École Polytechnique Fédérale de Lausanne

Predicting emotional response to visual stimuli, a machine learning approach.

by Robin Szymczak

Master Thesis

Prof. Süsstrunk Sabine
Thesis Advisor

Prof. Baroni Raphaël
External Expert

Aydemir Bahar, Pajouheshgar Ehsan
Thesis Supervisor

EPFL IVRL
CH-1015 Lausanne

January 28, 2022

Art is a lie that makes us realize the truth...
— Pablo Picasso

Abstract

We created an emotion predicting model capable of predicting emotions in images using OpenAI CLIP as backbone. Using the ArtEmis dataset which contains 80K paintings annotated on the base of perceived emotions (amusement, fear, etc..). We show that this method of predicting emotion is effective, outperforming previous methods in predicting dominant emotion (70% vs 60%) or positive/negative images (80% vs 77.7%). We leverage our method on text emotion prediction which allows to quickly identify the reason why an image has a certain affect. With this method we unveil CLIP's political preferences, and discovered a lean towards the democrats. We also created a affect based query tool that allows the users to search a set of pictures with a prompt as well as an emotion. The code for the data exploration, emotion encoder and query tool are publicly available : https://github.com/robinszym/EmotionPredictor.

Contents

Figure 1: **Example of emotion prediction with our method**

Chapter 1

Introduction

The research topic of this thesis came about from a request of journalists to have the possibility to perform searches on texts and images. In order to find a corresponding image for an article and article corresponding to an image. Within this broad problem highly related to two fields of machine learning, namely nlp (natural language processing) and computer vision. The affective search, the search with an emphasize on emotional inputs rather than object, is an under studied field and could be beneficial to many other fields. In psychological research, to find images containing a specific stimuli for psychological experiment. In Art history, by performing data intensive research on different art corpora. In design and photo editing, having a tool to predict the emotion of an image can guide the conception of a logo or the filters to apply to an image. In stock photography search engines, to find images with specific emotional response. And finally to general users, who has not wondered what is the most exciting or saddest picture in his phone?

Yet, usual search engine as well as deep learning efforts are focused on image content recognition rather than high level concepts. Emotion is a complex topic and making a predictive tool on emotion sparks the question on how an artificial intelligence can understand or conceive emotions. It is not a deep understanding per se. A human being with a defect in a region of the brain called the amygdala[2] can become enable to feel any fear. This human is yet still perfectly capable of identifying scary images or situations, and even flee appropriately, just not to experience fear like other humans[18]. Artificial intelligence will behave similarly, learning to predict what triggers an emotion, but not how to experience it. Emotions are highly personal and carry a lot of subjectivity, which poses a challenge to find good labeled dataset. Annotations might vary and confusion often rises when asked how one feels about an image. Some emotion dataset with pictures labeled as *anger* inducing are made entirely of angry looking persons[14] which are more scary than infuriating. For this reason other psychological emotional representations such as the circumplex model of affect, uses a continuous emotion representation rather than a discrete one. With two scales, one of valence, going from highly negative to highly positive. And one of arousal representing the excitation/agitation one feels. Other name for the dimension of the continuous emotional space have been proposed with for example tension and energy[24],

approach and withdrawal[13].

To predict the emotional effect of a picture on a group of individuals. Previous work have described pictures using visual features[20] from computer vision, such as the fourier transform, colourfulness as well as artistic features[26] inspired by art theory such as texture and form. But deep learning has a tendency to outperform feature based methods[15]. The learning based methods[22] have also participated in the affective race but the lack of high quality dataset has been the limiting factor.

The ArtEmis dataset released in 2021 is a novelty in that it is a large (eighty thousand images) publicly available dataset, annotated and designed for deep learning purposes. The dataset contains art work, mostly paintings and drawings ranging from the 13^{th} century to the 20^{th} century, each image is described with a basic emotion[5] and a short description. Dataset with similar annotations are oriented for psychological studies[9]. They are gathered with the objective of triggering a specific emotion on the public and are not large enough to perform deep learning.

Open-AI released in 2021 a new model called CLIP[19] (Contrastive Language-Image Pre-training). CLIP has impressive abilities to match text and images, attaining state of the art performances on dataset it was not previously trained on. With the release of CLIP and the dataset ArtEmis it is a natural step to bridge the two and create an emotion predicting model. The paintings from the different art periods carry emotional meaning in the form of basic features like brightness and colors, as well as high level concepts like scene composed of real, mythical and imaginary subjects. We also aim at understanding which from the low level or high level features have the most impact on emotional perception. We answer this question by opening a neural network. Each layer of a network can be understood as a step in comprehension. Like an image forming in the brain distills more and more information. We isolate each layer and assess their emotional predictability. The results indicates that the *disgust* category requires more cognitive capacities than the others.

Since CLIP's feature space is shared by both image and text we can translate them to emotion by only training on one of them. We only train the translator using image features and we are able to make emotional prediction on text as well (40% accuracy on artemis test set against 60% for humans). On Images we achieve 70% accuracy on the artemis test set, the ResNet50 pre trained on ImageNet and fine tuned on ArtEmis achieves 60%. We achieve 80% accuracy on the ArtPhoto (positive or negative), which was previously at 77.8%[22] . The text translation can be used to understand what word or sentence sparks what emotion and we surprisingly discovered that CLIP has a political opinion. Indeed by giving the text prompt "Donald trump", the model reacted with *disgust, anger,* and *amusement,* whereas for "Barack Obama" the model outputs with more than 80% confidence *contentment*. The same is observed by with the words "republican" and "democrats". This test-emotion translator allows to quickly understand why the model reacts the way it does by directly inputting the suspected word. For example a picture with a pile of cucumber stacked on a table had a surprisingly strong *disgust* output, imputing the text "cucumber" in the translator we realised it had an aversion for the cucurbit, and to food

in general.

By merging clip and an affective art dataset we have created a tool with various application such as affective search. Large scale affective evaluation and provided an extra tool to quickly understand the link between a word and his corresponding emotional perception by CLIP.

Figure 1.1: **Example of emotion based search with the word "landscape" using our method.**

Chapter 2

Background

2.1 CLIP - Contrastive Language-Image Pre-training[19]

CLIP is a recent powerful model that came out in 2021 as a new multi modal model capable of handling text and image data. Its design aims for generality to perform state of the art zero-shot predictions. Zero shooting stands for the task of predicting class labels without training on them. For example when performing zero shot on imagenet, the 1000 classes are encoded using CLIP's text encoder. The encoded classess are compared with the encoding of an input image using cosine similarity. The resulting similarities are softmaxed and results in a probability for each class label. Using zero shooting, CLIP beats existing model trained for a task. CLIP training uses an impressive 400M image-text pair scraped from the web, in total 32 datasets were combined to create the training set.

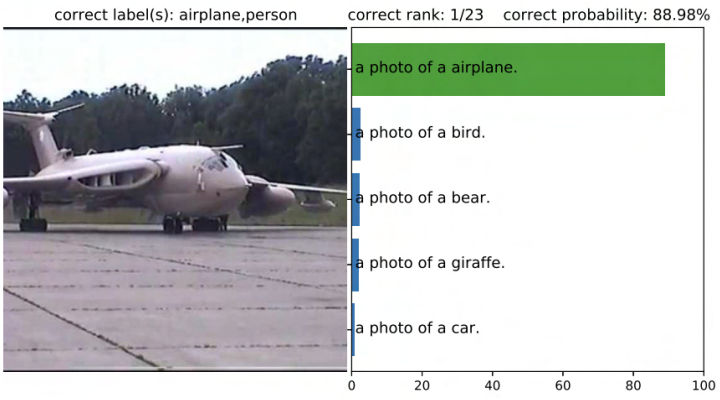


Figure 2.1: **Example of successful zero-shot prediction**, taken from CLIP's original paper[19]

2.2 Training

CLIP's novelty comes in part from their training process that we briefly expose here. CLIP is trained using contrastive learning, instead of accurately predicting an image-text pair the objective is to maximise the distance with the other pairs of the same batch (Figure 2.2). They gathered 32 different datasets containing images and text. The loss is computing by (highlighted in blue in the diagonal) while simultaneously maximising the similarities with the other. This allowed them to efficiently train on the huge dataset at a "reasonable" cost. The training still took between 13 and 18 days on 250-500 GPUs depending on the architecture. They trained on both transformers as well as convolution nets (ViTs and ResNets).

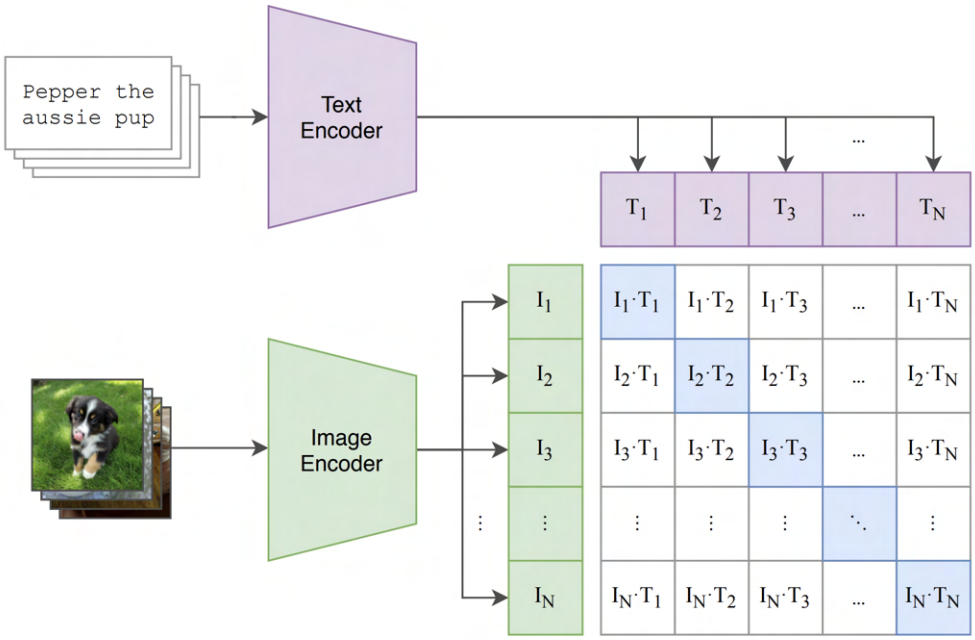


Figure 2.2: **CLIP training process**. The matrix is generated by input images as well as text tokens. The loss is computed based on the difference of diagonal (blue) with the rest of the matrix. Rather than minimizing of the loss of matches, it maximises the difference with respect to the other cells of the matrix. Taken from the original paper[19]

.

2.3 Breakthroughs

By releasing this powerful model the team allowed a leap forward in different areas of computer vision. In visual art generation : by generating paintings with a given emotional prompt AffectGan generated paintings [8], CLIP-guided GAN created realistic pictures [21] and CLIPdraw generates drawings [7], all optimize the similarity between a text prompt and the generated image both encoded in CLIP's feature space. CLIP also helped in video captioning [16] and generation[6]. The

latent of clip was explored and disentangled in [17] and the author note the "the extraordinary visual concept encoding abilities of CLIP", some results can be seen in Figure 2.3. The ALIGN[11] team claims to have obtained better results than CLIP, using a similar training but on a much a bigger dataset (2 billion samples). Which indicates the general direction of the field and the efficiency of the method. They unfortunately have not disclosed their model.

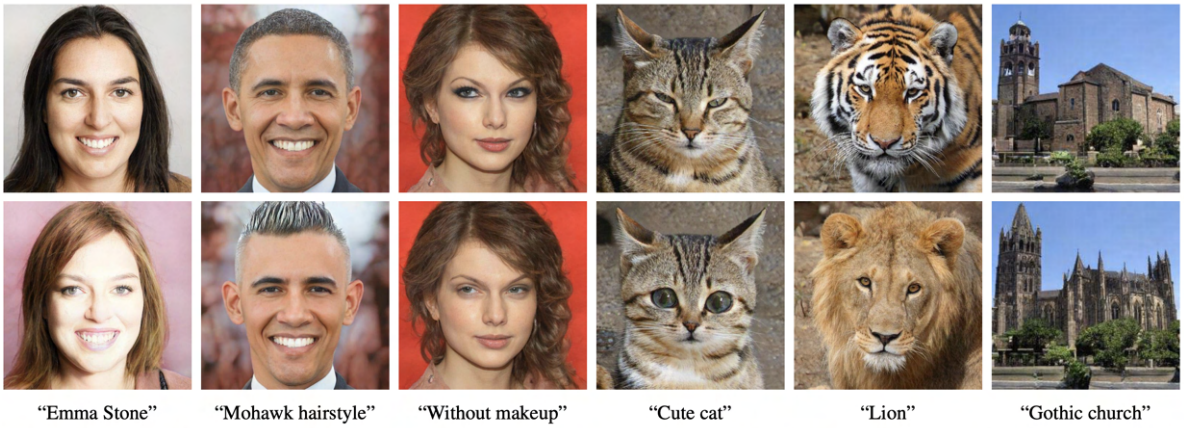


Figure 1. Examples of text-driven manipulations using StyleCLIP. Top row: input images; Bottom row: our manipulated results. The text prompt used to drive each manipulation appears under each column.

