Leveraging CLIP for Image Emotion Recognition

Alessandro Bondielli[1] and Lucia C. Passaro[2]

[1] Department of Information Engineering, University of Pisa, Pisa, Italy
alessandro.bondielli@ing.unipi.it
[2] Department of Computer Science, University of Pisa, Pisa, Italy
lucia.passaro@unipi.it

Abstract. Multi-modal neural models that are able to encode and process both visual and textual data are becoming more and more common in the last few years. Such models enable new ways to learn the interaction between vision and text, and thus can be successfully applied to tasks of varying complexity in the domain of image and text classification. However, such models are traditionally oriented to learn grounded properties of images and of the objects they depict and less suited to solve tasks involving subjective characteristics, such as the emotions they can convey in viewers. In this paper, we provide some insights in the performances of the recently released OpenAI CLIP model for an emotion classification task. We evaluate the model both under zero-shot settings and via fine tuning on an image-emotion dataset. We compare the performances of CLIP both in a zero-shot and fine-tuning setting on (i) a standard benchmark dataset for object recognition (ii) an image-emotion dataset. Moreover, we evaluate to which extent a CLIP model adapted to emotions is able to retain general knowledge and generalization capabilities.

Keywords: Affect · Emotion Classification · Computer Vision · Natural Language Processing · CLIP

1 Introduction

The ever-increasing production and spread of multi-modal content over the internet requires new analytical tools to deal with them. Although many issues related to the multi-modal analysis of text and images have already been addressed in the literature, it is still unclear whether and to what extent state-of-the-art multi-modal systems can be exploited to explore the affective characteristics of the visual contents.

Several multi-modal resources, systems and architectures have been proposed in the literature to approach a wide range of natively multi-modal tasks, such as Image Captioning [5, 12], Visual Question Answering [20, 21] and Image Generation [18]. However, traditional literature in the field of Computer Vision and

specifically of Image Classification typically focuses on the recognition of objects and concrete entities depicted in the images. In this context, several large-scale resources useful to train neural models have been released [8, 10, 13]. On these benchmarks, the literature is plentiful of systems that have been proven effective in solving tasks of various levels of complexity [2, 7, 19, 20, 16].

On the contrary, the field of Natural Language processing has addressed problems related to the affective properties of texts for many years. The literature is filled with approaches dealing with sentiment, opinion and affect. For example, several studies have been proposed to analyze the sentiment and the emotions expressed and evoked by texts from several perspectives [4, 6, 14, 15].

The sentiment encoded in images has attracted a lot of interest due to its various applications [3, 9], ranging from human-robot interaction to social media analysis, but the results are not on par neither with systems working only on text nor with computer vision systems focused on concrete aspects of visual contents. This may be due to the fact that images convey rich semantic properties and can induce, as textual inputs can and possibly even more, emotional reactions to users who are exposed to them. Thus, it is important to develop new benchmarks to assess the ability of systems to classify images from an affective point of view.

This aspect is also very relevant in the field of Industry 4.0. Companies are in fact expected to constantly communicate with their customers using new and effective forms of communication, such as the visual ones. On the one hand, it is important to study the emotional content conveyed by an image. On the other hand, especially for web marketing purposes, it is crucial to analyse the emotions "elicited" by images in viewers.

To the best of our knowledge, a fully multi-modal dataset that includes real-world image samples addressing this issue is still lacking. However, a large scale visual dataset labelled for the emotions evoked by images has been proposed in [23]. This dataset is suitable to challenge state-of-the-art multi-modal neural models in predicting subjective, abstract labels for a given image like emotions. Thus, the analysis of the performances on this dataset may be seen as an early attempt to exploit pre-trained multi-modal systems to bridge the gap between computer vision and affect.

To study aspects related to the emotions evoked by images, we decided to base our experiments on the recently released and well-known OpenAI CLIP model [17], a multi-modal Neural Network learned on text-image pairs. CLIP adopts an architecture that includes an image encoder and a text encoder. The peculiarity of CLIP resides in its *contrastive* training strategy. CLIP is trained on a dataset of 32,000 image-caption pairs. Its training objective is to predict, given an image, which of the captions was actually paired with it in the training dataset. The goal of this pre-training is to provide the network with a wide array of visual concepts found in images and enable it to learn how to identify proper associations between these visual content and their textual descriptions or presentations [17].

