data_exploration_PAMI2019

February 6, 2024

1 Data exploration on PAMI 2019 interim dataset

In this notebook are shown the main features of the dataset and the distribution of the data fot the PAMI dataset from emotic. The generated interim dataset from the raw data will be used to perform the data exploration, as in this format it will be easier to compute the necessary operations.

First I will import the necessary libraries and load the dataset and see an example of the data and its general annotations. Afterwards, I will perform a data sanity check to avoid having issues latter on. Then I will see the distribution of the data in the dataset and the distribution of the annotations in general. Finally, I will analyze it in more detail for each data split given and for each original db. Here is the main schematic of the data exploration: * 0. Interim data loading * 1. Data sanity check * 2. Data exploration over all data * 3. Data exploration over each split * 4. Data exploration over each original db * 5. Conclusions

Note: For visualizations I used blue to denote quantities or counts, *red* to show ratios and *orange* to show probabilities or normalized values*

2 0.Interim data loading

First I will load the libraries needed and set some debugging variables:

Now I will load the interim data and check the archives found in the folder. I expect to have three archives: `train.pkl`, `test.pkl` and `validation.pkl`, corresponding to the train, test and validation sets respectively given by the raw data annotations.

```
The annotations found are for the datasplits: ['val', 'train', 'test']
```

I will order the data splits into the following structure: train, validation and test.

```
The annotations order: ['train', 'val', 'test']
```

The available databases are:

```
The avaliable databases are:  ['ade20k', 'emodb_small', 'framesdb', 'mscoco']
```

The standard column format for all annotations is:

```
Standard column format: ['path', 'orig_db', 'img_size', 'people', 'bbox',
'label_cat', 'label_cont', 'gender', 'age']
```

And here it can be seen the first rows of the train set:

```
                                                  path orig_db    img_size  \
0  /home/usuaris/imatge/armand.de.asis/emotion_re…  mscoco  [640, 640]
1  /home/usuaris/imatge/armand.de.asis/emotion_re…  mscoco  [640, 480]
2  /home/usuaris/imatge/armand.de.asis/emotion_re…  mscoco  [640, 480]
3  /home/usuaris/imatge/armand.de.asis/emotion_re…  mscoco  [480, 640]
4  /home/usuaris/imatge/armand.de.asis/emotion_re…  mscoco  [500, 333]

   people                    bbox                              label_cat  \
0       1    [[86, 58, 564, 628]]     [[Disconnection, Doubt/Confusion]]
1       1  [[485, 149, 605, 473]]                         [[Anticipation]]
2       1   [[305, 92, 461, 465]]  [[Engagement, Excitement, Happiness]]
3       1   [[221, 63, 448, 372]]                  [[Aversion, Pleasure]]
4       1   [[44, 143, 150, 288]]             [[Confidence, Excitement]]

     label_cont gender  age
0   [[5, 3, 9]]    [M]  [A]
1   [[6, 4, 7]]    [M]  [A]
2   [[7, 8, 8]]    [M]  [T]
3   [[8, 9, 8]]    [M]  [K]
4  [[7, 9, 10]]    [M]  [A]
```

Now I will take an example of the train set and check the labels for each photo. Each photo is encoded in BGR instead on RGB, so I will need to convert it to RGB to visualize it correctly. For each person , there is annotated: the gender, the age, the emotion categories and the continuous labels.

Regarding the continuous labels, it shall be remembered that valence tells how good or bad the person is feeling (0 to 10), while arousal tells how calm or excited the person is feeling (0 to 10) and dominance tells how much control the person feels they have (0 to 10).

```
The path of the example image is: /home/usuaris/imatge/armand.de.asis/emotion_re
cognition/data/raw/PAMI/emotic/emotic/mscoco/images/COCO_train2014_000000288841.
jpg
The image orig DB is: mscoco
The image shape is: (480, 640, 3)
There is a total of 1 annotated people in the image
```


```
+--------+--------+-----+--------------------+--------------------------------
------------------+
| Person | Gender | Age | Emotions categories | Continuous emotions [Valence,
Arousal, Dominance] |
+--------+--------+-----+--------------------+--------------------------------
------------------+
|   0    |   M    |  A  |   ['Anticipation']  |                     [6, 4, 7]
|
+--------+--------+-----+--------------------+--------------------------------
------------------+
```

In this initial sample, it can seen that not all the people are annotated and it does not relate to the closeness between the camera and the people. I will need to take this into account when I are training the model.

```
The path of the example image is: /home/usuaris/imatge/armand.de.asis/emotion_re
cognition/data/raw/PAMI/emotic/emotic/mscoco/images/COCO_train2014_000000251754.
jpg
The image orig DB is: mscoco
The image shape is: (457, 640, 3)
There is a total of 3 annotated people in the image
```


```
+--------+--------+-----+-----------------------------+-------------------------
-------------------------+
| Person | Gender | Age |     Emotions categories     | Continuous emotions
[Valence, Arousal, Dominance] |
+--------+--------+-----+-----------------------------+-------------------------
-------------------------+
|   0    |   M    |  A  | ['Engagement', 'Happiness'] |                      [4,
4, 4]                      |
|   1    |   M    |  A  |        ['Excitement']       |                      [6,
6, 6]                      |
|   2    |   F    |  A  |        ['Engagement']       |                      [7,
4, 7]                      |
+--------+--------+-----+-----------------------------+-------------------------
-------------------------+
```

3 1.Data sanity check

Now I will check if the data is correctly loaded and if the annotations are correct. I will check the following:

- 1.1. Check NA values
- 1.2. Check img size and channels
- 1.3. Check bbox

3.1 1.1.Check NA values

I will check the NA values for continuous and categorical labels. First I will search for the nan values for all dataset.