Figure 2.3: Example of the effectiveness of a disentangled clip latent space, this figure is directly taken from the styleclip paper [17]

2.4 Limitations

CLIPs training was performed on data taken from the internet, and with such data comes its share of biases. It is therefor not surprising to have a gender bias as well as racial bias, and a dedicated section on the subject is discussed in the original paper. An other problem found by a fellow lab member is the incapacity of clip text encoder to identify relationship with noun and color adjective. As shown in Figure 2.4 the predictor fails to understand the concept of a red shirt, (we tried several prompts with similar results). So a surprising limitation of clip is the communication with model.

2.5 Interest

Clip offers a new powerful model trained on noisy web data enabling new representations differing from models trained for object recognition. Impressively, this model outperforms models specifically trained for classification. Nevertheless limitations exists, indeed CLIP speaks its own language. Even though it could have the capacities to categorize images the intended way. The accuracy greatly depends on the quality of the text tokens, especially for abstract concepts.

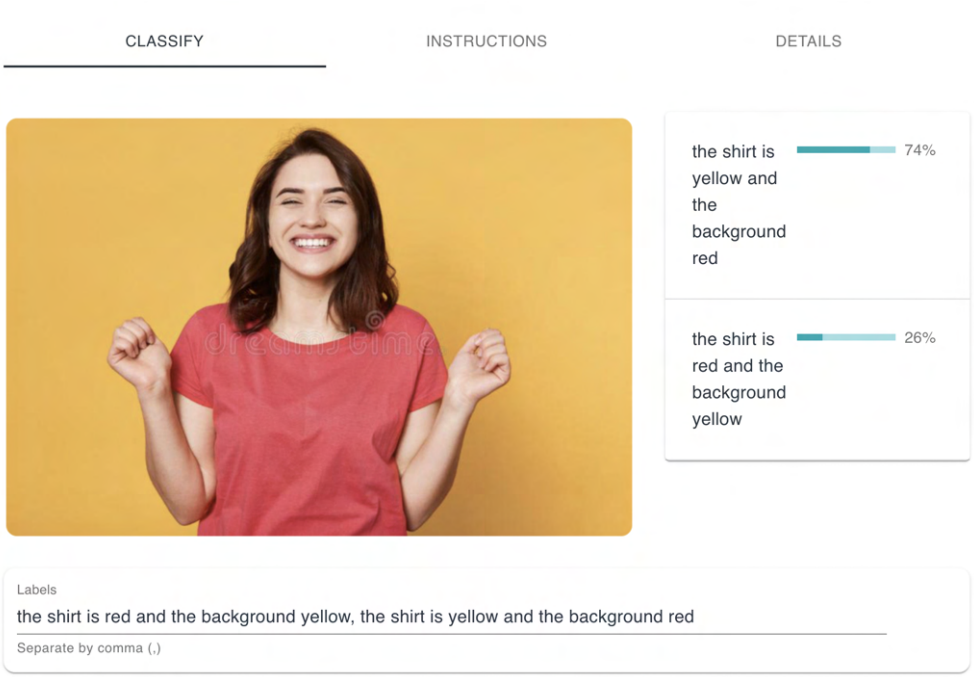


Figure 2.4: **CLIP failing** to link the colors to their corresponding words.

This observation is the motivation for the method employed in this thesis. That is, training a model to find in CLIP's latent space the directions corresponding to the perceived emotions.

Chapter 3

ArtEmis Dataset

3.1 ARTEMIS

The ArtEmis dataset [1] is a collection of emotional reaction of annotators towards art pieces found in the WikiArt[1] collection. It ranges from the 14^{th} to the 20^{th} century. The annotators were asked to select the emotion they felt looking at an art piece from nine possibilities : *amusement, awe, contentment, excitement, anger, disgust, fear, sadness,* and a *something else* category and to provide a small corresponding affective description (Figure 3.1). The dataset collected 455K emotional reactions and descriptions towards 80K paintings. Most paintings were annotated by 5 or 6 annotators (75% by 5, and 96% by 5 or 6 annotators (Figure 3.2). Having a small set of annotators can lead to oversimplification, emotion histogram associated to paintings could fail to capture their true overall responses[2]. The next sections identify the limits of small sets of annotators by comparing the dataset with a reference subset. .

3.2 Accurate reference subset

The accurate subset is the subset of the 703 paintings (0.9% of the dataset) that received more than 40 annotations. It has the same overall distribution as the whole dataset. The distributions can be considered closer to a real distribution of affects. Figure 3.4 shows an example of a very sparse distribution attainable only with enough annotators. This subset is still only an approximation of the real effect of a painting, but is the best one available. This set is used as reference for computing entropy in section 3.3, and as approximation for underlying distributions

[1] https://www.wikiart.org/

[2] The true overall response, or underlying distribution of painting can be defined in two different ways: an artwork annotated as 60% fear inducing and 40% awe, scares 60% of the population and impresses the rest. Or evokes both emotions simultaneously, like a raging volcano evokes awe and fear. An ongoing debate in neuroscience about emotions in the brain, whether emotions each form in a separate circuit or if they originate from a common one.[3]

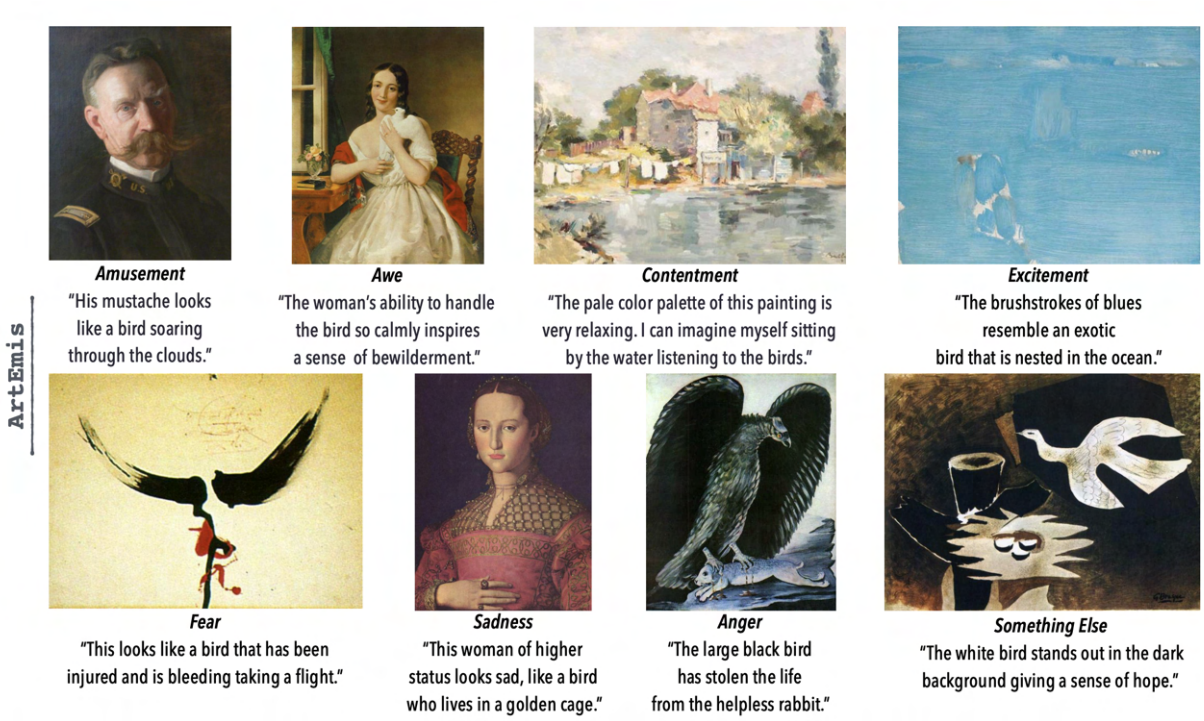


Figure 3.1: **Example** taken from the ArtEmis paper[1]. Each painting is annotated by at least five annotators. They entered the dominant emotion they felt between 9 categories (*disgust* is missing in this example) and provided a small explanatory sentence.

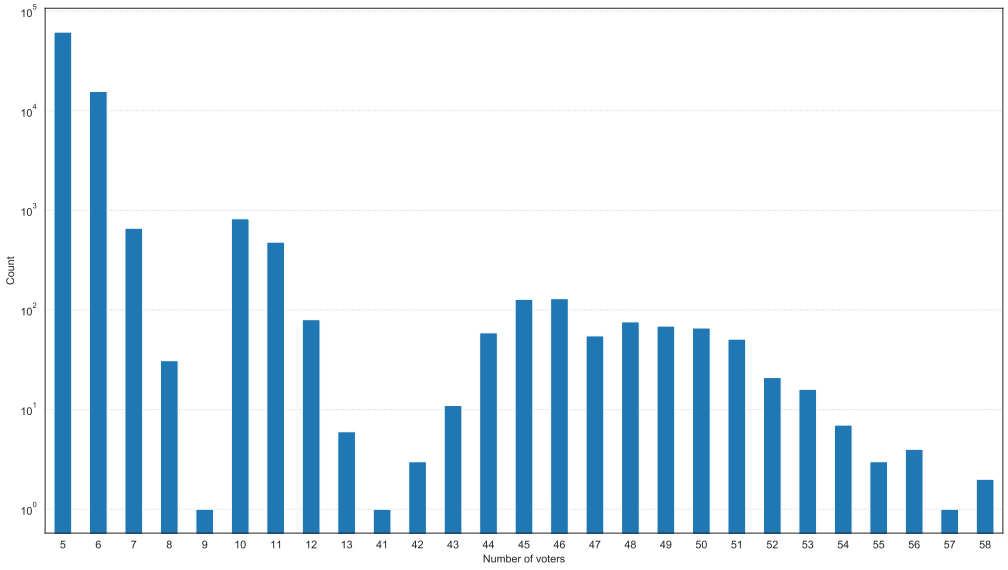


Figure 3.2: **Count of the number of painting per number of voters.** The y axis is in log scale, 96% of the data is annotated by 5 or 6 annotators

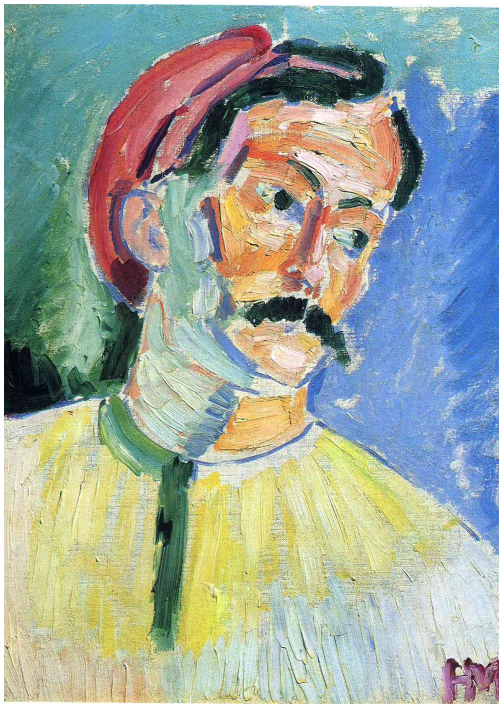
in section 3.4.

3.2.1 Overall distribution

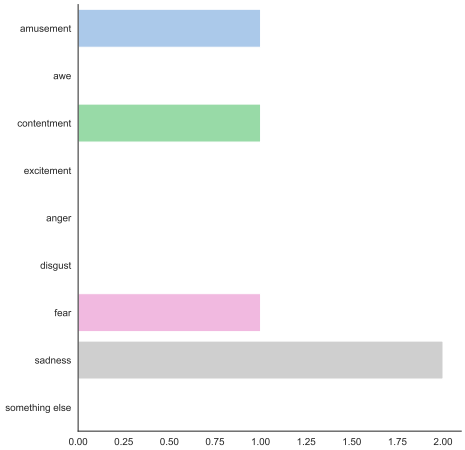
The compilation of all the votes (see Figure 3.5) indicates a high bias towards *contentment* and an under representation of *anger*. We hypothesis that this bias could be explained by the implicit objective of art to be appeasing and peaceful. An other factor is the time period of the art pieces. Indeed with time, subversive art becomes accepted[3] [4, p.295 302]. Any model trained on labeled art datasets will inherit the biases of the historical period of their making. An other possible explanation for the high level of *contentment* might be the absence of the null option or *no emotion*, which supposedly falls in the *something else* category which is ambiguous as it combines the *no emotion* category as well as *another emotion* category. To test this hypothesis a research with a more complex emotional representation would have to be used like the one[4] proposed by Hagtvedt et al.[10]..

[3] Bourdieu gives the example of the Impressionist movement that was vividly criticised at its debut in the 1870s, and ironically is the movement that produces the most *contentment* votes in the dataset. Whereas Analytical Cubism and Action Painting spark the most anger and disgust

[4] A model combining 17 emotions (e.g : Despair, Anxiety, etc..) and 19 perceived attributes (e.g elegant, symmetrical, etc..) to evaluate ones response towards an art piece.



(a) Portrait of andre derain by Henri Matisse 1905



(b) Corresponding histogram

Figure 3.3: **ArtEmis Example** taken from the dataset, two annotators reacted with sadness, the three others for amusement, contentment and fear. It illustrates the possible different reactions towards the same art piece.