In this context, the motivation of our choice to adopt CLIP is twofold. On the one hand, the model has been trained to efficiently learn visual concepts

by exploiting natural language supervision. We can argue that it may directly encode latent emotive concepts. On the other hand, CLIP authors claim that it can be used to nearly arbitrary visual classification tasks [17] under zero-shot setting. Moreover, from an implementation perspective CLIP and CLIP-like models have a very interesting property that stems from their training approach: representations of images and texts (e.g. captions) can be easily compared in terms of cosine similarity between their vectors. For example, classification on a 10-class dataset can be faced with CLIP by simply encoding labels in the form of captions, and then by identifying the closest caption (i.e. label) in terms of cosine similarity for each image. Representations of images could be stored in memory and queried at inference time for their similarity with either another image or a piece of text, thus drastically reducing the computational cost at inference time.

We are conscious that analyzing affect elicited by images is a very challenging task because people with diverse social and cultural backgrounds may have different emotional reactions to the same image [22]. Moreover, we know that labelled datasets addressing this issue are scarce. However, for our preliminary studies we consider the Image-Emotion dataset [23] as suitable to draw the first insights to approach emotion classification of images.

In this work, we propose to exploit and analyze the performances of CLIP for the task of image emotion recognition. CLIP can be leveraged either as a pre-trained model for zero-shot classification, as intended by the authors [17], or by further fine-tuning it on specific downstream tasks. Our goal is to explore how CLIP models perform on highly subjective tasks out-of-the-box and how they can be adapted to them via fine-tuning. Moreover, as the task of image emotion recognition is rather challenging, we aim to compare it also with a more standard classification task on a computer vision benchmark, namely the popular CIFAR100 benchmark dataset [11].

The contributions of this paper are the following:

- We evaluate the zero-shot performances of CLIP on two different benchmark datasets, namely (i) a dataset for image emotion recognition and (ii) a dataset for a more standard image classification problem;
- We evaluate CLIP in a fine-tuning setting on two different tasks, namely (i) image emotion recognition and (ii) image classification, and compare the obtained results;
- We evaluate to which extent CLIP is able to retain general knowledge and generalization capabilities to other tasks after being fine-tuned.

The rest of this paper is organized as follows. Section 2 thoroughly describes the performed experiments. In Section 3 the results of the experiments are presented and discussed in order to shed some light into the capabilities of CLIP for image emotion recognition. Finally, Section 4 draws some conclusions and discusses future work.

2 Experiments

In order to provide some insights into the capabilities of CLIP for emotion recognition we perform several different experiments. The experiments are devised in order to fulfil two goals: first, we want to assess the performances of CLIP under zero shot settings; second, we want to evaluate the impact of fine-tuning in the performances of the model, both for the specific task and in its generalization capabilities. In addition to this, we try to address the differences between the performances of the CLIP model on more abstract and more concrete tasks across all of the performed experiments.

For the image emotion recognition task we employ the dataset described in [23], that we refer to as the Image-Emotion dataset. The dataset includes 23,308 images labelled with an emotion among Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sadness. The images are collected and weakly labelled by searching for the emotion keywords on Instagram and Flickr. The weak labelling is then verified with a crowdsourcing experiments. Concerning the more concrete image classification task, we employ the popular CIFAR100 dataset [11]. It includes 60,000 images labelled with one of 100 classes of objects such as for example Dolphin, Road, and Boy.

Our experiments are organized as follows. First we perform zero-shot classification on the two dataset using the pre-trained `ViT-B/32` CLIP model. Second, we fine-tune the CLIP model on the two datasets, and evaluate its performances in a cross-validation experiment. Third, we again perform zero-shot classification on each of the two datasets using the model fine-tuned on the other one. This means that the model fine-tuned on the Image-Emotion dataset is applied to CIFAR100 and vice versa. This aims to understand how fine-tuning affects zero-shot performances on other tasks.