Now, the results are shown in a table:

Original DB	NA on categorical label	NA on continous label
ade20k	0	0
emodb_small	1	2
framesdb	0	357
mscoco	0	0

It can observed that there is only one NA value in the categorical labels. But there are a lot of NA values in the continuous labels. Now I will show which are the images with NA values in the continuous labels. Concretely the `show_NA_images` to avoid cluttering the notebook:

```
The image ' train 23 ' with the annotated people labels: [[7, 4, 7], [nan, nan,
nan]]
The image ' train 106 ' with the annotated people labels: [[7, 7, 6], [nan, nan,
nan], [7, 8, 4], [6, 5, 6], [nan, nan, nan], [5, 7, 5]]
The image ' train 212 ' with the annotated people labels: [[8, 6, 7], [8, 5, 7],
[nan, nan, nan]]
The image ' train 250 ' with the annotated people labels: [[5, 2, 9], [nan, nan,
nan]]
The image ' train 263 ' with the annotated people labels: [[nan, nan, nan]]
```

Here it can be observed one of these images:

```
The path of the example image is: /home/usuaris/imatge/armand.de.asis/emotion_re
cognition/data/raw/PAMI/emotic/emotic/framesdb/images/frame_dy0xnq4fy1sn1ra8.jpg
The image orig DB is: framesdb
The image shape is: (2416, 4288, 3)
There is a total of 6 annotated people in the image
```

Person	Gender	Age	Emotions categories	Continuous emotions [Valence, Arousal, Dominance]
0	F	A	['Anticipation', 'Engagement']	[5, 2, 8]
1	F	A	['Disconnection']	[nan, nan, nan]
2	F	A	['Anticipation', 'Engagement']	[5, 2, 9]
3	M	A	['Disconnection']	[nan, nan, nan]
4	F	A	['Engagement']	[nan, nan, nan]
5	F	A	['Disconnection']	[5, 4, 5]

Even though there are annotated people in the image, there are no continuous annotations for them. And most of the bbox annotations are poor and with no logical approach (as there are many people of similar size and locality that is not being annotated). I will need to take this into account when preprocessing the data.

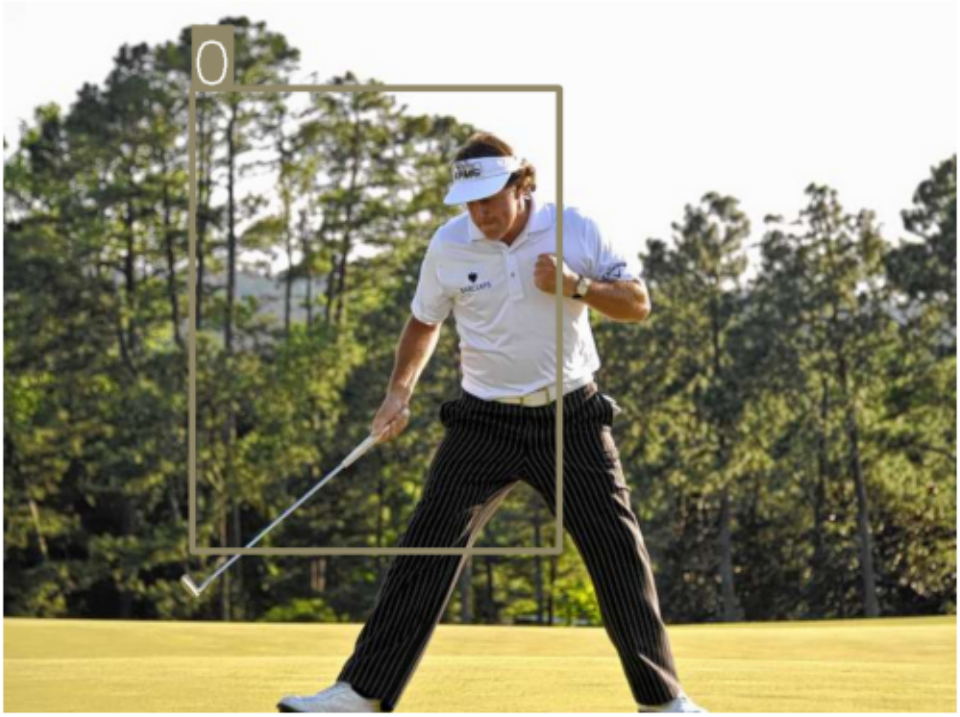
3.2 1.2.Check img size and channels

Now I will focus on image sizes and check if they are given in a correct format and size. From experimenting, I discovered that emodb_small has inverted the height and width of the images, instead of the standard width and height. I will keep in mind this in order to not have problems when making the checks.

```
The image 398 in the data split 'train' has a different size than the one stored
in the annotations.

libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: extra compressed data
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile

The image 6838 in the data split 'train' has a different size than the one
stored in the annotations.

Corrupt JPEG data: 44 extraneous bytes before marker 0xd9

The image 8594 in the data split 'train' has a different size than the one
stored in the annotations.
The image 9978 in the data split 'train' has a different size than the one
stored in the annotations.
The image 14846 in the data split 'train' has a different size than the one
stored in the annotations.

libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
+-------------+------------------------------+----------------------------------
+
| Original DB | Number of different img_size | Number of black and white images
|
+-------------+------------------------------+----------------------------------
+
|    ade20k   |              0               |                0
|
| emodb_small |              0               |                0
|
|   framesdb  |              5               |                0
|
|    mscoco   |              0               |                0
|
+-------------+------------------------------+----------------------------------
+
```

There are only 5 images that can be considered different as the given img size. I show two of these images and see how they are annotated:

```
The image size stored in the annotations is: 566 x 730
The path of the example image is: /home/usuaris/imatge/armand.de.asis/emotion_re
cognition/data/raw/PAMI/emotic/emotic/framesdb/images/frame_0cs63q843g5bdvq5.jpg
The image orig DB is: framesdb
The image shape is: (724, 971, 3)
There is a total of 1 annotated people in the image
```




```
+--------+--------+-----+-----------------------------------------------------
+-------------------------------------------------+
| Person | Gender | Age |                  Emotions categories                  |
Continuous emotions [Valence, Arousal, Dominance] |
+--------+--------+-----+-----------------------------------------------------
+-------------------------------------------------+
|   0    |   M    |  A  | ['Engagement', 'Excitement', 'Pleasure', 'Yearning'] |
[6, 9, 9]                      |
+--------+--------+-----+-----------------------------------------------------
+-------------------------------------------------+

The image size stored in the annotations is: 4928 x 3264
The path of the example image is: /home/usuaris/imatge/armand.de.asis/emotion_re
cognition/data/raw/PAMI/emotic/emotic/framesdb/images/frame_apztd5rqirir5vuz.jpg
The image orig DB is: framesdb
The image shape is: (3264, 4928, 3)
There is a total of 2 annotated people in the image
```


```
+--------+--------+-----+--------------------+--------------------------------
------------------+
| Person | Gender | Age | Emotions categories | Continuous emotions [Valence,
Arousal, Dominance] |
+--------+--------+-----+--------------------+--------------------------------
------------------+
|   0    |   M    |  A  |    ['Engagement']   |                    [6, 8, 7]
|
|   1    |   M    |  A  |    ['Engagement']   |                    [5, 7, 8]
|
+--------+--------+-----+--------------------+--------------------------------
------------------+
```

It can be seen that the images bbox make no sense at all, so maybe the given image is incorrect. I will need to take this into account when preprocessing the data.