3.2.2 Agreement

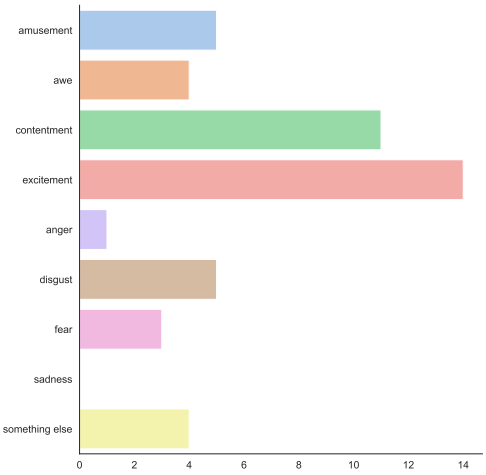
To visualize how much annotators agree with each other we take for each painting the maximum agreement. Since most painting are annotated by 5 annotators it results in increments of 0.2, which is not fine grained enough for good distribution approximation. By only taking pictures from the accurate subset (section 3.2), the agreement distribution results in a Gaussian centered at 40.06% (see figure 3.6), which is in accordance with the supplementary paper of ArtEmis, where they asked annotators to rate the ArtEmis annotations and found an overall strong agreement of 46.5%.[5].

3.2.3 Feeling different

To test if one emotion is more likely to be agreed upon, we compare the observed probability of having more than one vote per emotion knowing that one vote is already in this category (Equation 3.1) versus its expected probability (Equation 3.2) (by posing that annotators annotate

[5]Strong agreement means that annotators would have reacted with the same emotion, as opposed to the 51% weak accept where they understand why someone could feel this way while not feeling it themselves.

(a) Abstract landscape by Audrey Flack



(b) Corresponding histogram

Figure 3.4: **A painting with more than 40 annotations** , one vote per annotator.

randomly following the overall distribution of Figure 3.5). The greater the difference between the expected and observed value, the greater the agreement between annotators on an emotion.[6]

$$p_i = P_i(n > 1 \mid n >= 1) \tag{3.1}$$

$$E(p_i) = 1 - (1 - P(e_i)^{N-1}) \tag{3.2}$$

Where p_i is the probability of having more than one vote for an emotion i knowing that someone voted for it. N stands for the total number of annotator for an image, n for a number of annotator between 0 and N, E is the expectation, and $P(e_i)$ is the probability of picking the i^{th} emotion, and i is the emotion index ranging from 1 to 9. We compute $E(p_i)$ by setting $N = 5$ as this is the most common scenario. Negative emotions have a higher relative agreement than positive ones (Figure 3.7). Despite being less numerous, the negative images spark more relative agreement than positive ones. [7].

3.3 Entropy

The entropy of a distribution is defined as :

[6]This aims at testing the subjectivity of voters on each emotions. If voters answered *angry* only on the base of purely personal motives, like the profound aversion towards blue for instance. Then the voters will not agree with each other and the votes should be totally independent.

[7]This finding is linked to the correlations between the different emotions and will be more thoroughly analysed in chapter 4

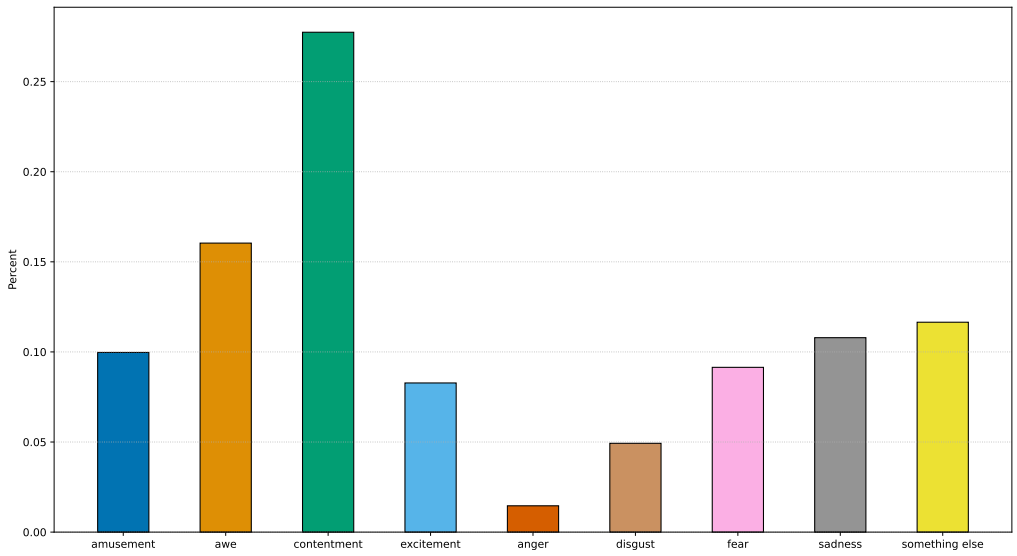
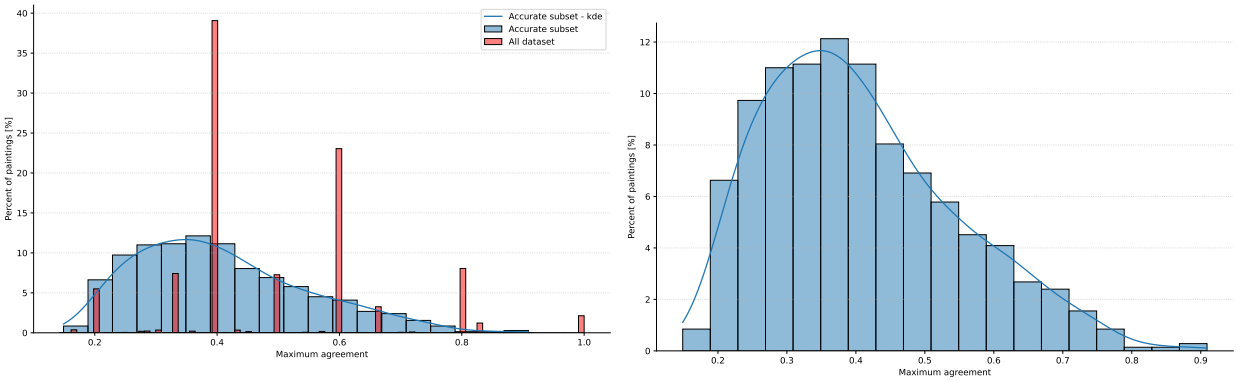


Figure 3.5: **Distribution of all votes**. 62.0% of the votes fall in the four positive emotions (bars 1-4) especially for contentment which represent 27.7%, more than negative emotions combined that account for 26.3% (bars 4-8) the remaining 11.7% fall for the something else category (9^{th} bar).

$$H_d = -\sum_{i=1}^{M=9} p_{i,d} \log(p_{i,d})$$

Where H_d is the entropy of a data point d, $p_{i,d}$ the probability mass of the i^{th} emotion of a data point d. H_d provides an understanding of the compactness of a distribution. An entropy of 0 implies all the mass of a distribution is concentrated on a single point, and maximum entropy implies the mass is equally distributed. A painting with an entropy of 0 sparks only one emotion, maximum entropy means that all labels are equiprobable (zero agreement amongst annotators). So the smaller the entropy of the real underlying distribution the smaller the set of annotators needed to approximate the distribution. If the average entropy of the accurate subset matches the one of the whole subset then we can infer that a few annotators are sufficient. In object recognition, you don't need more than one annotator if a dog is in the picture. Comparing the entropy of the accurate subset with the entropy obtained with 5 or 6 annotators yields that at least 75% of the accurate subset paintings have an higher entropy than what is achievable by 5 annotators (Figure 3.9). Suggesting that 5 annotators do not capture the full complexity of the emotional reaction of a painting. However Having sparse and incomplete training data is not problematic for machine learning[25]. For testing however, the quality of the data is important to correctly assess a model's performance.



(a) Whole dataset and accurate subset

(b) Zoom on accurate subset

Figure 3.6: **Distribution of users maximum agreement**, for all the paintings in red on the left, and for the accurate subset in blue (zoomed on the right). The line is the Kernel Density Estimate (kde) of the accurate subset using Gaussian kernels. The spikes for the whole dataset are coming from the paintings with 5 annotations.

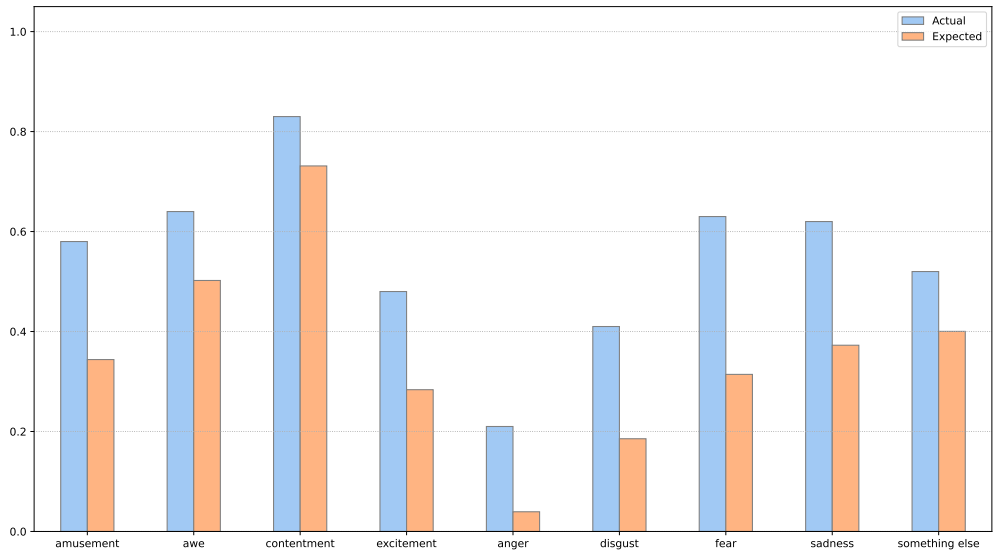


Figure 3.7: **Proportion of grouped annotations by emotion**, left blue bars are the observed and right orange bars are the expected value. The biggest differences are observed for negative emotions.

3.4 Effect of the number of annotators

In this section we propose an experiment to show the confidence one can have in the paintings labels. We sample n emotions from each distribution of the accurate subset, to mimic a set of annotators. We create 100 coarse distributions for each accurate one, so 70'300 samples. To evaluate any model the authors of ArtEmis took the paintings with at least 50% agreement on an emotion and labeled it as the dominant emotion. With this metric we evaluate how many set of generated annotations correctly predict the dominant emotion with a confidence of more than 50%. With $n = 5$ the results are 25% of False positive and 40% of False negative (Figure 3.1). Implying that 25% of our data is falsely labeled as having a dominant emotion[8]. The error decreases logarithmically when increasing the number of annotators (Figure 3.2). To reduce the False positive rate from 25% to 15%, one needs to go from 5 to more than 40 annotators.

	Dominant	Not dominant
Dominant	75.17	38.25
Not dominant	24.83	61.75

Table 3.1: **confusion matrix for n = 5 annotators**. 100 samples of 5 emotions are taken for each painting in the accurate subset. The samples are evaluated with respect to their original distribution. The sample is judged to be correct if he predicts with more than 50% confidence the same dominant emotion as his original distribution.

3.4.1 Implications

The models trained on the ArtEmis dataset will inherit the different cultural and historical biases of the annotators. *Anger* related utterance like "How can this be called Art? laziness" or "A plain, nearly white bent something is boring, and it almost hurts to see something so dumb."[9] is related to the art context. it is safe to assume that few get angry at their white walls when waking up every morning. Yet we expect the model to have a tendency to characterise monochromatic images as annoying the same way some humans do. It is valid in the art context as this is a perfectly plausible reaction. Context matters. Upon generalising to real life images, an ArtEmis trained model will look at everything as if in a museum. However the dataset is constituted by art coming from the past 6 centuries and the different focus point and techniques of the different movement hopefully captures a better understanding of what emotion are on a low and high abstraction level compared to real life pictures. The limited number of annotators is not a problem for the training process, but will limit the evaluation process. In future works one should think of creating an accurate annotated test set with enough annotators to precisely assess future models' performances.

[8] This result does **not** say that the data is badly annotated, simply that out of randomness some paintings are labeled as over or under expressing an emotion. So caution must be taken when using the annotation as ground truth in the test set.

[9] Real sentences found in the dataset.

