

# Feeling Your Images: Visual Emotion Recognition based on Image Attributes.

Matias Lessa Vaz – [up201900194@fc.up.pt](mailto:up201900194@fc.up.pt)  
Computer Vision – Universidade do Porto

---

A brief slide show about why that's important,  
the engineering behind the process  
and the achieve results.



# Objective?

---



Our primary goal is to move beyond conventional approaches that predominantly focus on facial features.



Instead, we aim to develop a robust system that comprehensively understands the visual context of an entire image, offering a nuanced interpretation of the scene.

---

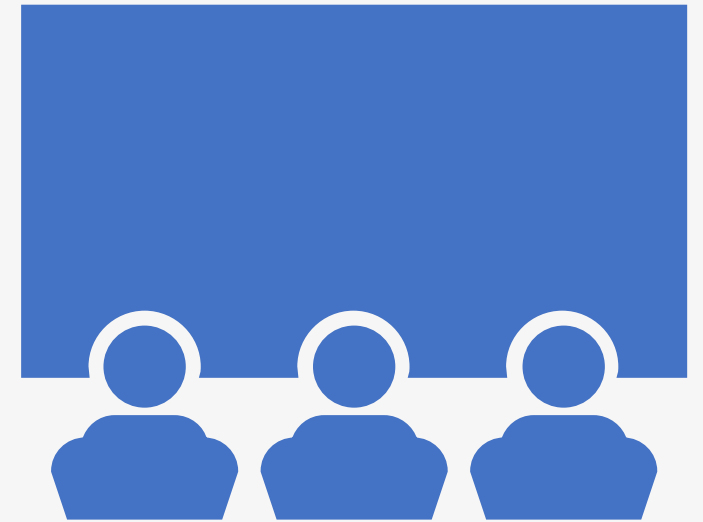
# Why?

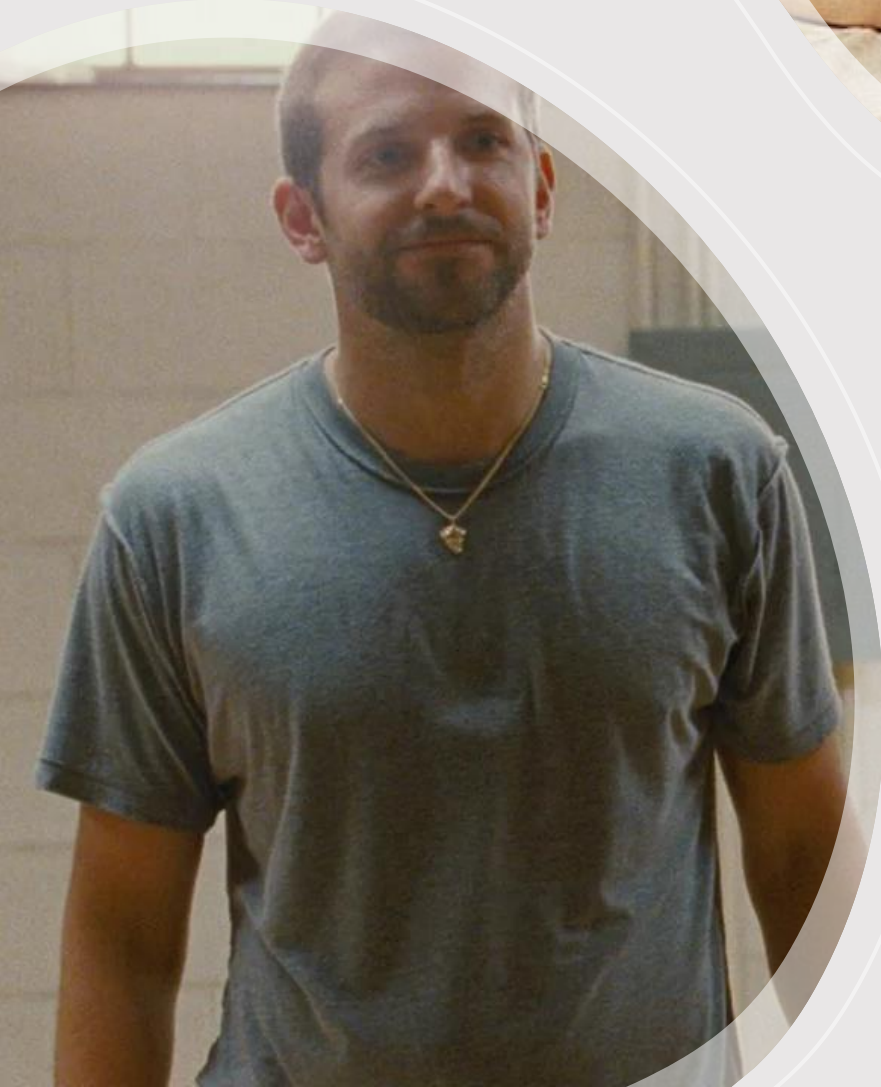
## Possible Applications!

**You can understand better the reaction from captured scenes!**

That can lead to:

- To improve advertises.
- To understand reactions through real-world captured conflicts.
- Predict movies spectators' reactions.
- To better link with images in newspapers.
- And many more...





## Beyond the meaning: Images as feeling carrier

The face expression is most obvious attribute for emotion recognition for images.