3.3 1.3.Check bbox

Now I will check if the bbox are inside the image.

```
The number of not valid bounding boxes is: 400
22      are too small to 0,0 origin
360     are too big with respect to the image size, with mean error of
4.377777777777778
18      are too small and too big
0       are in incorrect relation between x1, x2, y1, y2
```

```
The errors comes from the following original databases:
emodb_small : 400

The mean error for x1 < 0 or y1 < 0:             2.025
The mean error for x2 > width or y2 > height:    4.169
```

It can observed that many of these errors come from emodb that has bbox that are outside the image. But as the errors are small, I will not correct them now as they are not very important for the general distribution of bbox sizes.

4 2.Data exploration over all data

First I will analyze the distribution of the data over all the dataset. We will analyze the following: * 2.1. Total people and images available * 2.2. Gender and age distribution * 2.3. Bbox distribution * 2.4. Emotion distribution

4.1 2.1.Total people and images available

The total number of images and people is:

```
Total number of images: 23554
Total number of people: 34320
```

The number of images does not coincide with the paper, but the number of people does, so maybe the number of images is wrong in the paper or maybe there are some images that are not given.

4.2 2.2.Gender and age distribution

Now I will have a look on the age and gender distributions over all data. First I will compute the metrics and then visualize it:

The results are shown using a bar chart:

```
alt.HConcatChart(…)
```

It can be observed that the most common annotations is a male adult. For the people gender distribution it is more or less balanced, but for the age distribution it is not. There are much more adults than children and more children than elderly people. This is important to take into account when training the model, as it will be more biased to predict well on adults than to children or elderly people.

4.3 2.3.Body bbox distribution:

Now I will check the bbox distribution over all data. First I will compute the metrics and then visualize it using a histogram:

```
alt.HConcatChart(…)
```

It can be seen that many of the probability mass is centered around 0 to 500 px. This is important to take into account when choosing the model input size and be able to fit most images without making any distortion over them.

Now I will check the distribution of the aspect ratio using a histogram:

```
alt.Chart(…)
```

The ratio is defined as the width divided by the height. I decided to use the log scale to see the distribution better. This way I can expand the values between [0,1] (the bbox with vertical rectangle), and see with equal relevance the values situated at the right of the plot (the bbox with horizontal rectangle).

It can observed that the main kind of bbox is the vertical one, as the ratio is between 0 and 1. Also we observe the biggest bar is located at 1, so many values have a 1:1 ratio. Also we have a second spike around 1:2 so there are many bbox with a 1:2 ratio. This is expected as many people are standing up and the bbox are vertical rectangles.

4.4 2.4.Emotion distribution

Now I will check the emotion distribution over all data. I will focus on analyzing first the categorical labels and then the continuous labels.

For the categorical labels I will analyze: * 2.4.1.1. Number of emotions felt per each person * 2.4.1.2. Categorical emotion distribution * 2.4.1.3. Co-ocurrence matrix * 2.4.1.4. Mean color for each categorical emotion

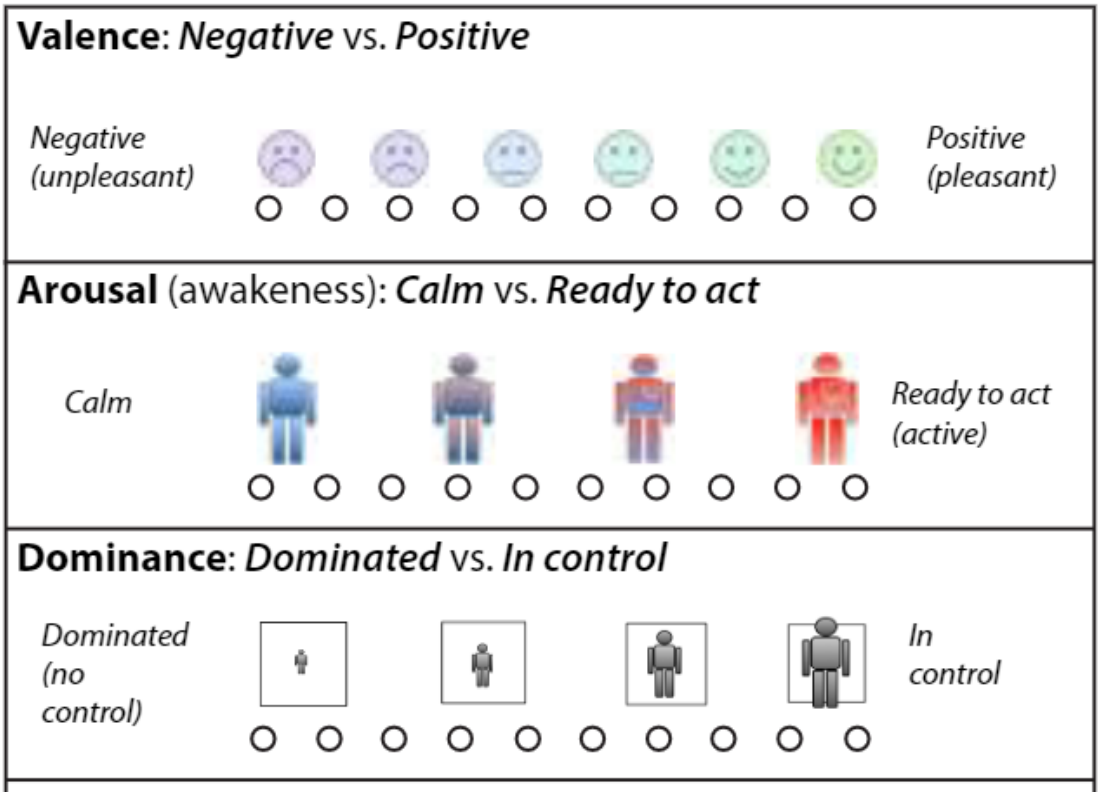
For the continuous labels I will analyze: * 2.4.2.1. Continuous emotion distribution * 2.4.2.2. Covariance matrix

4.4.1 2.4.1 Categorical annotations analysis and exploration

The main goal of this section is to understand the distribution of the categorical annotations in the dataset. I will check the distribution of the annotations in the whole dataset. The annotated categorical annotations are (I take train datasplit as example):

```
In total, there are 26 emotions annotated.
The emotions found are: ['Disconnection', 'Doubt/Confusion', 'Anticipation',
'Engagement', 'Excitement', 'Happiness', 'Aversion', 'Pleasure', 'Confidence',
'Peace', 'Fatigue', 'Pain', 'Sadness', 'Sensitivity', 'Suffering', 'Sympathy',
'Fear', 'Yearning', 'Disquietment', 'Esteem', 'Annoyance', 'Affection', 'Anger',
'Disapproval', 'Embarrassment', 'Surprise']
```