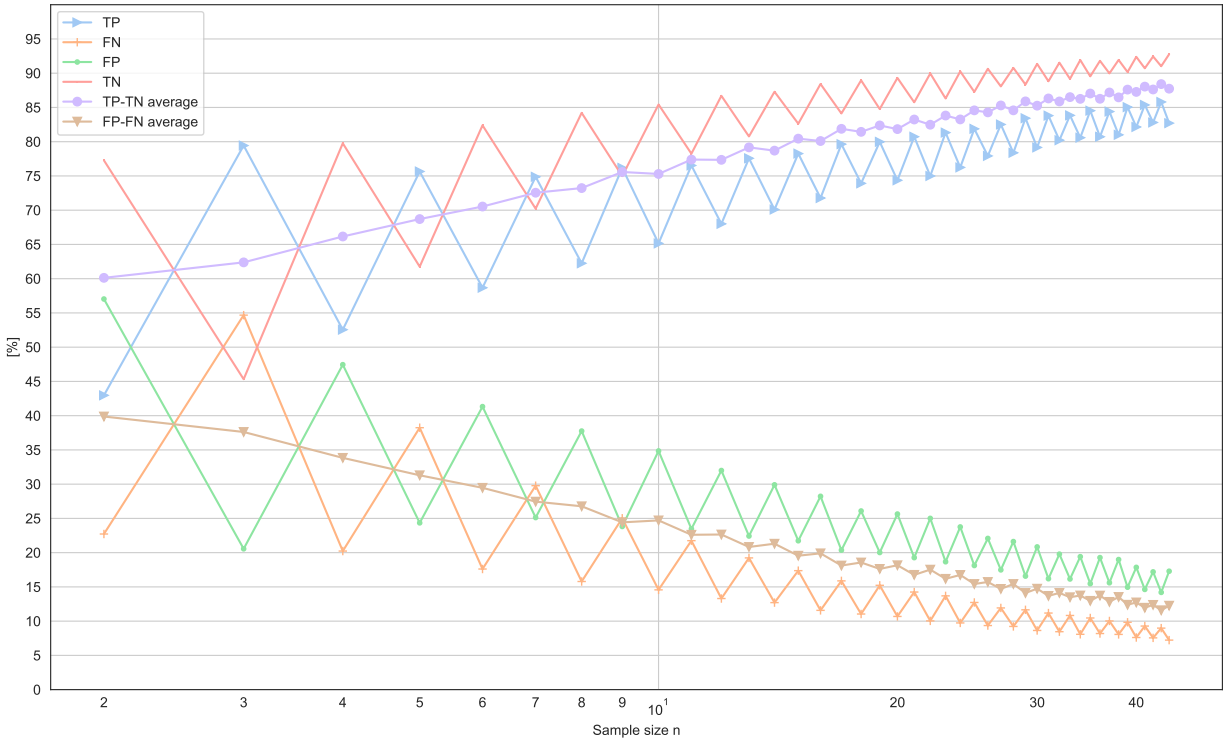


Figure 3.8: **Error as a function of sample size n**. Here the True positive (TP) is the share of correctly predicted dominant emotion. True negative (TN), the share of correct none dominant negative. False positive (FP), the share of wrongly classified as dominant emotion, False negative (FN) the share of missed dominant emotions. An odd number of annotators induces a bias for positive detection and inversely for pair numbers of annotators inducing the saw tooth pattern.

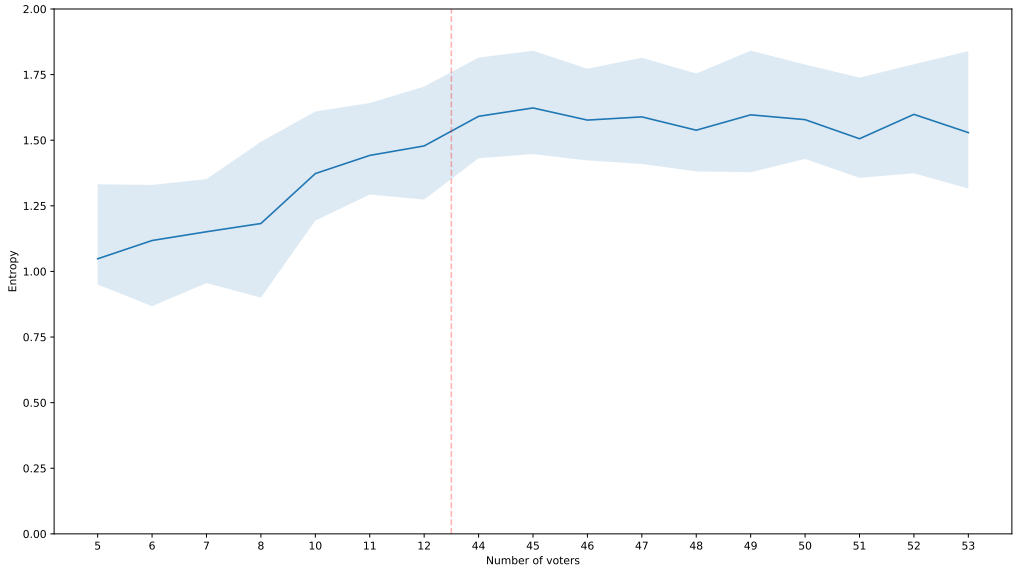


Figure 3.9: **Average entropy** by number of voters. The vertical dotted line indicates the jump from 12 to 44 annotators and the shaded area represents the mass between 25 and 75% of the entropy distribution. After the red doted line it is considered the reference entropy. Comparing the 96% of the data being annotated by 5 or 6 annotators, we realise that most of the dataset fails to capture the entropy of the paintings.

Chapter 4

Making a prediction on emotion responses

In this chapter we present the solution found to identify emotions in CLIPs latent and turn it into an emotional predictor. We start by creating a baseline with CLIP's zero shot capabilities, where no training is required, and further present the fully connected layer and evaluate its performance in various experiments.

4.1 Zero shot baseline

4.1.1 Method

In this experiment we try to determine the zero shot capacities of clip. As seen in chapter 2, CLIP can be leveraged to perform class prediction. By encoding the class names with the text encoder and computing the cosine similarity with the images, the predicted class is the one yielding the highest cosine similarity. The list of classes is given in the form of "an image evoking (one of the nine emotion)", e.g "an image evoking contentment". With those prompts both the "An image of an contempt man" and "A contempt image of a man" are expected to trigger an response.As The first one relates to the content of the image and the later to the effect of the image. ArtEmis has been labeled according to the image effect not its content.

4.1.2 Results

The zero shot mostly predicts *contentment* (Figure 4.1), and *something else* on a smaller scale. Leading to an overall accuracy of 0.48 and an f1 score of 0.16 (Table 4.1). Which leads us to

believe that this implementation of CLIP zero shot emotion detection can not be used for emotion prediction. It is understandable that the model categorises paintings as *something else*, since something[1] can refer to anything and obtain a fairly high predictive score compared to the other emotions. In CLIP's training, *something* must have referred to the presence of a recognisable entity. Whereas in ArtEmis, it is precisely when nothing is recognisable that one would be unsure on which emotion to feel and pick the *something else* option.

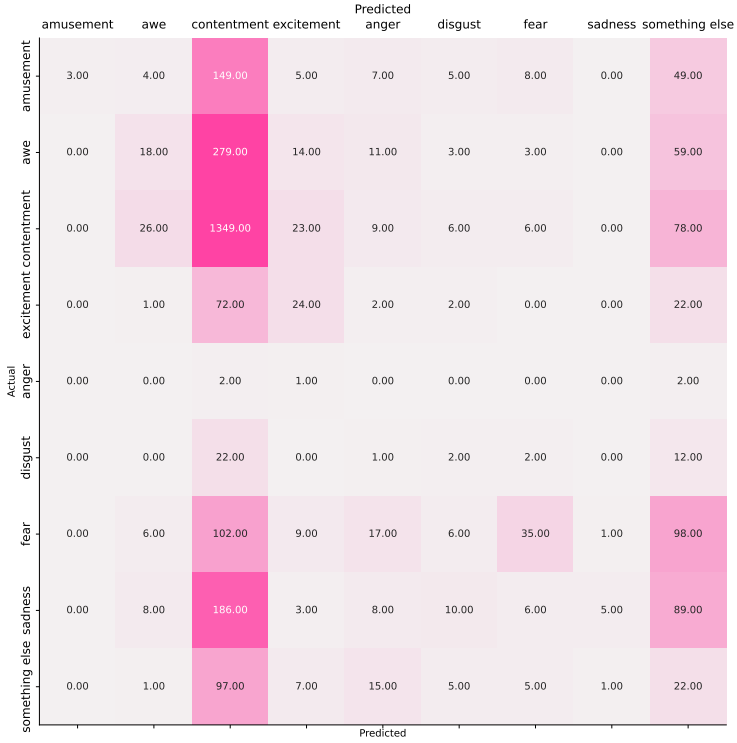


Figure 4.1: **Confusion matrix - Zero shot CLIP**.

accuracy	0.482
recall	0.166
f1_score	0.158

Table 4.1: **CLIPs zero shot performance on ArtEmis test set**. Overall accuracy, recall and f1 score both computed using macro averaging.

[1] As seen in chapter 2 CLIP text encoder struggles with the association of words. So *something else* is understood as something rather than not something.

4.2 Mapping CLIP's latent to the emotion space.

On the hypothesis that CLIP's latent space entails useful information for emotion prediction. We propose to use a single fully connected layer to translate the latent into nine emotion categories. The layer is trained and tested using the ArtEmis dataset.

4.2.1 Training

The fully connected layer is trained using Binary cross entropy (BCE) with logits loss or Mean Square Error (MSE) (both converge to equivalent weights), and Adam[12] as optimiser. We reuse the train-test-eval split of ArtEmis. All the architectures require preprocessing the input images (normalisation and center crop). We recommend to store the output of this preliminary step to save some computation time. The different architectures use the same preprocess step except for the RN50X16 which takes in higher quality images (the center crop yields 336 by 336 pixels images instead of 224 by 224). We use a batch sizes of 20 for the training and 40 for the other sets. The training takes 10 to 20 seconds per epoch on a GPU Tesla V100 on the train set of 70K images. After each epoch we decrease the learning rate when no validation loss improvement is observed, going from 10^{-2} to 10^{-5}. The models converge after a few epochs (5 to 10 depending on the feature size). We trained 8 fully connected layers, one for each of the 6 CLIP models (2 transformers : ViT-32 and ViT-16, 4 ResNets : ResNet50/101/50x4/50x16)[2], and as baseline a ResNet50 and Alexnet both pretrained on Imagenet.

4.2.2 Results

The same method used in Artemis is applied to evaluate the performance of each predictor. The method consists of taking a subset of the test set where a clear majority of annotator picked the same emotion. Using this method we note that most important factor for the classification is the training data, indeed a ResNet50 trained on ImageNet, ImageNet + ArtEmis, CLIP's dataset achieve an accuracy of respectively 56.9%, 59.7%, and 65.6%. By taking deeper models like RN50x16 the accuracy reaches 70% (Table 4.2 and Figure 4.2). The size of the features has little impact on the performance, suggesting that deeper model encode more relevant information in smaller feature space. The ResNnet50 and ResNet101 have the same performances yet the feature size of the ResNet101 is a quarter that of the ResNet50.

[2] Following EfficientNet[23] scaling

Training	Accuracy	Recall	f1_score	Feature size
Imagenet				
AlexNet	0.531	0.264	0.277	4096
RN50	0.569	0.270	0.289	2048
+ ArtEmis				
RN50	0.597	0.324	0.340	2048
CLIP-based				
RN50	0.656	0.385	0.409	2048
RN101	0.654	0.389	0.414	512
RN50x4	0.679	0.417	0.442	640
RN50x16	0.700	0.450	0.476	768
ViT-B32	0.672	0.409	0.434	512
ViT-B16	0.682	0.417	0.442	512

Table 4.2: **Performance comparison on the Artemis test set**. +ArtEmis indicates that the ResNet50 has been trained on imagenet and fine tuned on ArtEmis. The feature size is the size of the output feature vector of each model.

Effect of Annotator agreement

The previous evaluation is based on the strict 0.5 agreement threshold. As seen in section 3.4, we estimate that out of randomness around 25% of the paintings labeled as expressing a dominant emotion should not be. By changing the threshold we rise the confidence in the labeling at the cost of loosing some test samples. Changing the minimum agreement at 60% (which usually correspond to having 4 at of 5 annotators picking the same emotion) pushes the accuracy of all each model above 80%, the best model achieves 92% (Figure 4.3). As a comparison the ArtEmis ResNet50 achieves 75%. The rise in performance is either explained by the better quality of the annotation, or by sampling "easier examples". Above 50, 60 and 80% we keep only respectively 38, 15 and 3.4% of the test set. At more than 80% it is mostly the paintings expressing *contentment*, *fear*, and *sadness*.

Top 2 accuracy

Changing the agreement between annotator allows to raise the confidence in the dominant emotion but the set of images with high agreement is too small to be truly meaningful. A way to circumvent the limits of small sample size is to compare the top 2 emotions predicted by the model with the top 2 of the labels. Using this approach the best model finds at least one emotion for 94.1% of the test set against 90.6% for the ArtEmis model (see Figure 4.4). The two dominant emotions are matched only for 47.50% of the test set using the best model (42.81% for ArtEmis ResNet). We propose to use the top 2 emotion matching as benchmarking measure, as it tests

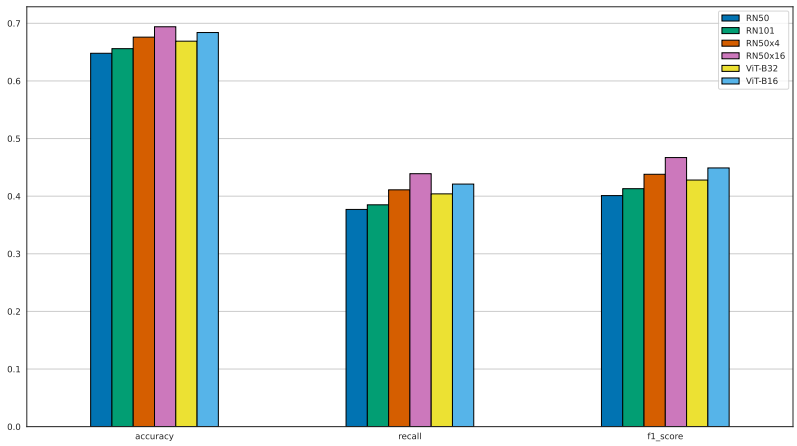


Figure 4.2: **CLIP architectures comparison**. The CNN based model performance grow according to their size as well as the transformers.