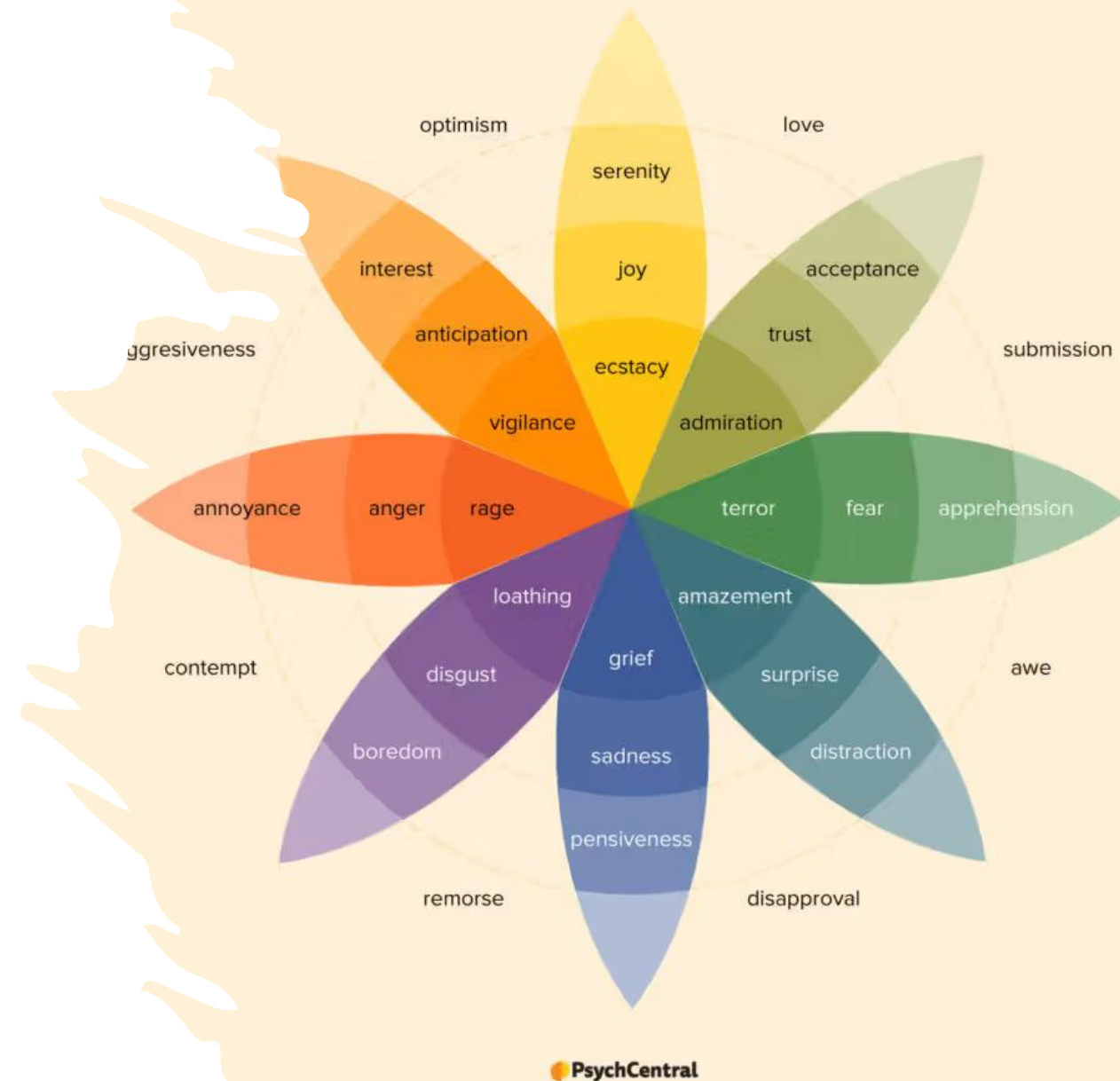
But could we find more?

Quoting former  
USA president  
Barack Obama:

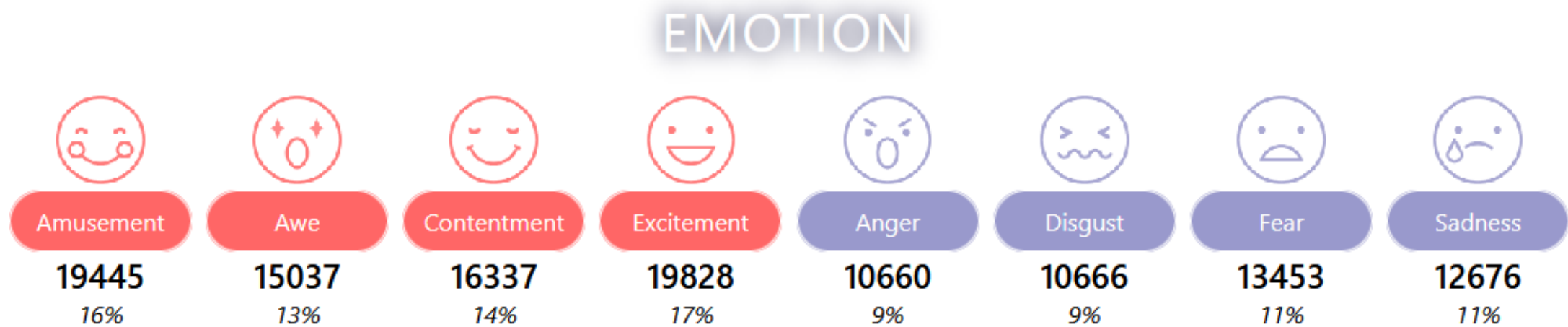
“Yes, We Can!”

Let’s play a game!

## Plutchik’s Wheel of Emotion



Choose one for the next images!





fear  
sadness  
disgust  
awe  
contentment  
anger  
amusement  
excitement





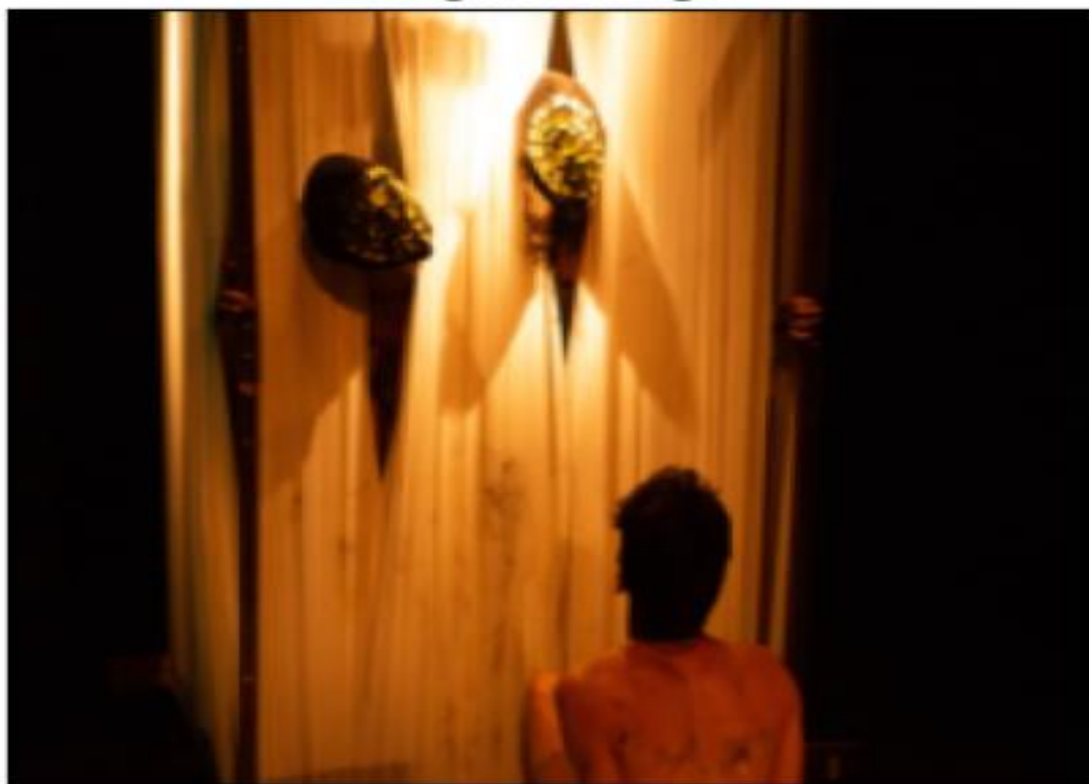
■ fear  
sadness  
disgust  
awe  
contentment  
anger  
amusement  
excitement