Here we can observe the meaning of each emotion following the paper:



2.4.1.1. Number of emotions felt per each person First I will check the number of emotions felt per each person, this way we can know the expected number of predictions that will be done by the model when we train it. I will compute the metrics and then visualize it using a histogram:

Now I analyze the number of emotions per photo:

```
alt.LayerChart(…)
```

It can be observed that many people has only one or two emotions per photo, but there are also people with more than 10 emotions per photo. This heavily weighted distribution towards the left will be important to take into account when training the model, as the model will try to predict around 2 or 3 emotions, and will have more difficulties with the less common ones above 8 as they are less frequent in the dataset.

```
The Q1 quartile (25%) is located at: 1.0
The Q2 quartile (50%) or median is located at: 2.0
The Q3 quartile (75%) is located at: 4.0
```

The distribution is heavily located in the left as it can be seen the quartiles. Around 75% of entries have less than 4.0 emotions per photo.

```
The probability to have 8 or more emotions is: 2.534965034965035 %
```

It can be seen in the distribution that less than 5% of the entries have 8 or more emotions per person.

2.4.1.2. Categorical emotion distribution Now I analyze the occurrence of all the emotions in the dataset:

```
alt.Chart(…)
```

It can be observed that the label distribution is not uniform. This is expected, as there are many emotions that has less occurrences than others. I will need to take care of this when training the model, as it could be biased towards the most common emotions.

I can also observe how the most usual emotion felt is engagement. I will take a closer look, to see the distribution of it to see if it's the only emotion felt or there are more:

Now I will plot the results:

```
alt.Chart(…)
```

It can observed that there are 5000 people that is labeled only as engagement. So I must take this into account when I are training the model and I must be careful with the results. Some data augmentation or label balance techniques could be useful to solve this problem.

2.4.1.3. Co-occurrence matrix I will now compute the co-occurrence matrix to see the distribution of the emotions felt together. This will be useful to see if there are emotions that are usually felt together or not and establish some relations between them.

Now I compute all the pairs, and add count them to add them to the co-occurrence matrix:

Now I will plot the result. I will use a heatmap to see the distribution of the emotions felt together. I set the max value to the maximum number of co-occurrences between different emotions, this way I can focus on the information of different co-occurrences (as the emotion ocurrences are detailed in the previous sections and make this information harder to see):

```
alt.LayerChart(…)
```

It can be seen that there are some emotions that are heavily related, as they have a great value on the co-occurrence matrix. For example, engagement and excitement are usually felt together. Also, there are some emotions that are not usually felt together, like sadness and excitement. This is expected, as they are opposite emotions. I will need to keep this in mind as the model will try to predict the emotions felt together, and it will be easier to predict the emotions that are usually felt together.

Now I will normalize each row of the co-occurrence matrix to see the the conditioned probability. This will help to see the probability of each emotion given another emotion. This way I can focus on the emotions that does not have as much entries and see how given a emotion (y axis), the probability of another emotion is affected (x axis).

Now I plot the result.

```
alt.LayerChart(…)
```

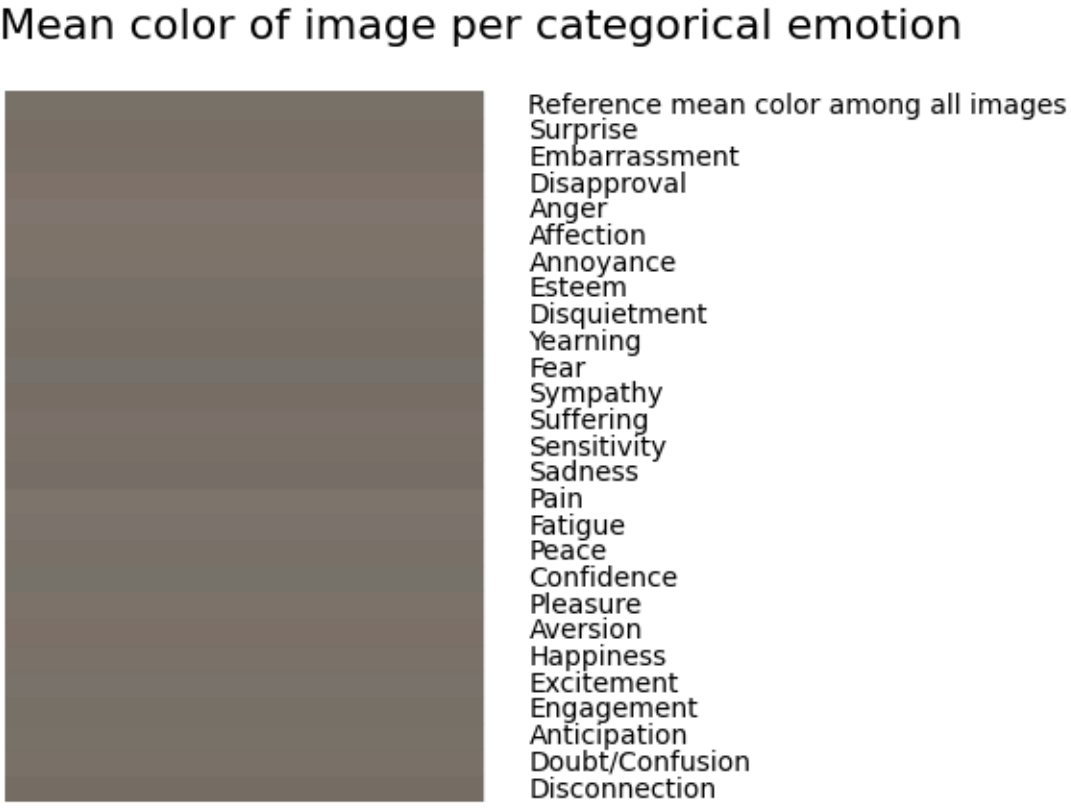
It can be observed that the most probable variable is engagement, this is expected as it is the most common entry in the dataset. Also we can see that the probabilites does not sum up to 1. This is

expected as the emotions are usually felt together. For example, if a person is feeling yearning is usually feeling anticipation and engament with a probability of around 0.5.

2.4.1.4. Mean color for each categorical emotion Now I will study if there is any relation between the mean color of the image and the emotion felt. I will compute the mean color for each photo to which is annotated the emotion and plot the results.