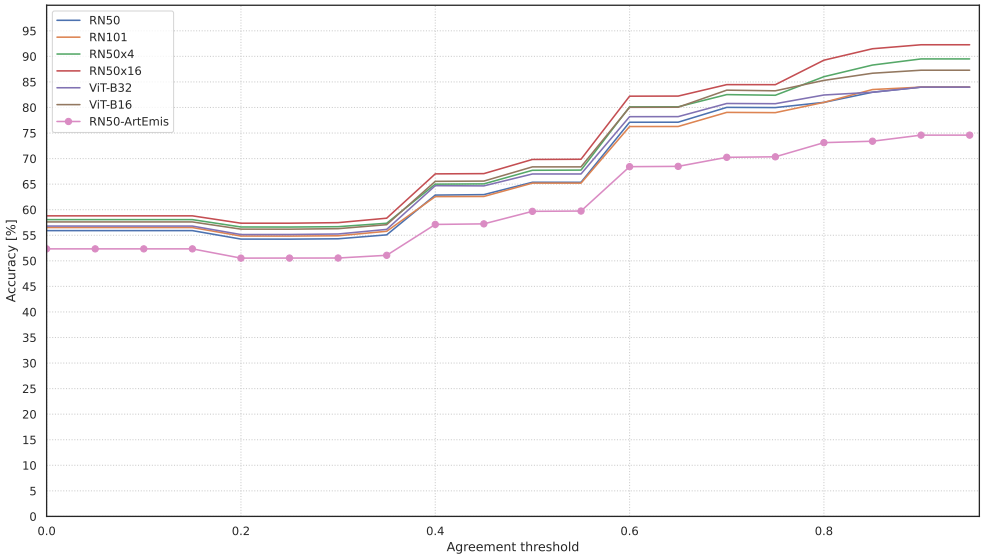


Figure 4.3: **Classification accuracy for paintings with agreement above a threshold** for each clip model. 96% of the images are labeled by 5 or 6 annotators, inducing noticeable step of 0.2 (samples with 5 annotators) and 0.17 (samples with 6 annotators). When we threshold the agreement of the annotators at strictly more than 50% we are in reality taking the threshold at the next step of agreement which corresponds to 60%.

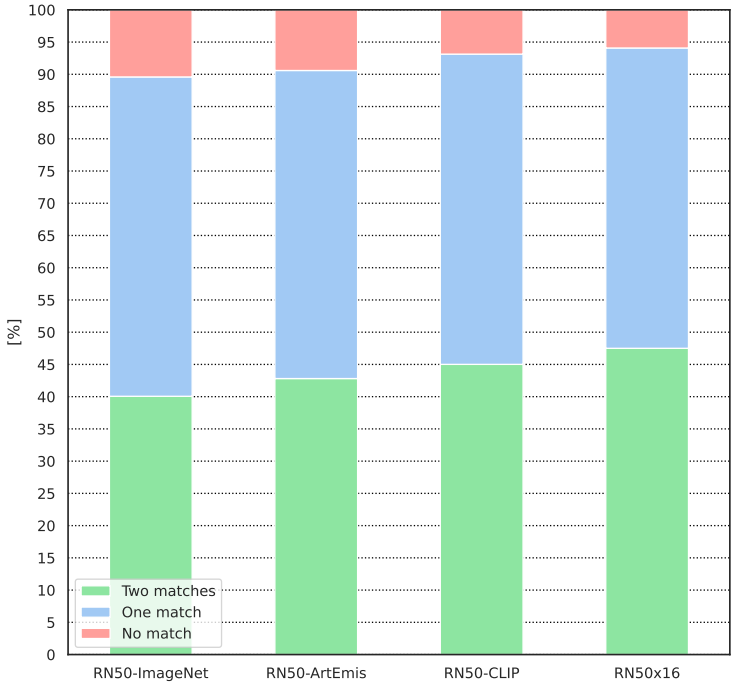


Figure 4.4: **Top2 results** for the three different ResNet50 as well as CLIP's RN50x16. From top to bottom : No match (red bar), one match (blue), and two matches (green). The one match happens if any of the top two prediction matches any of the top two labels.

the performance on the whole dataset and not just a subset like when using a threshold.

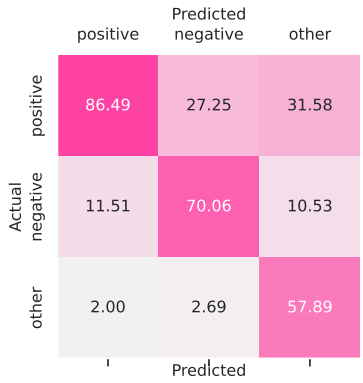


Figure 4.5: **Binary classifier confusion matrix.**

4.3 Maximising each emotion

This experiment tests the hypothesis that the model could gain in accuracy by maximising the emotions per paintings. Under the assumption that by training on a single label, the model performs a task similar to what is asked in test settings. The distributions are transformed in one hot encoded vectors based on the labeled dominant emotion. This data transformation turns out to be detrimental to the performance, on every architectures the accuracy drops below 50%. Even if the training mimics the test settings we are loosing a lot of information by removing the other emotions present in the dataset. Multi labels are found to be very beneficial for emotion prediction.

4.4 Binary classifier

Taking inspiration from text sentiment analysis we create a binary classifier to classify the paintings as positive, negative, or something else. Taking *amusement, awe, contentment*, and *excitement* as positive and *anger, disgust, fear*, and *sadness* as negative and *something else* as something else. The model is able to classify correctly 84% of the data and struggles with the something else category (Figure 4.5 and Table 4.3).

	positive	negative	other
precision	0.86	0.70	0.58
recall	0.94	0.54	0.07
f1_score	0.90	0.61	0.13
support	5192	1409	153

Table 4.3: **Binary classifier metrics.**

4.5 Making a query tool based on emotions.

In this section we showcase the application of our model by creating a query engine based on a prompt and an emotion. We first search for the images corresponding to the prompt, take the top 20 and then output the image maximising the desired emotion.

4.6 Text prediction

The strength of CLIP is to have a shared latent between the text encoder and image encoder. It is possible that the knowledge we obtained using images can be used on text as well. To test this we used the description provided by the annotators of ArtEmis, encoded them using CLIP's text encoder and translated the result using the layer trained on ArtEmis images. At 45% precision the results are promising, as noted by ArtEmis, their human baseline is at 61%, but their model reaches 64.8%.

4.7 Performances on other datasets.

4.7.1 The DIRTI dataset

The DIRTI[9] dataset contains images labeled as disgusting[3]. The disgust category appears to pose a challenge to the models, the models trained on imagenet have a net zero prediction on the *disgusting* images on the ArtEmis test set, even the fined tuned ResNet got 0. Making the DIRTI dataset a good evaluation tool. The testing is straight forward we extract the embeddings with CLIP and translate them to the emotion space with the layer trained on ArtEmis. The model conservative in assessing disgusting pictures, making no wrong *disgust* prediction (4.4). 26% of the disgusting images are classified in an other emotion category. The disgusting image that fooled it the most was of a mug of hot chocolate, and what at first glance appears to be milk bubbles is in fact mold. The model predicted this image to be mostly *contentment* and 10% disgusting. Interestingly the neutral image that received a stronger disgust reaction than the moldy hot chocolate was a picture of a pile of cucumbers (Figure 4.6). Have we created the first model with food taste ? We need a deeper analysis.

[3]They are indeed disgusting. For the readers sake we will restrain from showing any.

	other	disgust
other	60	0
disgust	63	177

Table 4.4: **Confusion matrix on the DIRTI dataset**

Figure 4.6: **A neutral image** predicted as 40% *contentment* and surprisingly 25% *disgust*.

Chapter 5

Deeper analysis

Understanding what truly sparks an emotion in a painting is a complex question. It can be correlated to the color, style or the subject of a painting. This chapter aims at understanding which one is predominant for the model's decision and if it made some spurious correlations. On the hypothesis that different emotion require different level of abstraction, we compare their predictability with different architecture and open up a ResNet50 to test each layer.

5.1 Correlation between emotions

The fully connected layer is represented by a matrix M of size (f_l, 9), and a bias vector of size 9 that we omit in this analysis, with f_l being the feature length of the output vector. M represents the projection from the feature space of a model to the 9 emotion space. Using the Cosine similarity (Equation 5.1), the correlation between the emotion projection learned by the model can be identified.

$$Cosine(x, y) = \frac{x \cdot y}{|x||y|} \tag{5.1}$$

Where x and y are two vectors of the same size. Figure 5.1 displays the correlations between each emotion learned by the model compared to the ones found in the accurate subset (section 3.2). *Anger* deviates the most from the annotators, probably due to the lack of good examples as exposed in section 3.1. It seems that *something else* is linked to *anger*. One should rethink the sentence "fear of the unknown" to "Anger of the unknown". *Amusement* is mostly correlated with *disgust* while being the opposite of awe. The features are well disentangled with the maximum correlation being linked to anger and disgust with 0.27. It would be interesting to find the direction corresponding to the continuous emotion representation (valence, arousal) and compute their correlation to the 9 discrete emotions.

5.2 About maximisation and failure

5.2.1 Maximisation

The paintings maximising each emotion (Figure 5.2) reveal some first insight on what the model learned. Here are our observation :

1. **Amusement**: Cartoons, people having fun, and strangely round characters.
2. **Awe**: Majestic scenery, religious icons and buildings.
3. **Contentment**: Gardens, calm and green forests, little rivers and ponds.
4. **Excitement**: Vivid colors, movement and action.
5. **Anger** the color red, some form of torture or injustice, and monochromes.
6. **Disgust** Rotten food, dead animals and anything sexual.
7. **Sadness** Sadness, tears, toil.
8. **Fear** hellish colors and figures.
9. **Something else**, minimalist paintings with dominance of white.

5.3 Using text

5.3.1 Artemis annotations

Using the short annotated sentences of ArtEmis we get an insight of the words maximising each emotion to confirm and supplement the observation of subsection 5.2.1. We sample 900 sentences, 100 for each label. And use CLIP's text encoder to map the sentences to the same space as the image and use the translator to infer the emotions of the sentences (same method as in section 4.6). With the predictions two things can be achieved, the first one is to take the sentences maximising the activation of a neuron (Table 5.2), or take the maximum relative emotion (Table 5.1). The first option maximises the absolute reaction. While the second option maximises the emotion relative to others, so that a sentence only expresses anger for example. We observe that prediction are much more concentrated for text inputs. With images, the usual output is very sparse and even the images maximising the emotion do not reach 100%. By contrast on text all the maximised prompts have at least 92% confidence (Table 5.1). The sparsity might come from the type of images, indeed on the DIRTI dataset, 75% of the *disgust* scores are above 30% against 4 art pieces in the ArtEmis test set.

Predicted	Description	Confidence	Actual
amusement	I am amused by this fellow's shappy clothes and his facial expression. Is he drunk?	100	amusement
	It's fun to see all of the people looking like their showing off their interesting eyewear.	100	amusement
	This woman clearly doesn't take much from anyone, and it is kind of amusing.	99	amusement
awe	This is awe-inspiring - the subject reaching for the sky, the hopeful colors of green, blue and yellow, the strength of the main subject's body	98	awe
	The colors look biblical and the picture reminds me that anybody can be inspired.	96	awe
	religious awe and woe combine here	95	sadness
contentment	I love the serenity of the painting, makes me feel peaceful and happy	100	contentment
	This looks like such a relaxing day out; these ladies look so calm and content sitting outside.	100	contentment
	The muted pastel colors reminds you of taking a walk on a hot humid day.	100	contentment
excitement	The bright colors against the black background reminds one of an abstract fireworks display.	100	excitement
	I love how the colors seem to be exploding out of this painting. It literally made my face smile.	96	excitement
	The commotion and different colors looks like a total mess.	95	anger
anger	It makes me angry that this man is pointing a gun at something with an angry look on his face.	100	anger
	The guy with the angry look on his face is about to do something bad to the man on the ground.	100	anger
	This is just red, a color that tends to invoke anger. It's also the most simplistic piece I've seen in this exercise.	100	anger
disgust	The man looks like his left eye is pulled out on his socket, which is disgusting.	99	disgust
	looks like the balls from a ball pit from chuck e cheese	94	amusement
	The abstract art is interesting and leaves me questioning.	92	something else
fear	it is scary and atmospheric	100	fear
	The darkness and her expression give an impression that she is a ghost.	100	fear
	The faceless figures in the painting make me feel very scare	100	fear
sadness	The woman looks rather sad with her forlorn expression and puffy lips.	100	sadness
	having the lady all dressed in black makes her look very depressed/sad	100	sadness
	He looks so sad, and I wonder what's going through his mind.	100	sadness
something else	The attention to the lines on his scalp encourage the viewer to think of this man has having an active and complex mind.	98	contentment
	the light gray area can be interpreted as a pond or sand	97	excitement
	This image is useless and lacks any sort of artistic design qualities or symbolic details	94	disgust