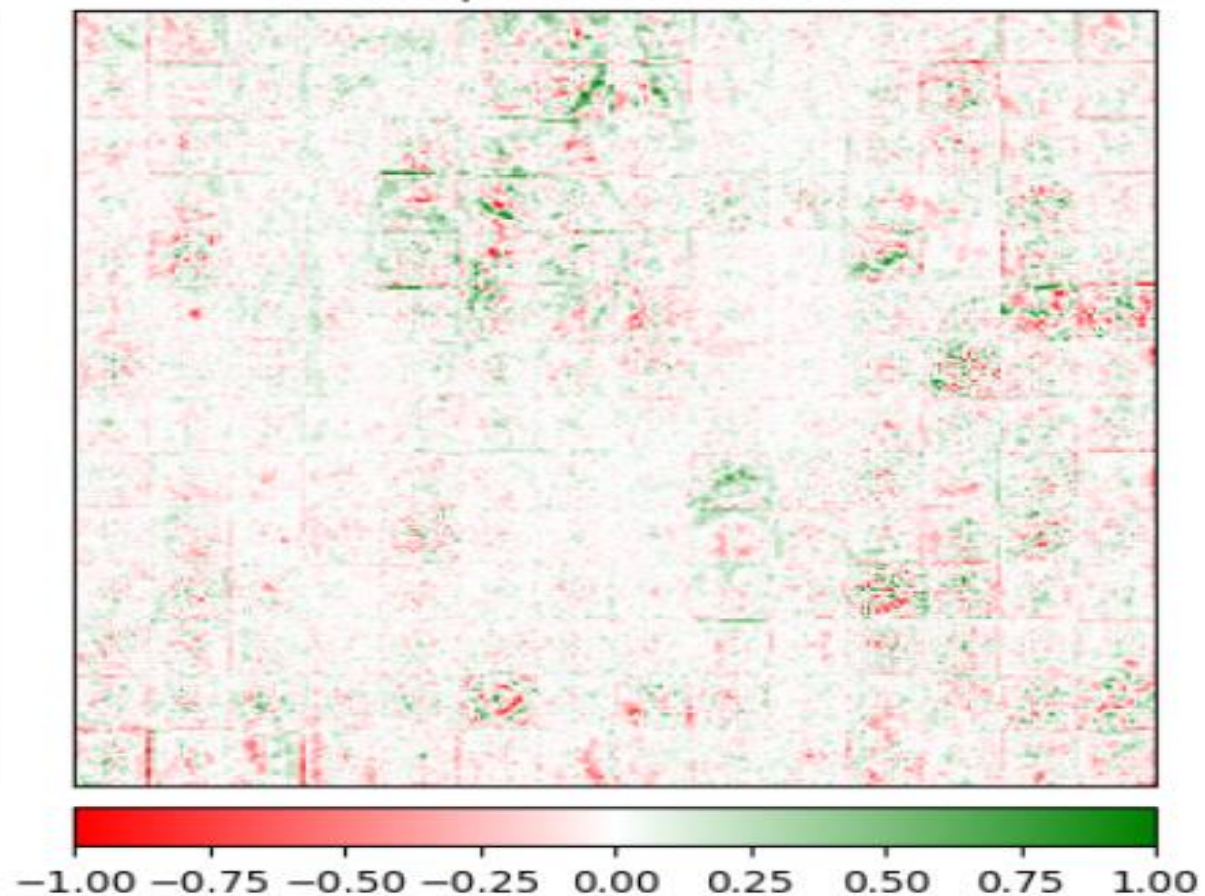


# The explanation:

Original Image



Heatmap IG. Prediction: fear



awe  
contentment  
sadness  
fear  
amusement  
anger  
excitement  
disgust





☒ awe  
☐ contentment  
☐ sadness  
☐ fear  
☐ amusement  
☐ anger  
☐ excitement  
☐ disgust