First I will compute the mean color for each photo:

```
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: extra compressed data
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
Corrupt JPEG data: 44 extraneous bytes before marker 0xd9
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
```

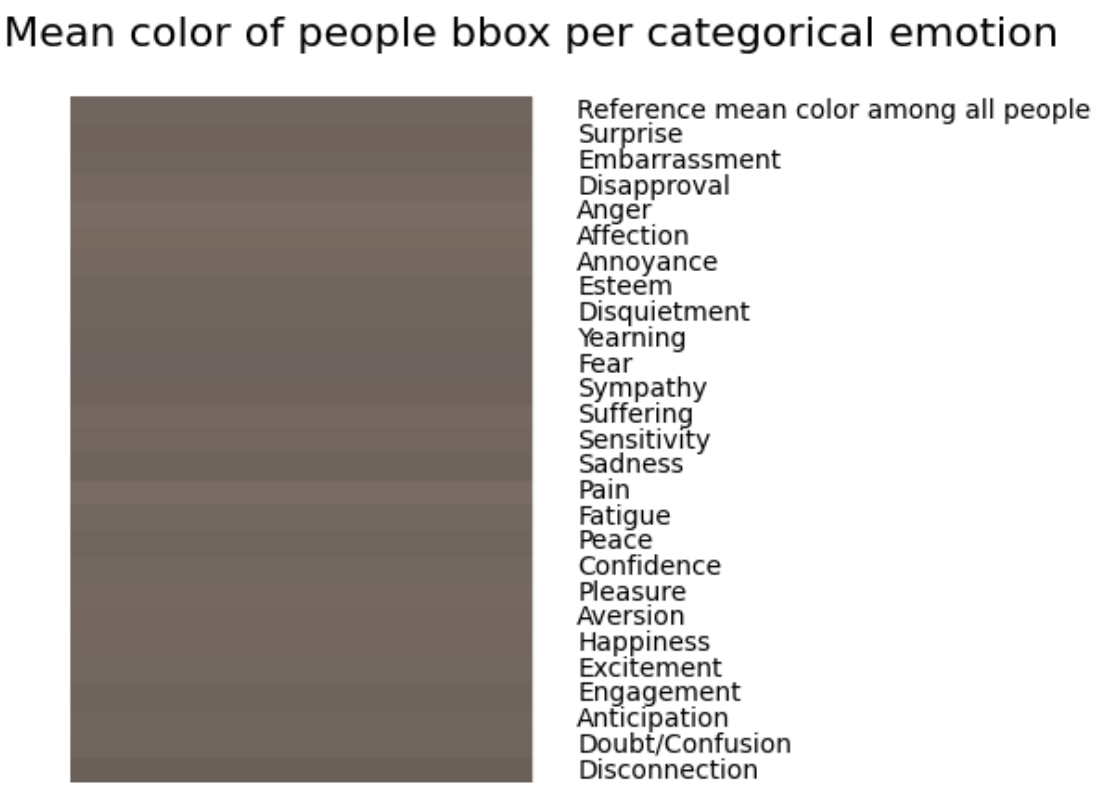


We can observe that there is some slight variation between the mean colors of each emotion. For example, the mean color of the photos with the emotion fear is a bit blueish than the mean color of the photos with the emotion anger. But as the the variation is not very big, I will not take this into

account when training the model (if the training give me the expected results), I will only perform a normalization along all the images.

Now I repeat for each person (not for whole image), if this effect is more noticeable:

```
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: extra compressed data
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
Corrupt JPEG data: 44 extraneous bytes before marker 0xd9
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
```



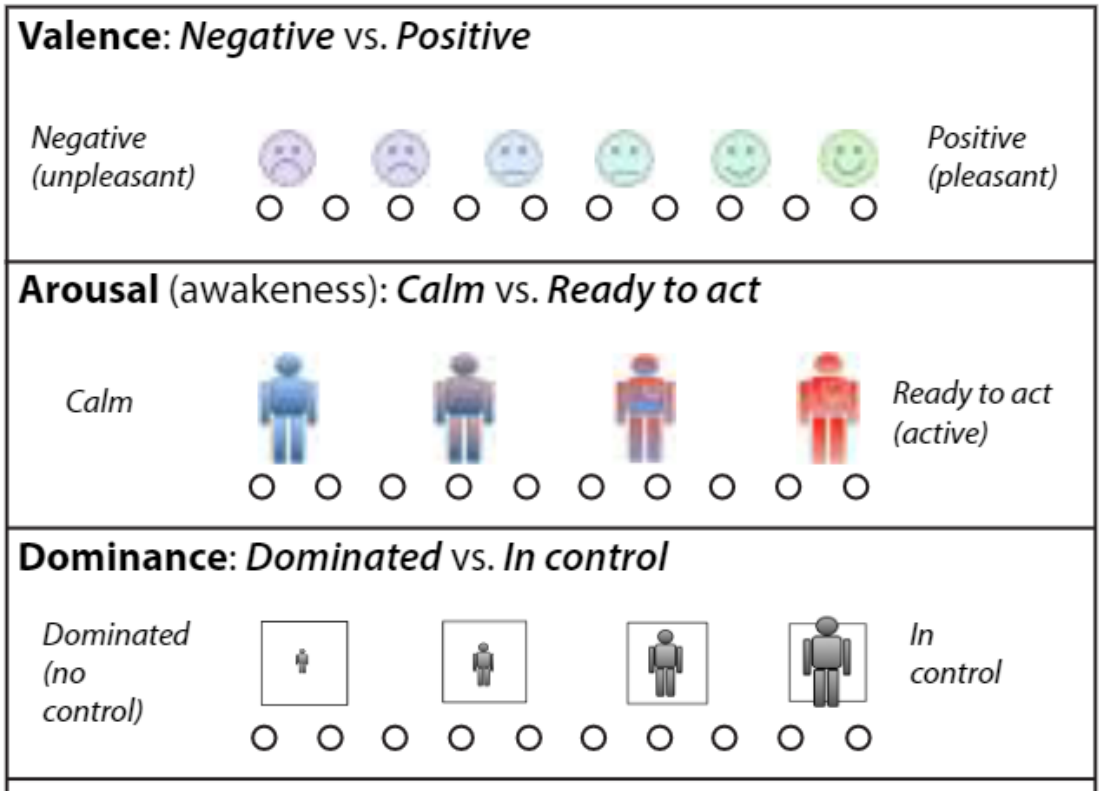
Now it can be observed that in general the mean colors are different from we have seen. For example, Surprise now is a bit reddish than the other emotions, and fear now more standard. But as the the variation is not very big, I have the same conclusion as the past section.

4.4.2 2.4.2. Continuous annotations analysis and exploration

To do that, first I will analize the general distribution continuous annotations. The continuous annotations are: 1. Valence: The negativity/positivity of the emotion felt. It is a value between -1 and 1, where -1 is the most negative valence and 1 is the most positive valence. 2. Arousal: The

intensitivity of the emotion felt. It is a value between -1 and 1, where -1 is the most calm and 1 is the most excited emotion. 3. Dominance: The dominance of the emotion felt. It is a value between 1 and 10, where 0 is the most submissive dominance and 10 is the most dominant dominance.