Table 5.1: **Sentences maximising confidence in an emotion.**

Predicted	Description	Confidence	Actual
amusement	The smiling lady and the dancing lady look like they're full of joy and having a great time. Vivid colors are very energizing.	64	amusement
	It's fun to see all of the people looking like their showing off their interesting eyewear.	100	amusement
	I am amused by this fellow's shappy clothes and his facial expression. Is he drunk?	100	amusement
awe	A family of higher status in the 1700 hundreds. Very well done and skillfully crafted. I never get tired of classical fine art.	93	awe
	This is awe-inspiring - the subject reaching for the sky, the hopeful colors of green, blue and yellow, the strength of the main subject's body	98	awe
	Heavenly visitors while young ladies contemplate a reading or being lost in thought. A very tranquil and serene scene.	8	awe
contentment	The nurturing mother and child image set in a landscape make this painting soothing	98	contentment
	The pastel blues from the water and greens from the tree are calming.	100	something else
	Heavenly visitors while young ladies contemplate a reading or being lost in thought. A very tranquil and serene scene.	92	awe
excitement	The bright colors against the black background reminds one of an abstract fireworks display.	100	excitement
	The smiling lady and the dancing lady look like they're full of joy and having a great time. Vivid colors are very energizing.	35	amusement
	The colors feel warm, bright and energetic while the pattern feels busy and alive	93	excitement
anger	It makes me angry that this man is pointing a gun at something with an angry look on his face.	100	anger
	This man looks sad and angry. I think is beard should be whiter to stand out more	12	anger
	The guy with the angry look on his face is about to do something bad to the man on the ground.	100	anger
disgust	These animals look scary and confusing based on the shape of their faces.	12	fear
	Odd but interesting painting of a nude woman laying down with a head dress. It's all done in muddy colors and seems old.	36	something else
	The man looks like his left eye is pulled out on his socket, which is disgusting.	99	disgust
fear	The sky is gloomy and the trees are giving off a spooky vibe, like something is hiding behind them.	100	fear
	The darkness and her expression give an impression that she is a ghost.	100	fear
	The forest appears dark, with a path leading in but not out. It is a bit eerie.	100	fear
sadness	The woman looks rather sad with her forlorn expression and puffy lips.	100	sadness
	This painting, with it's gray palate and lack of main subject, radiates loneliness.	100	sadness
	A solemn looking but pretty young African woman is the subject of the portrait. Her eyes are downcast and she looks like she is lost in thought.	100	something else
something else	Cows and horses graze peacefully in the sunlight	81	contentment
	it looks like a very sad boat, and if a boat is sad then I feel sad too	4	sadness
	The detail of this mans face, his features, even his skin is amazing. He looks tired and thoughtful.	1	awe

Table 5.2: **Sentences maximising emotion activation**

5.3.2 Predictions

The sentences maximising an emotion often contains the name of the emotion. Table 5.3 shows that *Fear, amusement* and *sadness* are empathy based, whereas *Anger* triggers*fear*. We test the different words of the list from item 9 and can confirm that the model associates "round" with amusement (Figure 5.3). The *disgust* emotion is strongly activated by food, too strongly even as, "steak" gets 58% *disgust* and "rotten food" 37%, right above "aubergine" at 35%. Figure 5.6 confirms that the model has learned to despise cucumber and especially if they come in piles. *Anger* proves challenging to trigger, the word "red" in itself does not pass the threshold of 10%. But associated with other words for instance "red injustice" we get 44%.

Politics

In this paragraph we attempt to unveil the political opinion of clip by giving some politician names. It is more an attempt of showcasing the possibilities of learned representation transfer on concept the model has not seen in the ArtEmis dataset. Interestingly CLIP seems to have preferences when it comes to politicians. With "Donald Trump" and "Richard Nixon" triggering respectively 74% and 76% *disgust* (even more than the pile of cucumbers!) and "Barack Obama" and "Elizabeth Warren" trigger 51% and 62% *contentment* (Figure 5.4). Future research has to be perform to understand on what bases the model is making those prediction. If it is from biases learned with CLIP or from some spurious emotional correlation inherited from ArtEmis. We hypothesis that those preference results from the web based crawling. If the hypothesis is true then some affective pattern towards words can be identify in populations using this tool.

Art

In this paragraph we go the full circle and ask the text predictor to predict the emotions of the art style name it has trained on, we don't expect CLIP to have the necessary cultural background to predict underlying emotions of art movements. Nevertheless Figure 1 shows surprising results. Indeed the "Naive Art Primitivism" is predicted 95% *amusement*, just as much as "Cartoon". When looking at the paintings maximising amusement, 4 out 9 originate from this movement. The paintings have predominant round characters[1]. This suggests that CLIP has possibly seen and learned about this movement. On an other note "Action painting" triggers excitement based on the word "Action" , as taken separately it triggers maximum excitement.

[1] https://www.wikiart.org/en/fernando-botero

Table 5.3: **Emotional reaction towards the emotion name as token**. This can be understood as how the model learned to react when recognising someone with such emotion. It learned that, the reaction towards anger is fear rather than anger. And *Amusement*, *sadness*, and *fear* perfectly correlate to their word.

5.3.3 failure

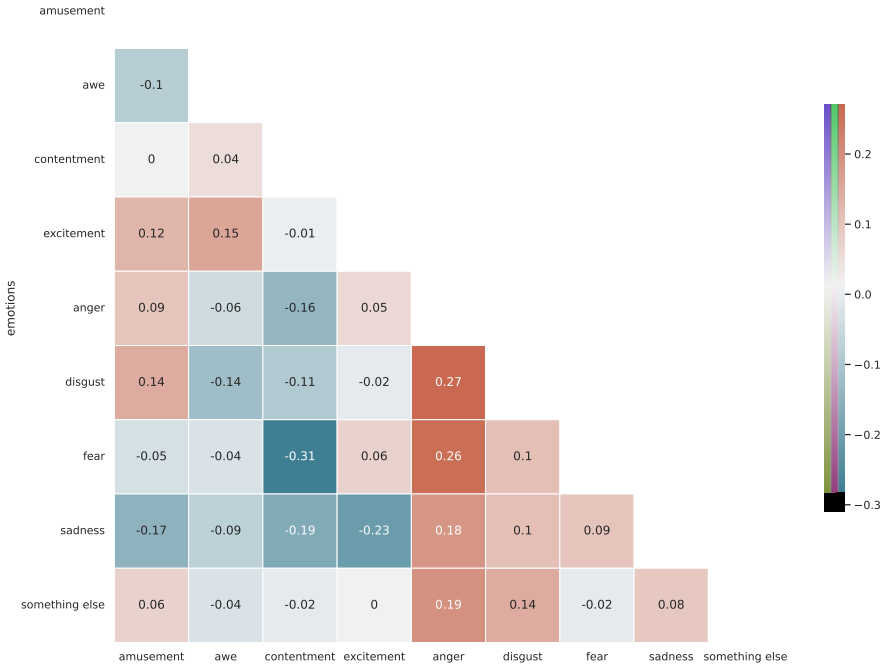
To identify the limits of our method we fetched the examples with the biggest difference with the labeling. The ones that strongly that had the biggest difference with their labeled dominant emotion. This allows us to identify a first limitation of the model which is the cropping in the preprocessing of the image. Indeed Figure 5.7 shows on the left the original image fed to the network and on the right the preprocessed image. The sad little boy on the bottom right corner is cropped out leading to misclassification. To mitigate this problem a better cropping method has to be considered. This problem is not specific to affect detection problems but affects all computer vision tasks. Future research should take care of this problem to optimise results.

5.4 Layers

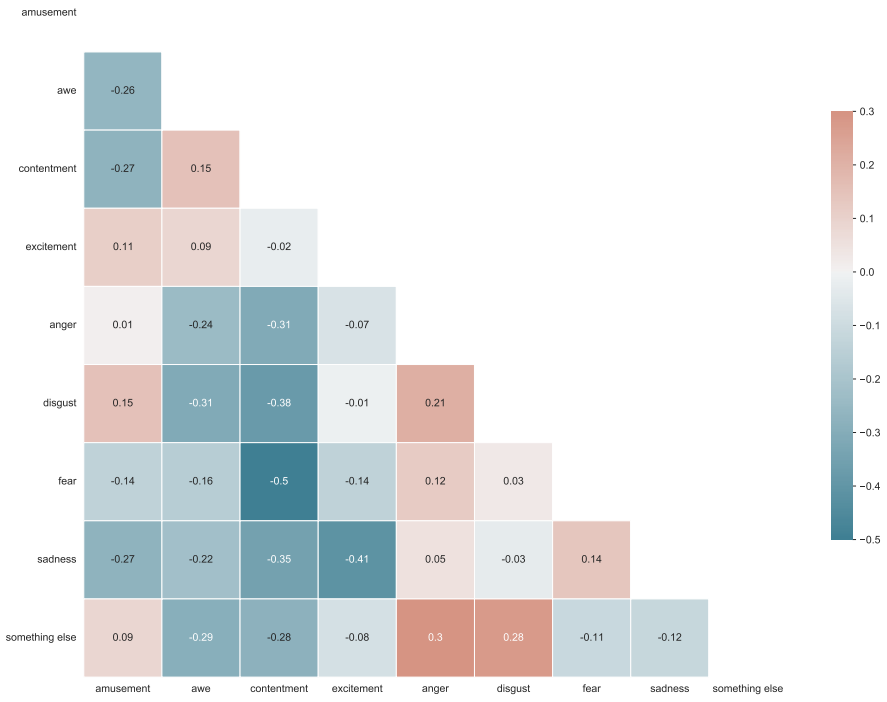
Methods[20] for affect prediction have used low level predictors like hue, saturation, colors, fourier transform, wavelets etc... We hypothesis that the low level features are understood in early layers of deep networks. If some emotions can be described with them, they should be similarly predicted regardless of the layer. Whereas emotion requiring complex cognition would be better predicted with features extracted closer to the output of a network.

We intercept the signal at each layer and train a layer on the extracted features using the same method seen in section 4.2. We extract the features of the 17 bottlenecks of CLIP's ResNet50. For each main layer the shape of their features are 256, 512, 1024, and 2048. We take the mean of each convolution image resulting in a significant loss but allow for fast training time.

Figure 5.8 shows the trend for each emotion at each stage of the network. The biggest improvement are observed for *sadness* and *disgust*, while being the smallest for *something else*. *Excitement* score rises in the second layer significantly confirming the observations that it is triggered by low level features like the color vibrance and diversity. In our observation of subsection 5.2.1 sadness seemed to be highly linked to empathy. The ability to recognize sad faces in latter stages, and indeed at each layer *sadness* is significantly improved compared to the other emotions.



(a) Model



(b) Annotators

Figure 5.1: **Emotion correlation**. On (a) the correlation between the emotions found by the fully connected layer (RN50x16 clips features), and (b) the correlations for human annotators on paintings with more than 40 annotators (1% of the dataset).

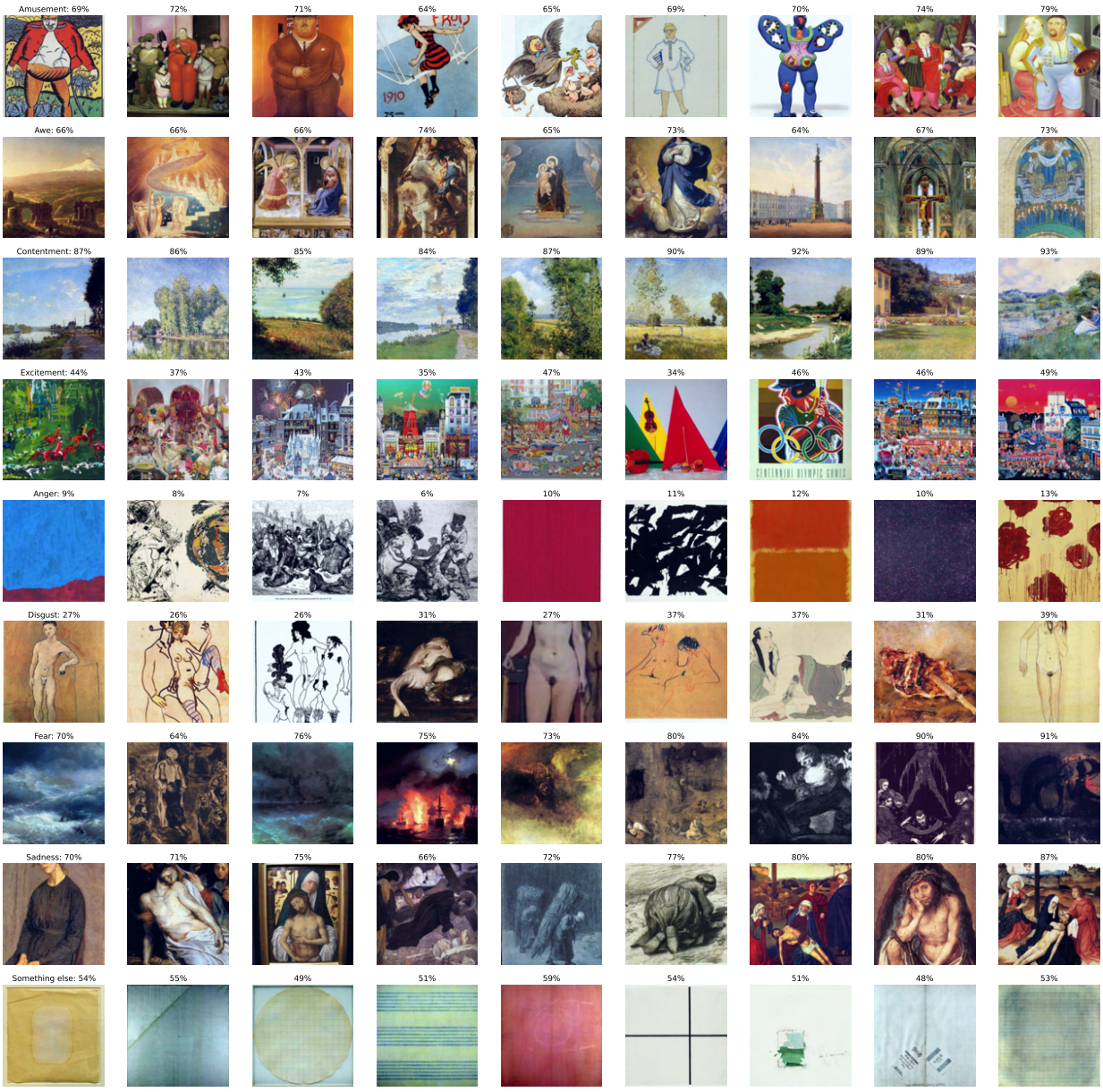


Figure 5.2: **Paintings maximising each emotion**, the paintings are taken from the ArtEmis test set.