All the experiments are performed by exploiting the CLIP python library[3] and the official pre-trained available models. In the following, we thoroughly describe the experiments and show the obtained results.

2.1 Zero-shot classification

In the first set of experiments, we simply employ a pre-trained CLIP model to classify images under zero shot settings. Following the original CLIP paper [17], we perform classification by means of cosine similarity between image representations and captions. Notably, since we have labels and not captions for both of the employed datasets, we first generate a caption for each label in the dataset. For CIFAR100, the employed caption is "a photo of a $< label >$", where $< label >$ is one of the 100 labels of the dataset. For the Image-Emotion dataset, the caption is "an image that evokes the emotion of $< emotion >$", where $< emotion >$ is one of the eight emotion labels. We use a different wording for the two datasets (i.e., image and photo) due to the fact that all the data in CIFAR100 consists of photos, while the Image-Emotion dataset includes also more abstract images.

[3] https://github.com/openai/CLIP

For both experiments, we encode all the images and all the captions with the CLIP model. Specifically, we use the `ViT-B/32` pre-trained model. Then, we compute cosine similarity between the representations of each image and each caption. To obtain the final label, we simply assign to each image the caption (label) with the highest cosine similarity to it.

2.2 Fine-Tuning CLIP

For the second set of experiments, our goal is to evaluate how much improvement could be obtained on downstream tasks by fine-tuning a base CLIP model. More specifically, we focus on the two downstream tasks of (i) image emotion classification on the Image-Emotion dataset and (ii) image classification on the CIFAR100 benchmark.

In order to obtain reliable and comparable estimations of the performances, we use 10-fold cross validation during training. However, since the two datasets are different in terms of size, number of classes, and distribution of classes, we also perform some hyperparameter tuning to obtain the best possible results on both datasets. For the sake of brevity we leave out the details of parameter tuning. However, in this regard it is very interesting to notice how the process of fine-tuning CLIP is extremely sensitive to different hyperparameters. For example, a slight change in learning rate or number of training epochs may lead to a decrease in performances of up to 0.20 in weighted and macro average F1-Score.

First, we experiment with the Image-Emotion dataset. We refer to the resulting model as `Emotion-CLIP`. As previously mentioned, we perform 10-fold cross validation on the whole dataset. Each fold is composed of 20,000 training examples and 3,500 test examples. The model is evaluated by predicting the most likely label for each image by means of cosine similarity with respect to the generated captions, as under zero-shot setting described in Section 2.1. As for the hyperparameters, we train each fold for 3 epochs with a batch size of 256. We use an Adam optimizer with a learning rate of 2e-5 and a 0.2 weight decay. Training each epoch took roughly 3 minutes on a Nvidia Titan RTX GPU. To obtain the final results for the classification, we average performances on each fold.

In order to further evaluate how the process of fine-tuning can be helpful also for zero-shot capable models, we propose to exploit a simpler and more grounded task of image classification on the CIFAR100 dataset. We refer to the trained model as `CIFAR100-CLIP`. As for the previous experiment, we perform 10-fold cross validation on the entire dataset (i.e. the concatenation of train and test set), with the predictions obtained by means of cosine similarity between images and captions. Each fold is composed by 54,000 training samples and 6,000 test samples. Note that the distribution of classes on the whole dataset is perfectly balanced (i.e. each label is associated with exactly 6,000 images). After tuning the parameters, we chose to train the model on each fold for 1 epoch with a batch size of 256. The same learning rate and optimizer used for `Emotion-CLIP` are employed also in this case.

2.3 Evaluation of fine-tuning on generalization capabilities of CLIP

While fine-tuning is a viable strategy for applying CLIP to downstream classification task, the original goal of CLIP is to take advantage of the interaction between natural language and images to perform image classification tasks without the need of direct optimization for the dataset at hand [17]. With the last set of experiments, our goal is twofold. On the one hand, we want to straightforward understand how and how much fine tuning on a benchmark task actually affect the zero-shot capabilities of CLIP. On the other hand, the experiments also serve to assess the extent to which a specific kind of benchmark data may affect zero-shot performances. In the original paper, authors clearly state that while zero-shot performances on simpler image classification tasks are very promising, the model encounters more difficulties when the task becomes more complex (e.g. counting specific objects in the image) or more abstract. In this context, we want to shed some light into how fine-tuning on a more challenging task such as emotion recognition would affect performances on simpler tasks, and vice versa.