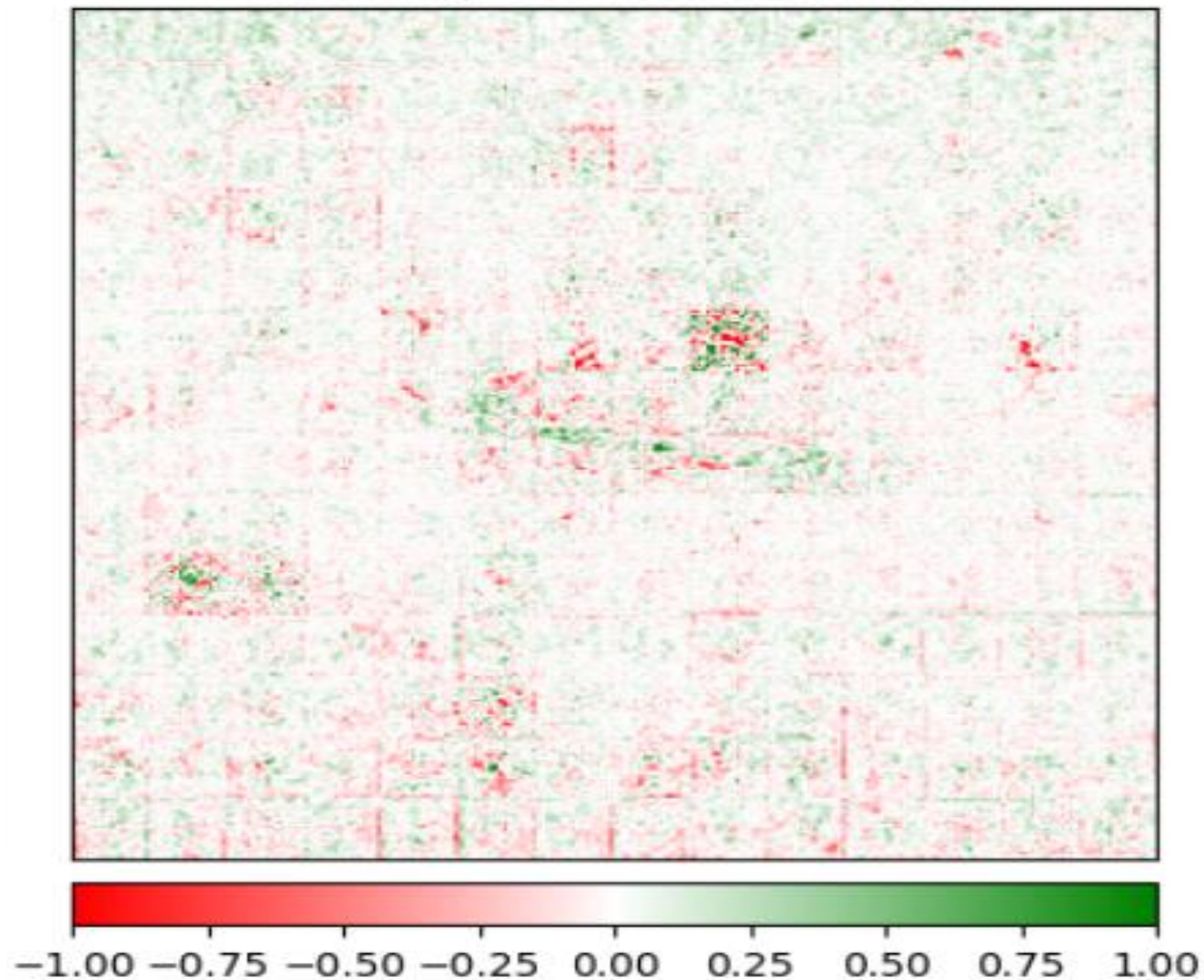


# The explanation:

Original Image



Heatmap IG. Prediction: awe



But...