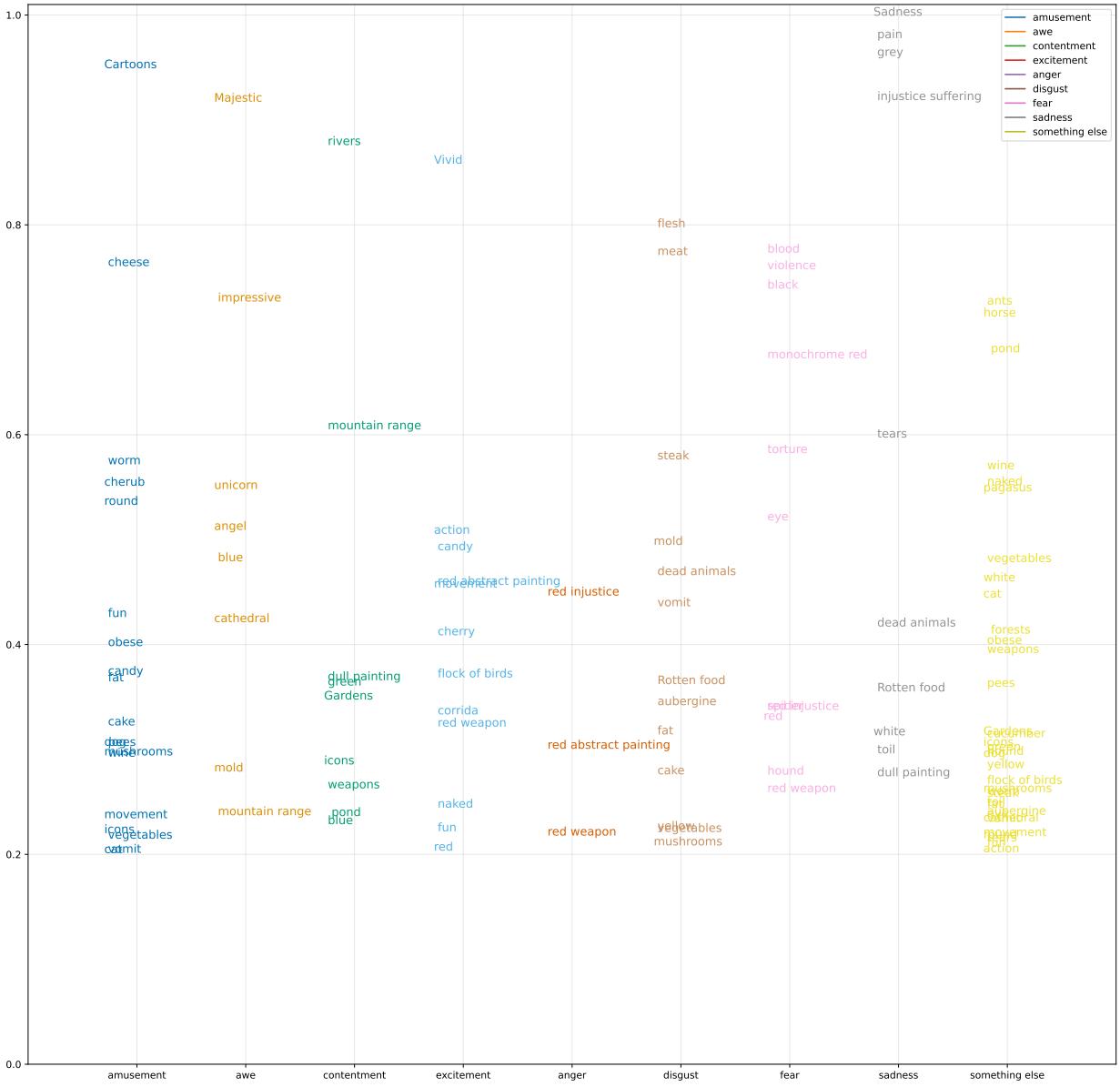


Figure 5.3: **The predicted main emotions** for different words. Only the top three reactions are displayed if they are above a 0.2 threshold.

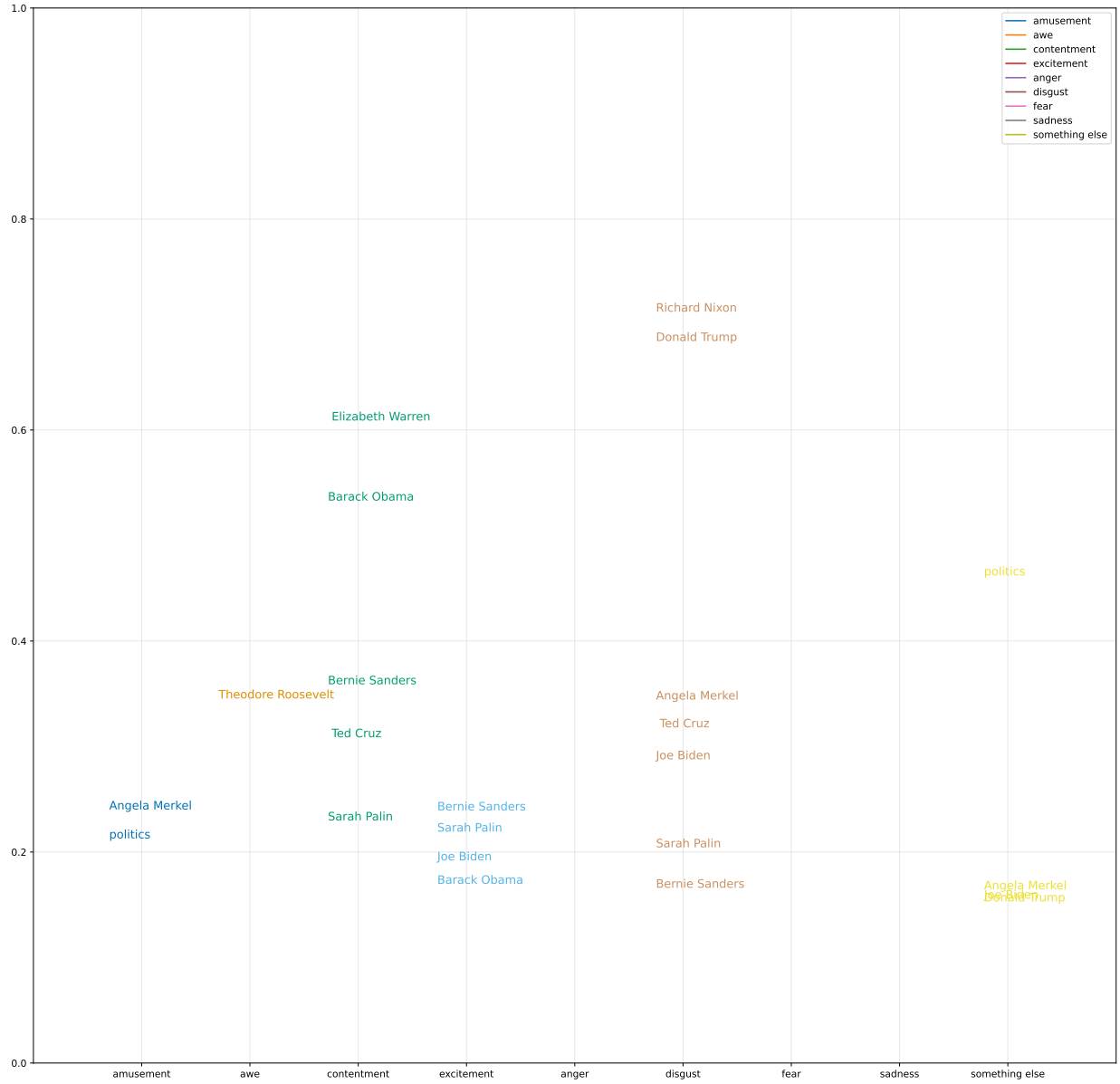


Figure 5.4: **Politicians main emotions predictions**. Only the top three reactions are displayed if they are above the 0.15 threshold. The graph displays the potential use the text based emotion prediction to have an initial idea about what CLIP, and by extension the web feels about different persona.

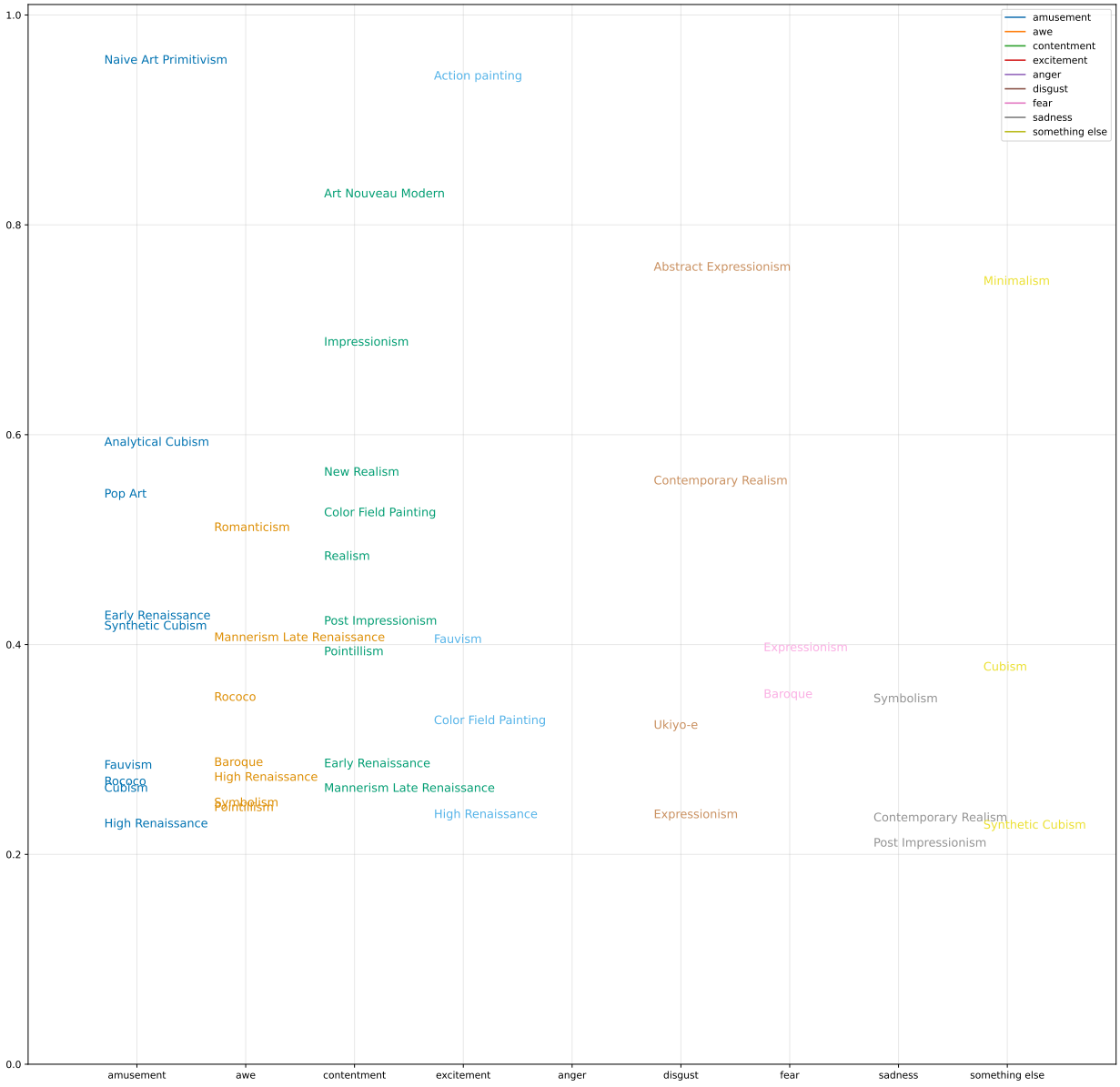


Figure 5.5: **Art movement three main emotions predictions**. Only the reactions above 0.2 are kept for visibility.

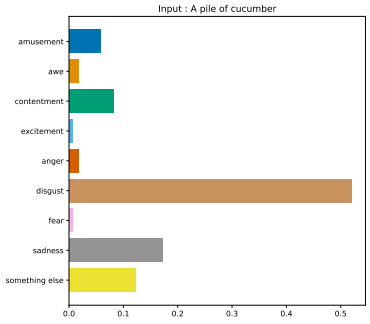


Figure 5.6: **Emotional effect of "A pile of cucumbers".** The model reacts with disgust and fear.