In order to pursue this goal, we propose the following experiments. We first fine-tune the `Emotion-CLIP` model on the whole Image-Emotion dataset, and test it under zero-shot settings on the CIFAR100 dataset for image classification. Then, we do the opposite, i.e. we train `CIFAR100-CLIP` on the CIFAR100 dataset and test it for image emotion recognition on the Image-Emotion Dataset.

Both `Emotion-CLIP` and `CIFAR100-CLIP` are trained on their respective datasets with the same parameters employed for the cross-validation experiments described in Section 2.2. The only difference is that, in this case, the model is trained on the whole dataset. As for testing, the models are deployed in zero-shot setting and labels for both the CIFAR100 and Image-Emotion Dataset are obtained by means of cosine similarity between images and generated captions.

3 Results and Discussion

In this Section, we provide the results obtained for each of the performed experiments and discuss them to shed some light on the performances of CLIP with the different settings and datasets.

3.1 Zero-shot classification

First, we evaluate the performances of the CLIP model under zero-shot settings both for the Image-Emotion dataset and for the CIFAR100 benchmark. As described in Section 2.1, for both the experiments the original CLIP `ViT-B/32` pre-trained model was asked to compare the cosine similarity between the generated captions and the images. As for the Image-Emotion dataset, we used the following captions: "an image that evokes the emotion of $< emotion >$", where $< emotion >$ stands for one of the eight emotion classes in the dataset. As for the CIFAR100 benchmark, the captions were of the form "a photo of a $< label >$", where $< label >$ is one of the 100 labels in CIFAR100.

Table 1: Experiments under zero shot settings

		CIFAR100	Image-Emotion
Accuracy		0.62	0.49
Precision	Macro Avg.	0.69	0.46
	Weighted Avg.	0.69	0.52
Recall	Macro Avg.	0.62	0.44
	Weighted Avg.	0.62	0.49
F1-Score	Macro Avg.	0.61	0.42
	Weighted Avg.	0.61	0.48

Results for the two datasets are shown in Table 1. We report accuracy, weighted-average and macro-average precision, recall, and F1-score for each dataset. We can see that, as expected, despite the much higher number of classes in the CIFAR100 dataset, the CLIP model under zero-shot settings is better able to predict its labels with respect to the emotion elicited by the image in the Image-Emotion dataset. We can argue that this is due to the fact that the training data for CLIP is much more akin to the CIFAR100 classification task. However, it is interesting to notice how the baseline model is nevertheless fairly able to face also a more complex and more abstract task such as emotion recognition out-of-the-box.

For the sake of completeness, we also report on class-level performances for the Image-Emotion dataset in Table 2.

We notice that there is a high variance in performances among classes, that is however not directly related to the sample size on each class. In fact, it seems that some emotions such as Disgust and Sadness are harder to model for the CLIP base model.

3.2 Fine-Tuning CLIP

In the second set of experiments, we evaluated the performances of fine-tuning the CLIP model for specific downstream tasks on the Image-Emotion dataset and on the CIFAR100 benchmark. The implementation details for the experiments are described in Section 2.2.

Table 3 reports on the results of the `Emotion-CLIP` model. For completeness, we also report the performances for each class.

It is interesting to notice how performances drastically improve by means of leveraging a fine-tuned model trained on images and small captions that describe and mention the emotion that is likely to be elicited when watching that image. Interestingly, the model and fine-tuning process is also rather sensitive to the

Table 2: Zero-shot classification results for `Emotion-CLIP` on the Image-Emotion Dataset.

	Precision	Recall	F1-Score
Amusement	0.80	0.45	0.58
Anger	0.46	0.37	0.41
Awe	0.38	0.75	0.50
Contentment	0.61	0.72	0.66
Disgust	0.36	0.10	0.16
Excitement	0.45	0.43	0.44
Fear	0.30	0.49	0.37
Sadness	0.30	0.20	0.24
Macro Avg.	0.46	0.44	0.42
Weighted Avg.	**0.52**	**0.49**	**0.48**
Accuracy			0.49

Table 3: 10-fold cross validation results for `Emotion-CLIP` on the Image-Emotion Dataset.