anger  
fear  
amusement  
sadness  
awe  
excitement  
contentment  
disgust



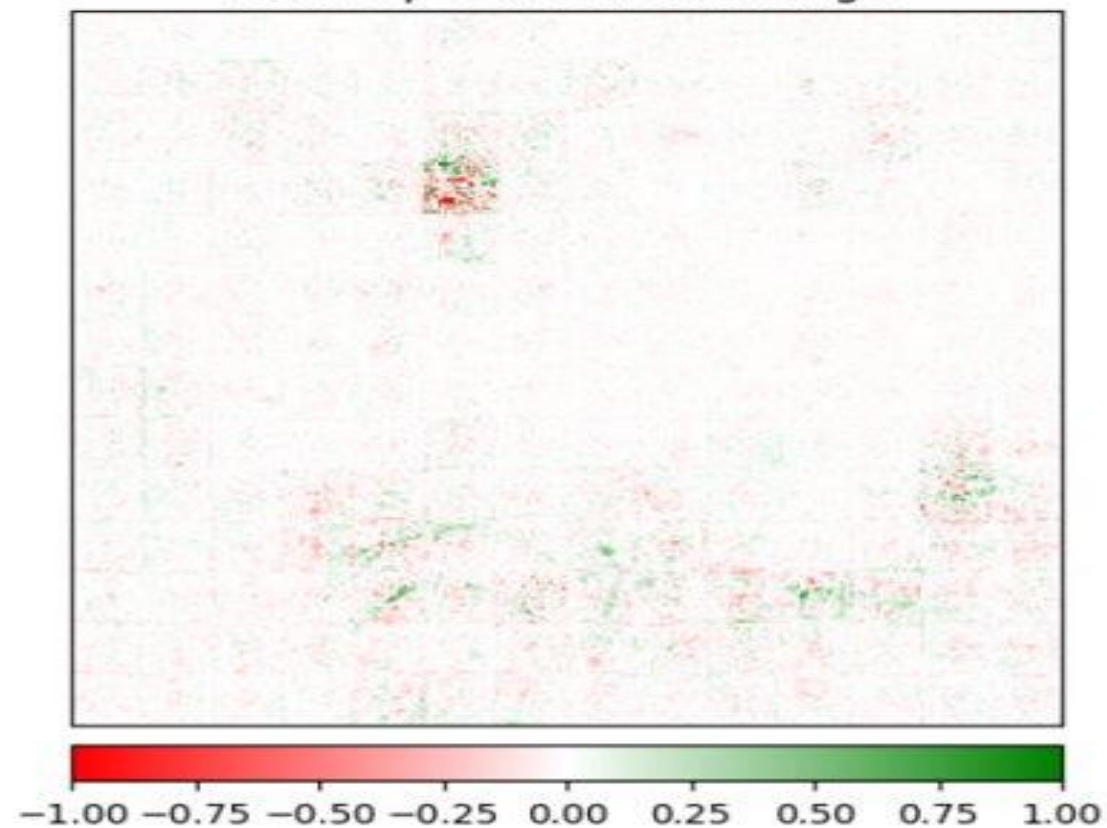


# The explanation:

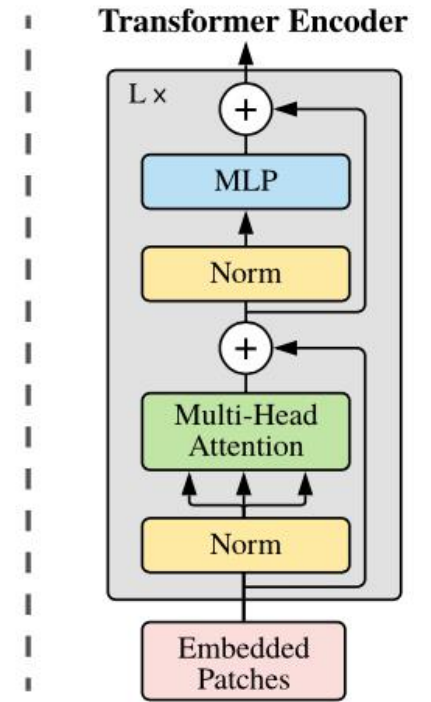
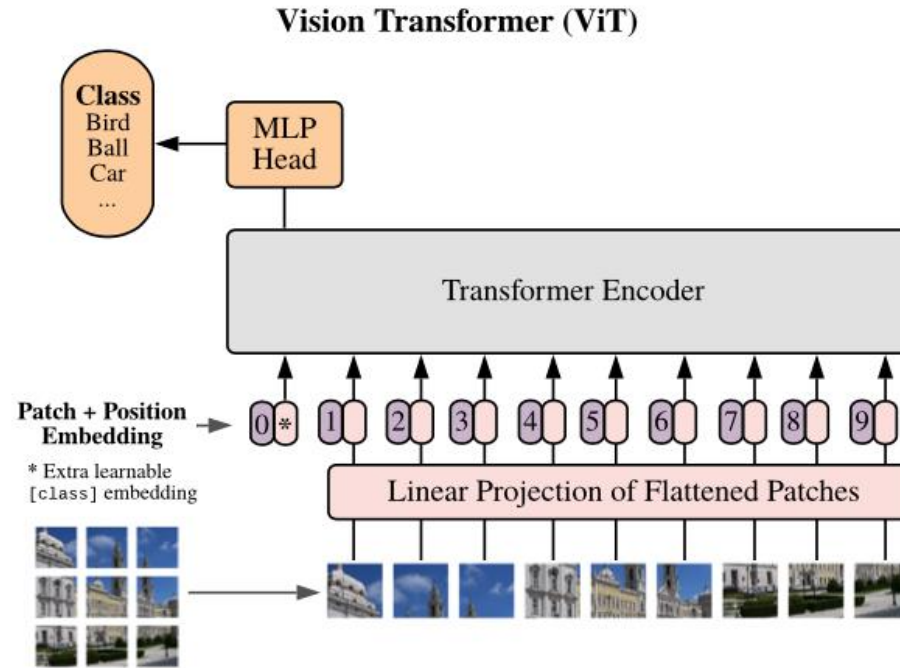
Original Image



Heatmap IG. Prediction: anger

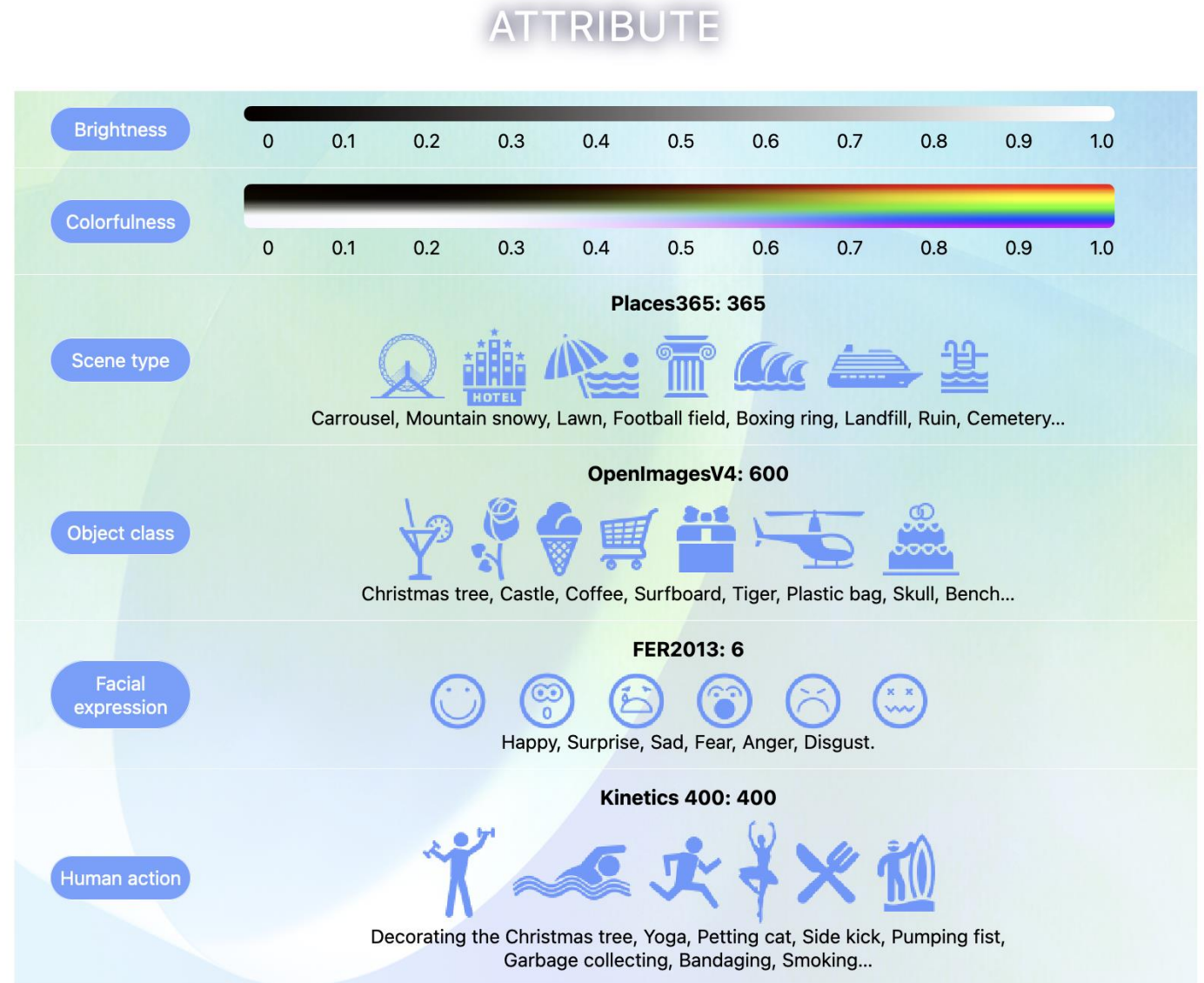


Using Vision Transformers to achieve our goals!