Here we can observe the meaning of each continuous annotation following the paper:



2.4.2.1. Continuous label distribution: Now I will compute the necessary metrics to see the distribution of the continuous annotations in the dataset.

```
alt.HConcatChart(…)
```

It can be seen that the distributions are far from the uniform distribution that we would like to see. The humans tend to annotate more towards tge middle of the range, and less to the extremes. This is important to take into account when training the model, as it will be more biased to predict well the annotations in the middle of the range than the annotations in the extremes. The distribution that shows a more spread weight is the arousal distribution, as it has more weight in 4-7 and not a huge spike in the middle of the range as valence and dominance.

Now I will show the mean and standard deviation of each continuous annotation:

Continuous annotation	Mean	Std. Dev.
valence	5.984	1.316
arousal	5.613	1.917

```
|       dominance       | 6.483 |   1.675   |
+-----------------------+-------+-----------+
```

The results are the expected ones from the seen distributions. The distribution that shows a bigger standard deviation is the arousal one. We can observe also that dominance is centered around 6.5 and valence around 5.9. This is important to take into account when training the model, as it will be more biased to predict the annotations in the middle of the range than the annotations in the extremes.

2.4.2.2. Covariance matrix: Now I will check the covariance matrix between the continuous annotations. This will help us to see if there are some annotations that are related between them. This will be useful to see if the variables are independent.

```
alt.LayerChart(…)
```

It can be observed that the variables are not strictly independent, as the covariance matrix is not diagonal. This is expected, as the annotations are a bit related between them. Now we will check the correlation matrix to see better the relation between the variables. This will help us to see if there are some annotations that are related between them. This will be useful to see if the variables are independent.

```
alt.LayerChart(…)
```

We can observe that the correlation in general is very low. On the one hand, valence and arousal have a value of 0.16 showing a very weak relation. On the other hand, arousal and dominance and valence and dominance have a value of 0.27 and 0.29 respectively, showing a weak relation. This is expected, as the annotations are a bit related between them. For example, if the arousal is high (high stimulation) is expected that the dominance is usually high too (control of emotion). This is maybe caused by the fact that the higher the stimulus the higher control over the emotion I have (example of skater).

5 3.Data exploration over each split

Here I will check the main features of each data split, such as the number of rows and columns, the data types of the columns and the number of null values. I will also store the insights to then create visualizations to ease the comparison between the splits.

I will check the following: * 3.1. Data splits main features * 3.2. Gender and age distribution for each data split * 3.3. Categorical annotations distributions on data splits for each data split * 3.4. Continuous annotations distributions on data splits for each data split

5.1 3.1.Data splits main features

Now I show the main features of each split and the ratio of each split with respect to the whole dataset:

```
+------------+------------------+--------------------------+-----------------
+------------------+
| Data Split | Number of Columns | Equal to standard columns | Number of images
| Number of People |
```

```
+------------+-------------------+--------------------------+-----------------
+------------------+
|   train    |         9         |           True           |      17077
|      23706       |
|    val     |         9         |           True           |       2088
|       3334       |
|    test    |         9         |           True           |       4389
|       7280       |
+------------+-------------------+--------------------------+-----------------
+------------------+
The ratio of images per each datasplits is: train 0.73 / val 0.09 / test 0.19
```

All data splits have the same number of columns and are equal to the standard columns. The number of images are different one to another as it can be seen in the table. Also the ratio of train, validation and test is unusual. When training the model and doing the hyperparameter optimization this is needed to keep in mind if we follow this structure.

Now I will observe the images in each data split and the ratio of people appearing on it:

```
alt.HConcatChart(…)
```

It can be observed that the ratio of people is very similar, having most in the test set, followed by the validation and train sets. This is expected as the test set is the hardest one. But when talking about sized we can see that the train validation is the one that has the biggest size. This is expected as the train set is the one that has the most images. Also it can be observed the strange proportion size of test and validation sets I talked about earlier.

5.2 3.2.Gender and age distribution for each data split

Now I will check the gender and age distribution across the data splits to see if they are similar or not.

Now I will plot the results:

```
alt.HConcatChart(…)
```

We can observe that across splits the results are very similar regarding gender. This is expected as the data splits are made to be similar one to another. Only there is a slight difference between age distributions.

5.3 3.3. Categorical annotations distribution for each data split

Now I will check the categorical annotations distribution across the data splits to see if they are similar or not.

```
alt.VConcatChart(…)
```

The results are similar one to another, but there are some differences. For example, the engagement emotion is more common in the train set than in the validation and test sets. Regarding the other most common emotions, they are more or less similar one to another with slight differences but following the general distribution pattern.

5.4 3.4. Continuous annotations distribution for each data split

```
alt.VConcatChart(…)
```

It can be observed that the validation data split is the one with the most centered mass around the middle of the range. This is expected as the validation set is the one that has the less images. Regarding the train split, the arousal distribution is the one that shows more spread weight, as it has more weight in 4-7 and not a big spike in the middle of the range as valence and dominance. Regarding the dominance distribution, it is weighted a bit to the right, as it has more weight in the right side of the range. And finally, analyzing the test data split it can be also seen that it does not follow a uniform distribution, as it is weighted to the middle-right of the range.

Now I will show the mean and standard deviation of each continuous annotation:

```
+------------+--------------+-------------+--------------+-------------+--------
--------+---------------+
| Data Split | Valence Mean | Valence Std | Arousal Mean | Arousal Std |
Dominance Mean | Dominance Std |
+------------+--------------+-------------+--------------+-------------+--------
--------+---------------+
|   train    |     6.02     |     1.38    |     5.56     |     2.1     |
6.57       |      1.86     |
|    val     |     6.07     |     1.03    |     5.81     |     1.25    |
6.12       |      0.92     |
|    test    |     5.83     |     1.2     |     5.7      |     1.52    |
6.37       |      1.24     |
+------------+--------------+-------------+--------------+-------------+--------
--------+---------------+
```

6 4.Data exploration over each original db

Based on the PAMI2019 paper, it is known that the EMOTIC dataset was obtained using two methods. The first one, was done manually collecting from the Internet by Google search engine the photos. For that, he used a combination of queries containing various places, social environments, different activities and a variety of keywords on emotional states. The rest of images belong to 2 public benchmark datasets: COCO and Ade20k.