	Precision	Recall	F1-Score
Amusement	0.83	0.79	0.80
Anger	0.49	0.53	0.50
Awe	0.66	0.73	0.69
Contentment	0.80	0.64	0.70
Disgust	0.70	0.71	0.70
Excitement	0.68	0.62	0.65
Fear	0.37	0.55	0.44
Sadness	0.37	0.55	0.44
Macro Avg.	0.64	0.65	0.64
Weighted Avg.	**0.70**	**0.67**	**0.68**
Accuracy			0.67

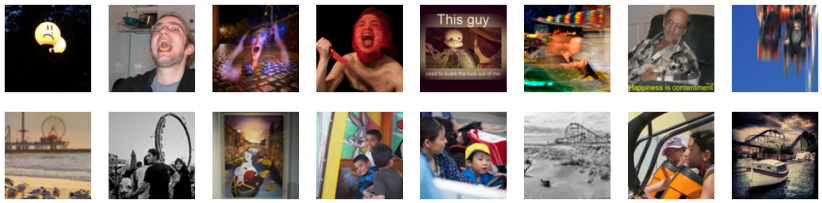
input captions that describe the labels. During the experiments we noticed in fact that captions that use more complex words, such as for example "an image that elicit < *emotion* >", or that are more direct in describing the image (e.g. "this image is about < *emotion* >") are consistently outperformed by models trained on a more simple yet specific and clear caption. While differences in performances are in the order of a few percentage points, i.e. 3-5%, it is nonetheless an interesting issue that could be explored further and more in-depth. Another interesting aspect that can be taken into account is the fact that performance vary rather widely across the different emotions. This may be due to the fact that describing (and thus recognizing) images eliciting certain emotions, such as Fear and Sadness, may be harder than with emotions such as Amusement and Disgust that may have more prominent visual features in the images. In addition to this, the size of the dataset and distribution of the labels must be taken into account as well. Interestingly, Disgust was the class for which performances were the worst in the zero-shot setting. Thus, in this case, it appears that the fine-tuning was rather helpful in pinpointing visual features of the emotion. Figure 1 shows some examples that highlight the differences between the zero-shot and the fine-tuned model. Specifically, we considered each caption (i.e. emotion) and show the top-8 images associated with that caption in the dataset extracted using zero-shot CLIP (top) and `Emotion-CLIP` (bottom). From the images, it is first and foremost clear that fine-tuning is very effective in learning better representations for the captions, and thus it is closer to images that actually represent the emotional content. Second, it is also interesting to notice that while the performances for classes such as Fear and Sadness are sub-par with respect to other emotions, the top-8 images actually represent them quite well. This may serve as an indication that fine-tuned CLIP models may be extremely helpful also for retrieval purposes.

Table 4 reports instead on the results of the `CIFAR100-CLIP` model. In this case, due to space concerns we report only the overall average performances of the model.

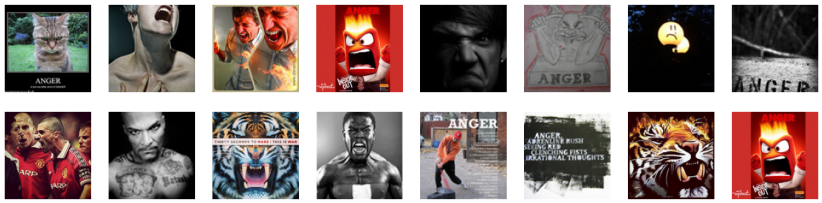
Table 4: 10-fold cross validation results for `CIFAR100-CLIP` on the CIFAR100 dataset.

	Precision	Recall	F1-score
Macro Avg.	0.82	0.81	0.81
Weighted Avg.	**0.82**	**0.82**	**0.81**
Accuracy			0.81

It is clear from the results that, even after only 1 epoch of fine-tuning, the model is closer to solve the CIFAR100 dataset with respect to the baseline CLIP model, with performances above 0.80 on all the considered metrics.