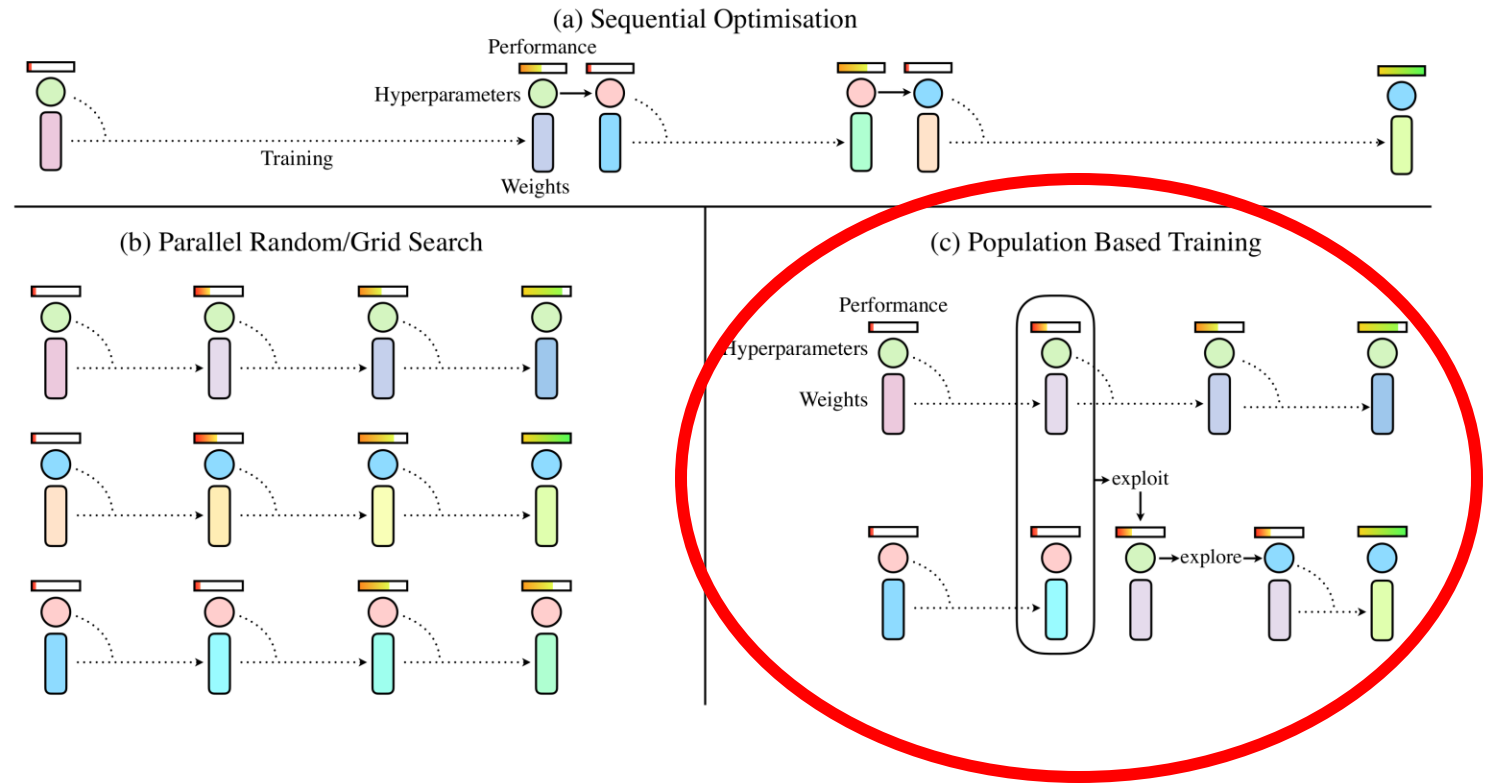
# EmoSet: Images with feelings as label.

---





# Hyper Parameter Tuning: Why PBT?



- Faster
  - Best results
- In many benchmarks

+

•

○

# Image processing: How to?

The best practices are employed:

- Normalized!
- With random crops!

# The benchmark:

- 78.40 of accuracy  
(based on the dataset paper)