First I will see the number of original db I have at our disposal:

```
The avaliable databases are:  ['ade20k', 'emodb_small', 'framesdb', 'mscoco']
```

I will check the following: * 4.1. Original db main features * 4.2. Gender and age distribution for each data split * 4.3. Categorical annotations distributions on data splits for each data split * 4.4. Continuous annotations distributions on data splits for each data split

6.0.1 4.1.Original database main features

First I will see the count of images per dataset:

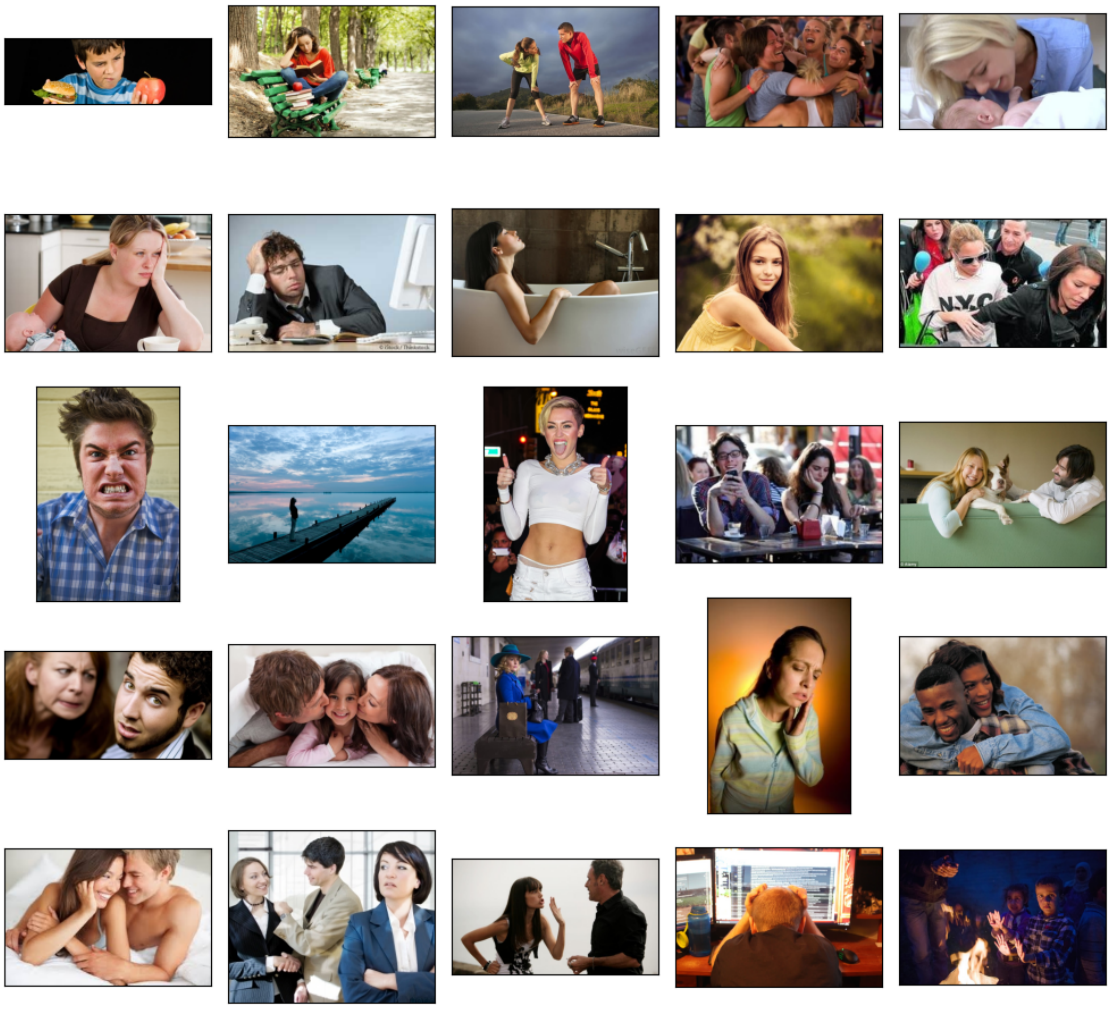
```
alt.HConcatChart(…)
```

Orig_DB	Number of images	Number of People
ade20k	432	648
emodb_small	1374	2417
framesdb	5252	10553
mscoco	16496	20702

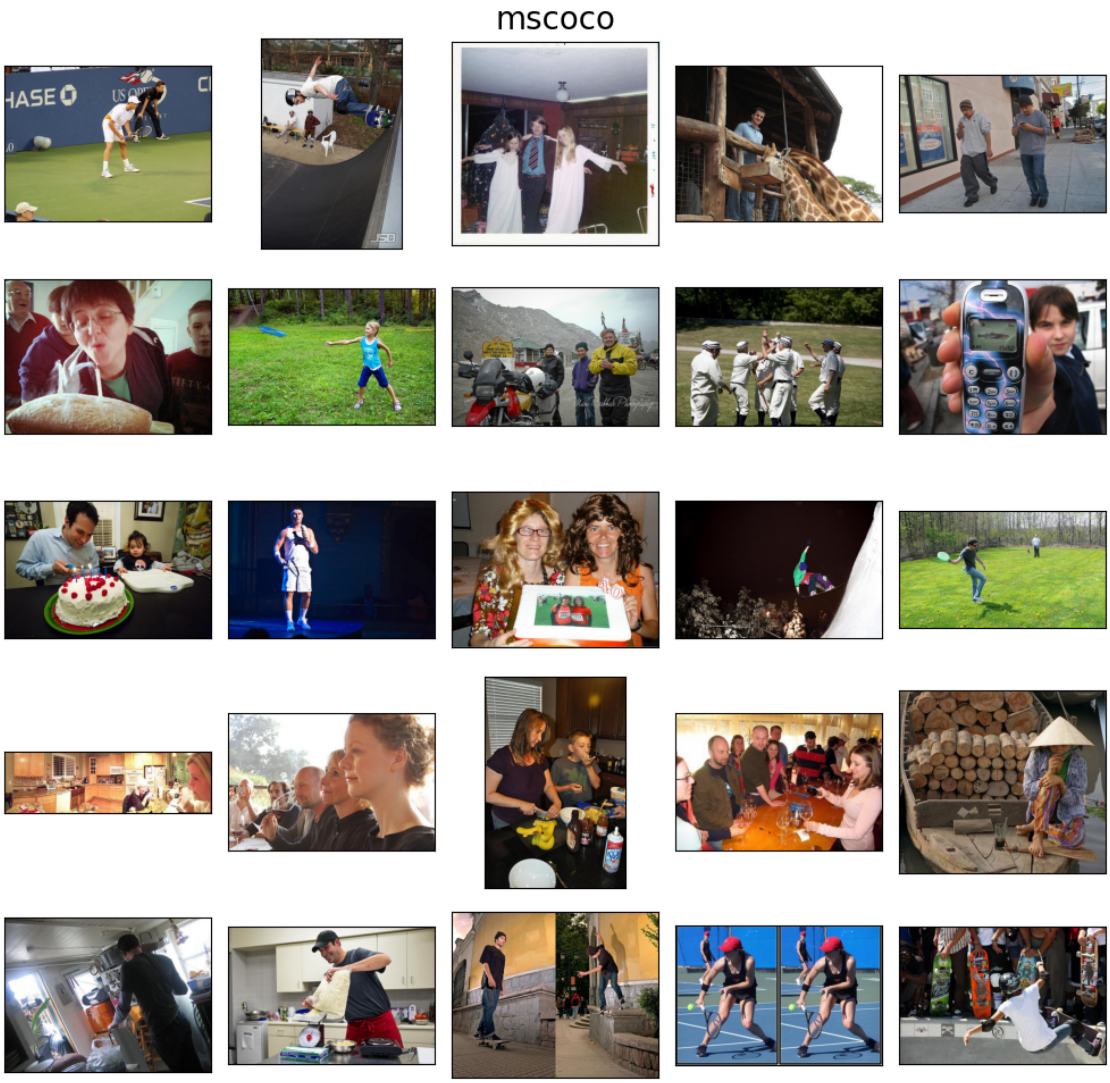
It can be observed that the dataset that shows the biggest number of images is the COCO dataset, followed by the framesdb dataset, emodb_small and finally the Ade20k dataset. The one that has more people per photo is the framesdb dataset, followed by the emodb_small dataset, the Ade20k dataset and finally the COCO dataset. This is expected as the mscoco is a general dataset and the framesdb is a dataset that has been created by the author of the paper for emotion recognition.

