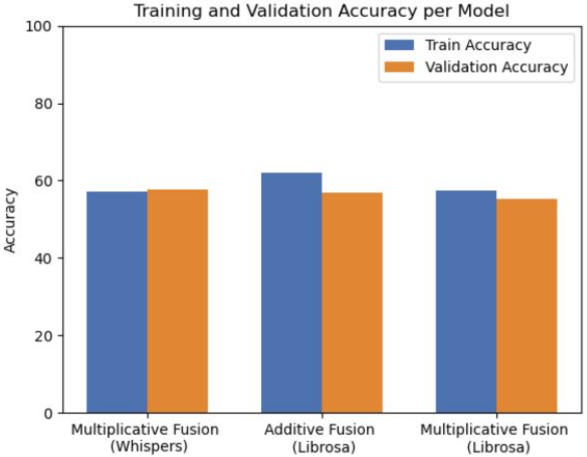
Multimodal Classifier: Best *Fusion Model and Overfitting example with Whispers*





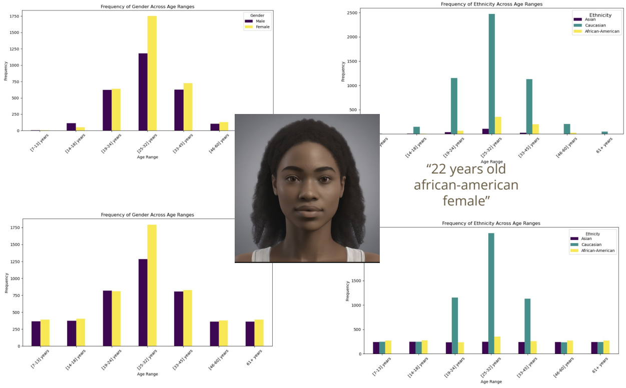


- We have only trained **three different multimodal classifiers**. Two with multiplicative fusion and one with additive function.
- We have obtained similar performances on all three models during training with the best **validation accuracy of 57,65%** without overfitting
- During **testing** we have obtained only a **49,22% of accuracy**, predicting mostly the most represented age range for most test images. Seems to be the most “fair” prediction



Multimodal Human Analysis

Summary



"22 years old african-american female"

- Train dataset **is unbalanced** specially on middle-age caucasians.
  - New images of the most disadvantaged groups have been generated with **stable diffusion model XL**.
  - Caucasians have been **downsampled** to have a more similar distribution in all groups.
- Two different ways to generate **acoustic embeddings**:
  - With **Librosa** library to extract MFCCs
  - Using **Whispers model** from Hugging Face's Transformers library
- For **text embedding extraction** we have used **BERT** from Hugging Face's Transformers library.
- For training we have applied extra regularization techniques as classes weight to face the imbalance and the label smoothing to avoid possible biases.
- For **test evaluation** we have used two strategies:
  - **Standard:** Different samples from the same person **may have different age predictions**.
  - **Combined:** All samples from the same person **must have the same age prediction**.
- We proposed a self domain adaptive method and we started to get some good results. However we couldn't dive more in this architecture due the lack of resources (2h per epoch) and time

Final Presentation:
Visual Recognition

Group 1:
Carles Pregonas
Marc Pérez
Pau Vallespí
Carlos Boned