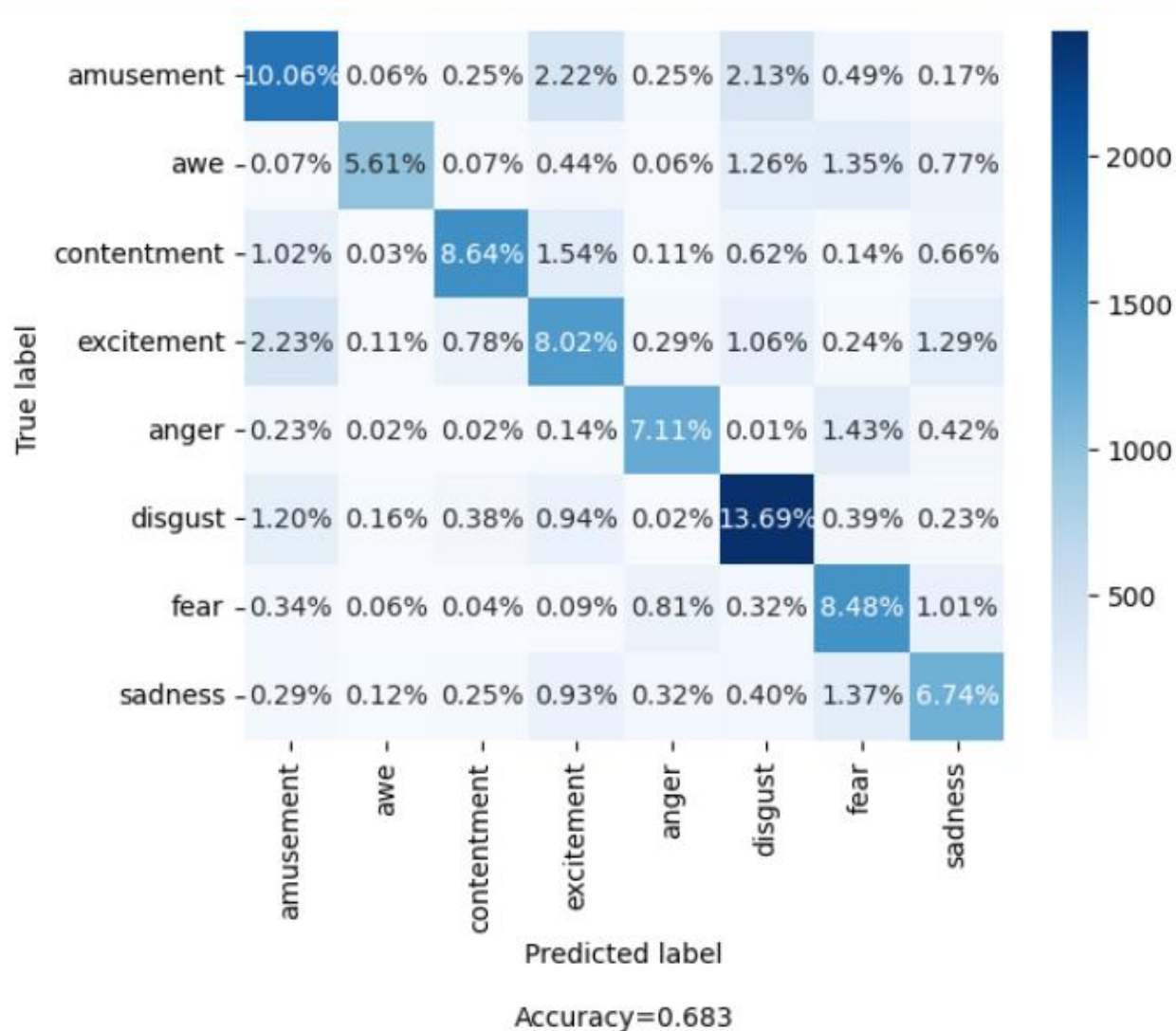
Method	Twitter I-2	Twitter II-2	Flickr-2	Instagram-2	Emotion6-6	FI-8	EmoSet-2	EmoSet-8
AlexNet [21]	75.20	75.63	79.73	77.29	44.19	59.85	<b>89.28</b>	<b>67.80</b>
VGG-16 [43]	78.35	77.31	80.75	78.72	49.75	65.52	<b>93.40</b>	<b>72.27</b>
ResNet-50 [15]	79.53	78.15	82.73	81.45	52.27	67.53	<b>93.48</b>	<b>74.04</b>
DenseNet-121 [16]	80.71	78.99	84.87	83.76	53.79	67.24	<b>92.92</b>	<b>72.32</b>
WSCNet [53]	84.25	81.35	81.36	81.81	58.25	70.07	<b>94.16</b>	<b>76.32</b>
StyleNet [59]	81.50	80.67	85.02	84.53	59.60	68.85	<b>93.93</b>	<b>77.11</b>
PDANet [61]	80.71	77.31	85.36	83.80	59.34	68.05	<b>94.01</b>	<b>76.95</b>
Stimuli-aware [52]	82.28	79.83	85.64	84.90	61.62	72.42	<b>94.58</b>	<b>78.40</b>
MDAN [48]	80.24	83.05	84.26	83.52	61.66	76.41	<b>93.71</b>	<b>75.75</b>