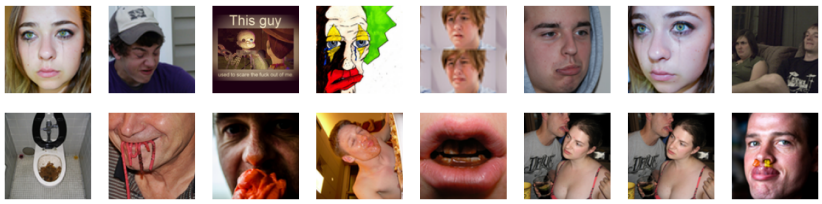
(a) Amusement



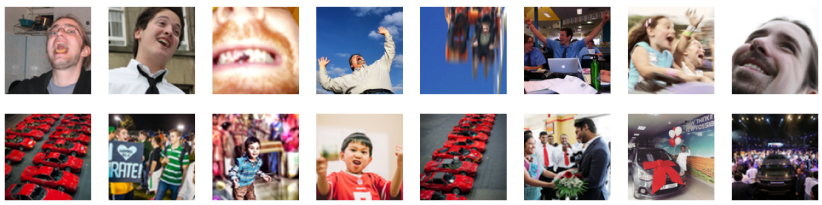
(b) Anger

(c) Awe

(d) Contentment



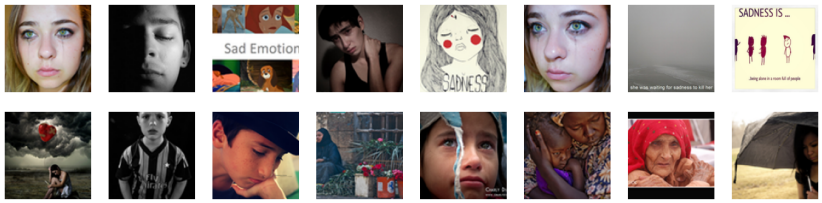
(e) Disgust



(f) Excitement

Fig. 1: Top-8 Images for each emotion (cosine similarity with the caption) with zero-shot CILP (top) and `Emotion-CLIP` (bottom).

(g) Fear



(h) Sadness

Fig. 1: Top-8 Images for each emotion (cosine similarity with the caption) with zero-shot CILP (top) and `Emotion-CLIP` (bottom) (continued).

It is also very interesting to notice that if we compare the results of `Emotion-CLIP` with those of `CIFAR100-CLIP`, we see that the differences in performances before and after the fine-tuning are similar for both experiments, with an improvement of around 0.20 across all metrics. This is interesting considering that the original model is much better suited to perform image classification tasks similar to the one of CIFAR100. We could speculate that, given a zero-shot capable model such as CLIP, the improvements in performances on downstream tasks and benchmark data may be limited by the architecture of the model itself.

3.3 Evaluation of fine-tuning on generalization capabilities of CLIP

In the final experiments, we evaluated the zero-shot capabilities of CLIP after fine-tuning on a different dataset, i.e. the extent to which fine-tuning on specific data may affect the zero-shot performances on different dataset. Recall that in order to do so, we first trained `Emotion-CLIP` and `CIFAR100-CLIP` on their respective dataset, with the same settings described in Section 2.2. Then, we exploited the fine-tuned models to perform classification on the other considered dataset. The details of the experiments are described in Section 2.3.

Results of the experiments are shown in Table 5.

If we analyze the results of leveraging fine-tuned CLIP for different tasks, we can identify an interesting trend. We saw in Section 2.2 that fine-tuning for a specific task is effective in improving performances. In this case, both fine-tuned models perform worse than the `ViT-B/32` CLIP pre-trained model on a task they are not fine-tuned on. This is clearly expected as the models' weights are shifted towards the end goal of the downstream tasks. However, it is interesting

Table 5: Results of applying fine-tuned models to a different dataset under zero-shot settings.