Now we will check the main look of the photos in each dataset as in general each database presents certain biasing towards one style of photo. First I will sample some photos from each dataset and then I will show them:

emodb_small




framesdb
Book Your Own Private
Bourbon Street Balcony



We can observe that the COCO and Ade20k show more "in the wild" photos. The framesdb dataset shows many watermarks. The emodb_small dataset shows many "stock" alike photos, so the emotions or situations are a bit forced.

6.0.2 4.2.Age and gender distribution:

```
alt.HConcatChart(…)
```

We can observe that the distributions differ from one database to another. Regarding gender, the one with the highest difference between male and female persons is mscoco. In emodb small, is the only dataset where women have more entries. Regarding age, the one with the highest difference between adults and children is ade20k, with most entries being adult. On the other hand, emodb_small is the dataset with more kids in proportion.

6.0.3 4.3.Bbox distribution:

Now I will study the bbox distribution for each original dataset. First I will compute the metrics and then visualize it:

Now I show the distribution of the bbox for each dataset:

```
alt.VConcatChart(…)
```

We can observe that in general all distributions are heavily weighted towards 0. This is expected, as many bbox are from small photos. This is the case for ade20k and mscoco datasets. For the emodb_small and framesdb datasetm we can see that there is a tail along the x axis, this is because the bbox are bigger as the images are bigger, reaching up to 4.000 px.

Now I will check the ratio of image:

```
alt.Chart(…)
```

We take the same approach as the section 2.3, and apply the log-scale to see better the 1:1 ratio and see if bbox are in general vertical or horizontal. In general, we can observe a similar distribution, more heavy weighted towards the vertical bbox. But we can observe that the emodb_small dataset has a more balanced distribution. Regarding framesdb, we can observe that the distribution mostly vertical, having a spike around 0.5 (1:2) bigger than 1:1, that is the case for all orher distributions.

6.0.4 4.4.Categorical annotations distribution on original databases

Know I will analyze the categorical emotion distribution for each original database. First I will compute the metrics and then visualize it using a histogram:

```
alt.VConcatChart(…)
```

We can observe that the distributions are not uniform, except for emodb_small, that has the weight more distributed. In mscoco almost 0.2 of the entries are engagement. This is important to take into account when training the model, as it could be biased towards the most common emotions.

6.0.5 4.5.Categorical annotations distribution on original databases

```
alt.VConcatChart(…)
```

We can observe that the emodb_small is the database having more weight distributed. Regarding the other databases, they are more or less similar one to another with slight differences but following the general distribution pattern. The ade20k is the one with the most centered mass.

7 5.Conclusions on data exploration

After the long data exploration on the emotic PAMI 2019, I have concluded the following:

1. For the categorical emotions labels some label weighting or data augmentation techniques should be done as there the label distribution is unbalanced. It can observed that there are 5000 people that is labeled only as engagement, so we will need to see if this is a problem when training the model.

2. For the continuous emotions labels most labels are centered around the middle of the range, so the model will be more biased to predict the annotations in the middle of the range than the annotations in the extremes.
3. Gender and age distributions are unbalanced (mostly age). So the model it is expected to have better predictions on adult mans (and a slightly worse on adult women).
4. The bbox distribution is heavily weighted towards 0 to 1000 px, but few above that threshold. This is needed to keep in mind when choosing the model input size and be able to fit most images without making any distortion over them.
5. Also the ratio of train, validation and test are unusual. When training the model and doing the hyperparameter optimization this is needed to keep in mind if I follow this structure.
6. This dataset cannot be used to train the bbox regressor, as not all people are not annotated.

So after all this, I decided to do the following for the data preprocessing and training: 1. Do some downscaling of the images or erase the images that are above the 1000px threshold of bbox. 2. Do some data augmentation or label weighting techniques to balance the categorical labels. 3. Take only the `mscoco`, `emodb_small` and `ade20k` dataset. `Mscoco` is the biggest dataset and it comes from the well studied photos. Even thought the labels are not well distributed, it will help the model to see more "in the wild" cases. `Emodb_small` showed a better distributed labelling and, even though many of its photos are a bit unreal emotions, it will help the model to see more rare cases. `Ade20k` is the smallest dataset, but it has almost the same distribution of labels as `mscoco` and it will help the model to see more "in the wild" cases. `Framesdb` didn't show astonishing annotation distributions and it's photos are not as well studied as mscoco (many of them include watermarks). Also it has a lot of NA values, incorrect annotations so this lead me to untrust the annotations present in this dataset. So i decided to discard it. 4. I will ignore the data splits and I will introduce all this data to other data splits following a 33-33-33 ratio. 5. Perform an image normalization to have a similar mean color across all the images.

8 References

R. Kosti, J.M. Álvarez, A. Recasens and A. Lapedriza, "Context based emotion recognition using emotic dataset", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2019. Emotic paper