# Our first results:

---

68.3 of accuracy  
Without hyperparameter  
tuning and 1 epoch

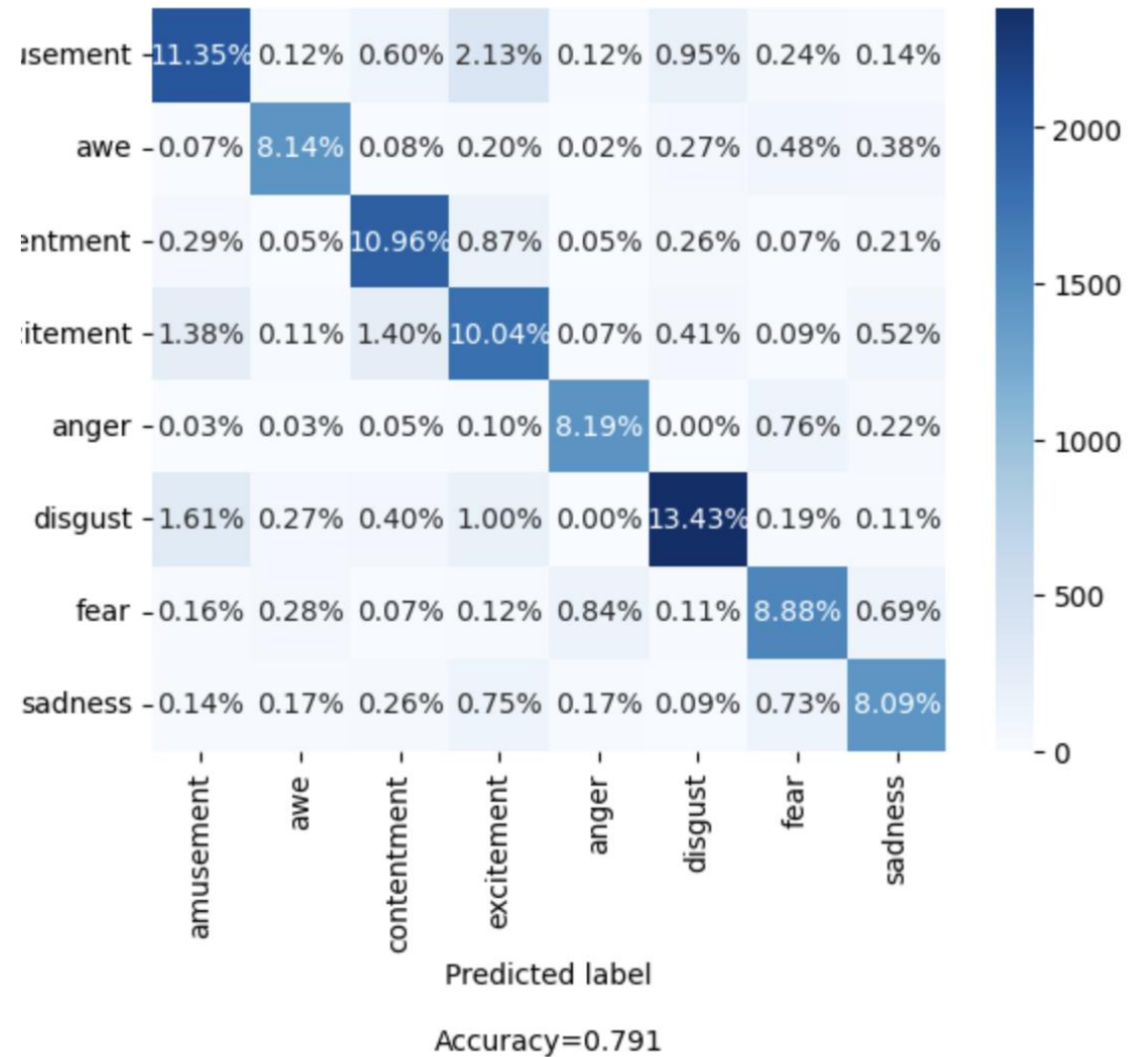


# Our middle results:

79.1 of accuracy

With hyperparameter tuning  
and 14 epoch

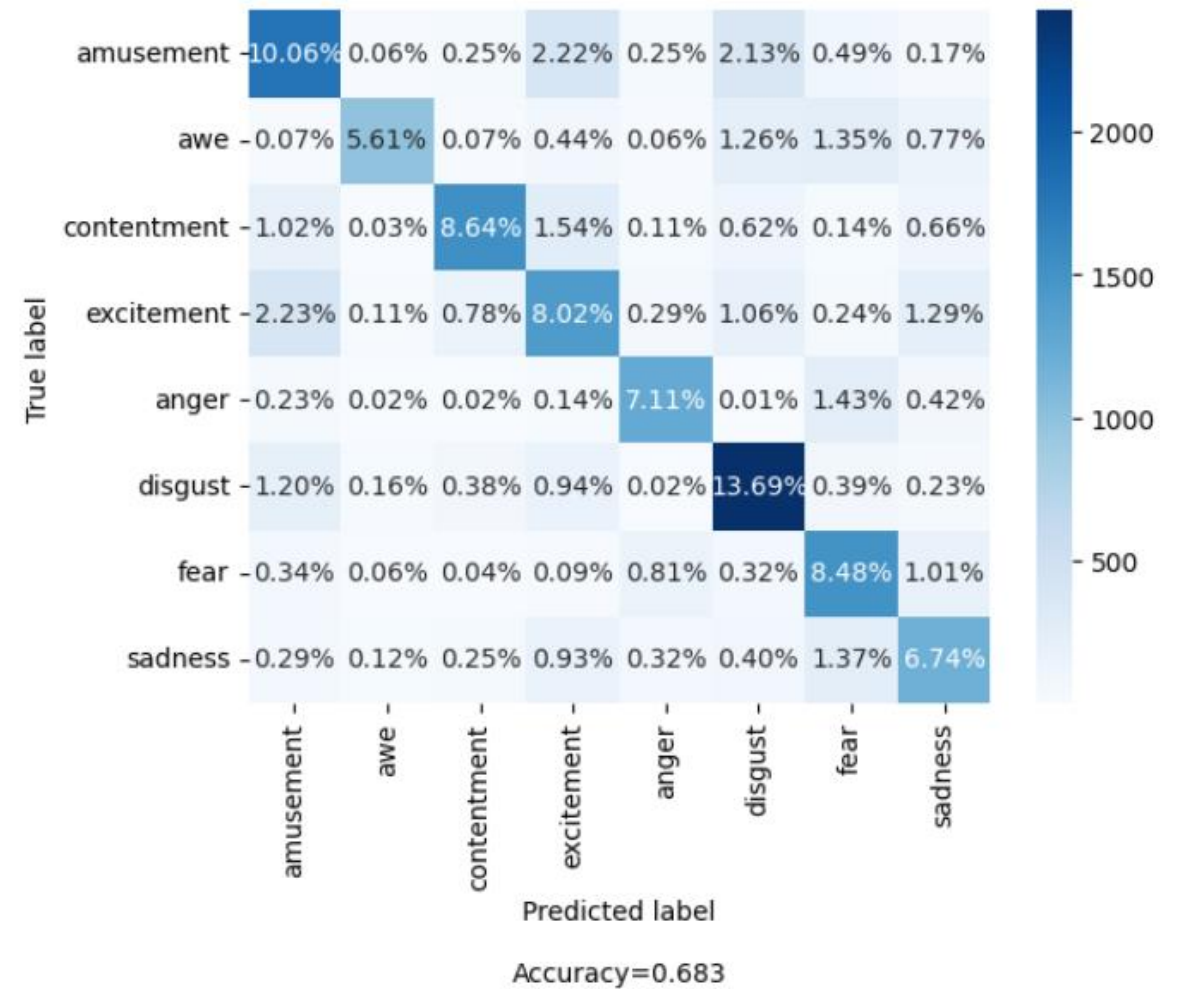
Better than the benchmark  
of 78.4!



# Our final results:

68.3 of accuracy

Without hyperparameter  
tuning and 1 epoch



+

•

○


# Visual Results:

- 01:26:30 – 01:27:00
- [Link](#)






# Conclusion – Good ones.

- (Really) good final results related with the best benchmarks.
  - Took care about all data processing.
  - Learned about more modern techniques of hyper parameter tuning.
  - A deeper understanding about transformers limitation of input data.
- 



## Conclusion — Bad Ones.

- After some tries to include categorical data using a multimodal transformer, was decide to focus only on the image data. (More tries?)
  - A video transformer could be also a option, but we didn't found any open dataset for this. (Data scriping maybe?)
  - Powerfull GPU's are a game changer (and really expensive).
  - Trying more transformers archititures (or deep learning in the general) could lead to better final results.
- 



*So Long, and  
Thanks for  
All the Fish*

---

Wait... That's another  
project!







# Bibliography

- <https://vcc.tech/EmoSet>
- <https://arxiv.org/abs/2010.11929>
- [https://huggingface.co/docs/transformers/model\\_doc/vit](https://huggingface.co/docs/transformers/model_doc/vit)
- <https://docs.ray.io/en/latest/tune/index.html>
- [https://www.youtube.com/watch?v=YAgjfMR9R\\_M](https://www.youtube.com/watch?v=YAgjfMR9R_M)
- <https://www.youtube.com/watch?v=dUzLD91Sj-o>
- <https://towardsdatascience.com/introducing-transformers-interpret-explainable-ai-for-transformers-890a403a9470>