(a) Original

(b) Preprocessed

Figure 5.7: **Faulty crop** of the small boy in the bottom right corner, explaining why the model classified this image as contempt instead of sad.

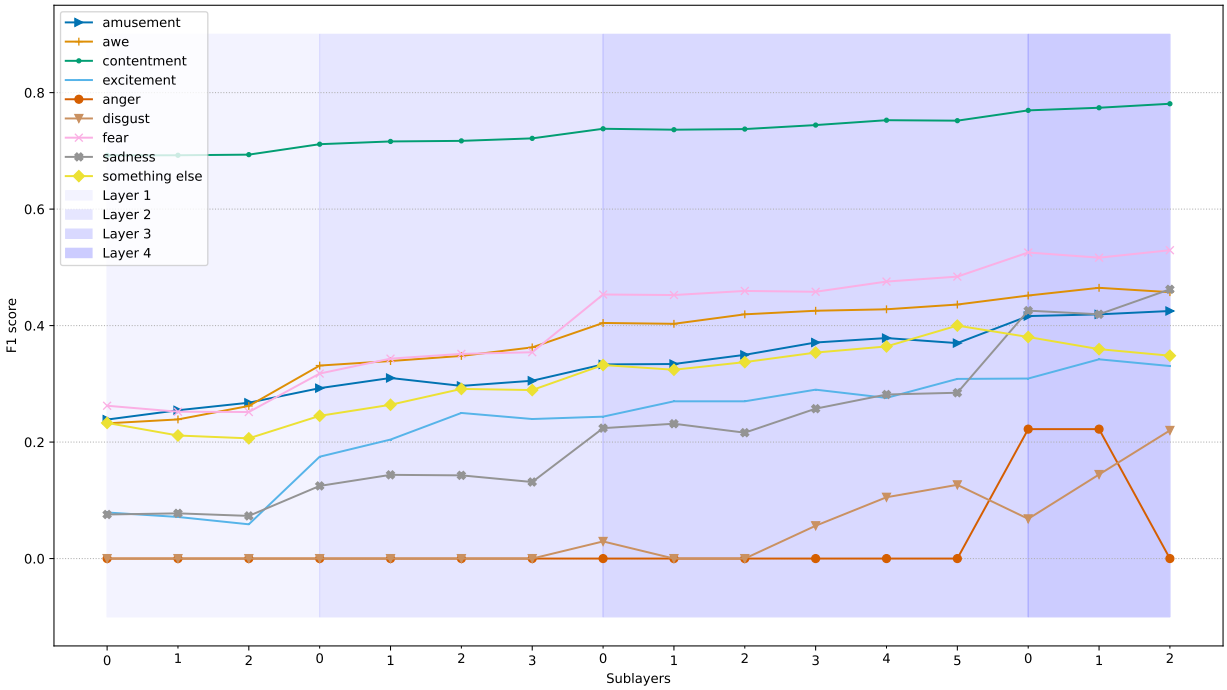


Figure 5.8: **The f1 score at each layer**

Chapter 6

Discussion and future work

This work has taken inspiration from art, psychology, sociology and machine learning fields combined. Numerous papers are using clip as backbone for many downstream task and we are one of them. Whether it be with no example or with eighty thousand CLIP is highly versatile. Nevertheless CLIP text encoder could be improved to allow for better correspondence between words. Especially negation and adjectives. A pink shirt should be understood as a pink shirt, and not as two separate entities.

Some improvement in the ArtEmis dataset could be made, which is to our knowledge the only large affective annotated dataset. A subset of a 2 to 3 thousand paintings annotated in the same fashion by more than 40 annotators would reduce dominant emotion error from 30% to less than 15% and provide better testing grounds. An other possibility would be to annotate the dataset or a subset of it in terms of valence and arousal, and create a mapping between the two. Expanding the dataset by including recent art movement or art movement entailing more negative emotions would provide better understanding and predictability for *anger*.

We have shown that it is straight forward to transfer knowledge learned on image on text using CLIP. The logical step would be to train on text and see its transfer capabilities on images. And further combine the two to achieve the best performance.

The political bias is certainly one of many, further investigation and intensive test should be made as CLIP and similar models are becoming the norm. The investigation would serve two purposes, a machine learning one to reveal hidden flaws of models to improve them. And a sociological one, to reveal the biases of the training set and infer conclusion on populations. Internet data brings its share of limitations and is is important to reveal them to avoid any mistakes. The same can be said with training on annotated art. The historical and cultural context of the annotators is ingrained in the predictor. ArtEmis holds many interesting data analysis related questions: The perception of art through time, the discrepancies of reactions toward art pieces. All these questions spark interest in sociology and the psychology of art, the emotional predictor could be a tool in large scale art analysis.

We tested the emotion model on 1000 real images, and it should be further tested by scaling up. Classifying a set of images and asking annotator if they feel the same as the model could be an effective way to test the model and further identify its limits. We have identified biases like overly classifying round objects and people as *amusement*. And other oddity certainly exist and should be found.

Chapter 7

Conclusion

Using CLIP as backbone we have created an emotion predictor and shown its capabilities to predict emotions in images by training a single fully connected layer. Since CLIP has a shared image-text feature space we were able to build an emotion translator on both by only training on images. Enabling Image emotion prediction (70% accuracy on the artemis test set, the ArtEmis trained model reached 60%). As well as text emotion prediction (40% accuracy on artemis test set against 60% for the human baseline). The text translation was used to assess which emotion distribution sparked from which words, and we surprisingly discovered that CLIP has a political opinion. The text prompt "Donald trump" outputted *disgust, anger*, and *amusement*, whereas for "Barack Obama" has 80% confidence in *contentment*. Furthermore by intercepting the flow of information at each bottleneck of a ResNet50 we showed that the *disgust* emotion requires more understanding than other emotion and suggests it be a good reference point to asses the performances emotion predicting models. By using CLIP and the ArtEmis dataset we have created a tool with various application such as affective search. Large scale affective evaluation and provided an extra tool to quickly understand the link between a word and his corresponding emotional perception by CLIP.

Bibliography

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. "Artemis: Affective language for visual art". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11569–11579.

[2] Ralph Adolphs. "The biology of fear". In: *Current biology* 23.2 (2013), R79–R93.

[3] Lisa Feldman Barrett. "Are emotions natural kinds?" In: *Perspectives on psychological science* 1.1 (2006), pp. 28–58.

[4] Pierre Bourdieu. *The rules of art: Genesis and structure of the literary field*. Stanford University Press, 1996.

[5] Paul Ekman. "An argument for basic emotions". In: *Cognition & emotion* 6.3-4 (1992), pp. 169–200.

[6] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. "CLIP2Video: Mastering Video-Text Retrieval via Image CLIP". In: *arXiv preprint arXiv:2106.11097* (2021).

[7] Kevin Frans, LB Soros, and Olaf Witkowski. "Clipdraw: Exploring text-to-drawing synthesis through language-image encoders". In: *arXiv preprint arXiv:2106.14843* (2021).

[8] Theodoros Galanos, Antonios Liapis, and Georgios N Yannakakis. "AffectGAN: Affect-Based Generative Art Driven by Semantics". In: *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE. 2021, pp. 01–07.

[9] Anke Haberkamp, Julia Anna Glombiewski, Filipp Schmidt, and Antonia Barke. "The DIsgust-RelaTed-Images (DIRTI) database: Validation of a novel standardized set of disgust pictures". In: *Behaviour research and therapy* 89 (2017), pp. 86–94.

[10] Henrik Hagtvedt, Vanessa M Patrick, and Reidar Hagtvedt. "The perception and evaluation of visual art". In: *Empirical studies of the arts* 26.2 (2008), pp. 197–218.

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision". In: *arXiv preprint arXiv:2102.05918* (2021).

[12] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[13] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. “Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology”. In: *Biological psychiatry* 44.12 (1998), pp. 1248–1263.

[14] Jana Machajdik and Allan Hanbury. “Affective image classification using features inspired by psychology and art theory”. In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 83–92.

[15] Matiur Rahman Minar and Jibon Naher. “Recent advances in deep learning: An overview”. In: *arXiv preprint arXiv:1807.08169* (2018).

[16] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. “MovieCuts: A New Dataset and Benchmark for Cut Type Recognition”. In: *arXiv preprint arXiv:2109.05569* (2021).

[17] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. “Styleclip: Text-driven manipulation of stylegan imagery”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2085–2094.

[18] JONATHAN POSNER, JAMES A. RUSSELL, and BRADLEY S. PETERSON. “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology”. In: *Development and Psychopathology* 17.3 (2005), pp. 715–734. DOI: 10.1017/S0954579405050340.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. “Learning transferable visual models from natural language supervision”. In: *arXiv preprint arXiv:2103.00020* (2021).

[20] Christoph Redies, Maria Grebenkina, Mahdi Mohseni, Ali Kaduhm, and Christian Dobel. “Global image properties predict ratings of affective pictures”. In: *Frontiers in psychology* 11 (2020), p. 953.

[21] Amy Smith and Simon Colton. “CLIP-Guided GAN Image Generation: An Artistic Exploration”. In: *Evo* 2021* (), p. 17.

[22] Kaikai Song, Ting Yao, Qiang Ling, and Tao Mei. “Boosting image sentiment analysis with visual attention”. In: *Neurocomputing* 312 (2018), pp. 218–228.

[23] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[24] Robert E Thayer. *The origin of everyday moods: Managing energy, tension, and stress*. Oxford University Press, USA, 1996.

[25] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. *Self-training with Noisy Student improves ImageNet classification*. 2020. arXiv: 1911.04252 [cs.LG].

[26] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. “Exploring principles-of-art features for image emotion recognition”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 47–56.

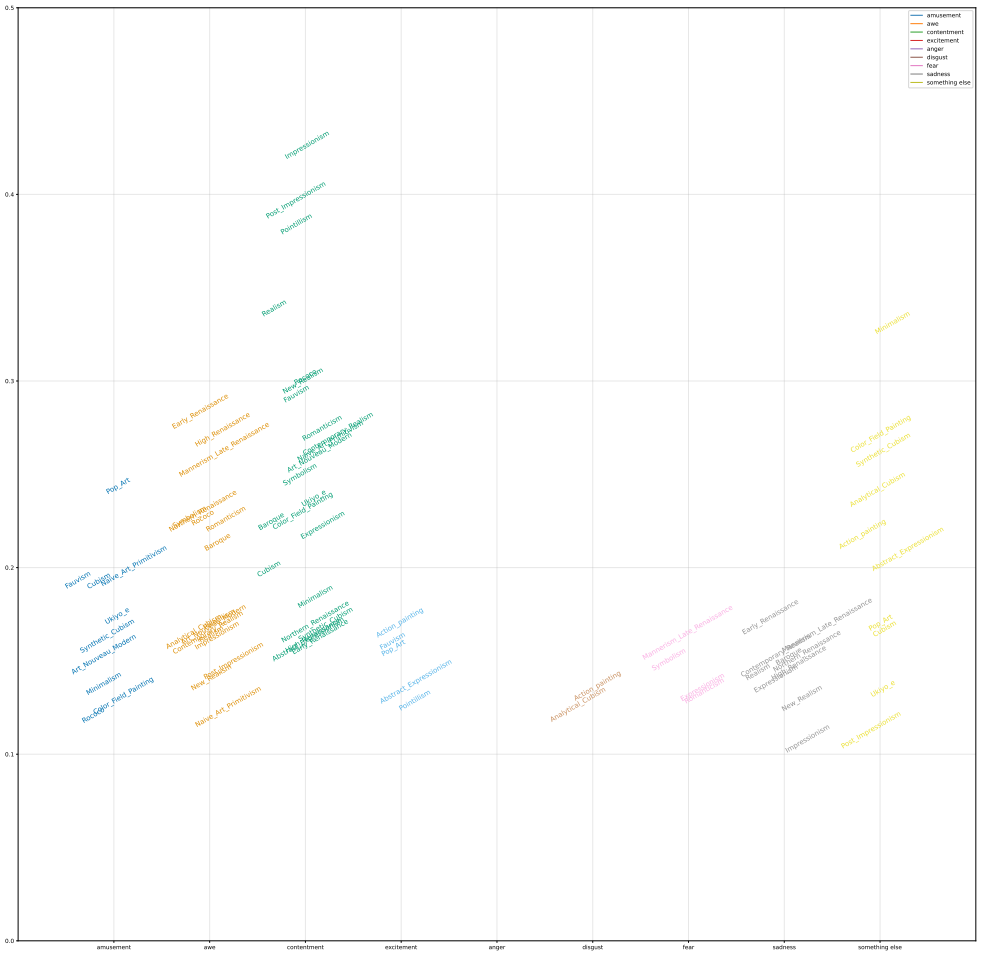


Figure 1: **The three dominant emotion for each style based on the annotation of Artemis. The y scale goes from 0 to 0.5.**

	amusement	awe	contentment	excitement	anger	disgust	fear	sadness	something else
amusement	1	1	47	0	0	0	0	2	9
awe	0	0	0	0	0	0	0	0	0
contentment	0	0	0	0	0	0	0	0	0
excitement	0	0	0	0	0	0	0	0	0
anger	0	0	0	0	0	0	0	0	0
disgust	6	6	14	1	0	177	9	8	19
fear	0	0	0	0	0	0	0	0	0
sadness	0	0	0	0	0	0	0	0	0
something else	0	0	0	0	0	0	0	0	0

Table 1: **DIRTI dataset confusion matrix**