Model	Test Data	Precision		Recall		F1-score		Accuracy
		Macro	Weighted	Macro	Weighted	Macro	Weighted	
`Emotion-CLIP`	CIFAR100	0.57	0.57	0.41	0.41	0.40	0.40	0.41
`CIFAR100-CLIP`	Image-Emotion	0.43	0.50	0.30	0.39	0.30	0.34	0.39

to notice that both the experiments show rather similar degradation of the performances. In fact, both the models lose between 15 and 20% of F1-Score when tested on a different benchmark. This is interesting if we consider the nature of the training set of CLIP and its performances on simpler tasks with respect to more complex and/or abstract ones. The CIFAR100 dataset is definitely more akin to the original training set with respect to the Image-Emotion dataset, thus the resulting model should be more similar to the original one in terms of weights, i.e. it has to learn less about the classes. On the other hand, addressing image emotion classification starting from a pre-trained model requires a deeper adaptation of the model. This is also proven by the fact that `CIFAR100-CLIP` needed only a training epoch to learn the dataset, while `Emotion-CLIP` needed three. However, the relative closeness between `CIFAR100-CLIP` and the original CLIP model does not avoid the performance degradation in zero-shot settings on the image emotion classification. Notably, such a degradation is similar to the one detected by performing zero-shot classification on the CIFAR100 dataset starting from a model specialized on detecting emotions.

4 Conclusions and Future Works

In this paper, we have provided an evaluation of CLIP for the detection of emotions elicited by images. We experimented with the model both under zero-shot settings and by leveraging a fine-tuning strategy, and evaluate the advantages and drawbacks of both also in comparison with a more straightforward computer vision task. Exploiting CLIP as a zero-shot classifier provides good and rather inexpensive out-of-the-box performances on image classification, while for image emotion recognition the obtained results still show a wide margin of improvement. By leveraging fine tuning, we saw a significant improvement, similar in both considered tasks, but at the cost of generalization. A fine-tuned model on a specific downstream task performs worse than the base CLIP model on a benchmark it is not trained on.

The obtained results provide an early insight into exploiting state-of-the-art multi-modal models to characterize the emotions elicited by images, and thus on more abstract and subjective tasks. In the future, we plan to extend this line of research by leveraging diverse models and datasets. To this extent, we plan to create a new dataset in which the emotive labels associated with images are provided with textual information describing the choice of the labelling,

according to the annotation schema adopted in the ArtEmis [1] dataset, which is focused on art. Moreover, we plan to face the emotion recognition task as multi-label problem, in order to better learn how emotional texts can be associated to images and vice versa. Finally, we plan to perform a more in-depth and systematic study on the impact of the generated captions on the final model quality.

References

1. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.: Artemis: Affective language for visual art. CoRR **abs/2101.07396** (2021)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
3. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 223–232 (2013)
4. Chatterjee, A., Narahari, K.N., Joshi, M., Agrawal, P.: SemEval-2019 task 3: EmoContext contextual emotion detection in text. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 39–48. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10578–10587 (2020)
6. Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., Davis, B.: SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 519–535. Association for Computational Linguistics, Vancouver, Canada (Aug 2017)
7. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) **51**(6), 1–36 (2019)
8. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
9. Jia, J., Wu, S., Wang, X., Hu, P., Cai, L., Tang, J.: Can we understand van gogh's mood? learning to infer affects from images in social networks. In: Proceedings of the 20th ACM international conference on Multimedia. pp. 857–860 (2012)
10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
11. Krizhevsky, A.: Learning multiple layers of features from tiny images pp. 32–33 (2009)
12. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020)

13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
14. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: SemEval-2018 task 1: Affect in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 1–17. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
15. Passaro, L.C., Lenci, A.: Evaluating context selection strategies to build emotive vector space models. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož (Slovenia) (May 2016)
16. Passaro, L.C., Lenci, A.: Less is more: a multimodal system for tag refinement. In: Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020). pp. 44–58 (2020)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
18. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8821–8831. PMLR (18–24 Jul 2021)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5100–5111. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
21. Teney, D., Anderson, P., He, X., Van Den Hengel, A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4223–4232 (2018)
22. Yang, J., She, D., Sun, M.: Joint image emotion classification and distribution learning via deep convolutional neural network. In: IJCAI. pp. 3266–3272 (2017)
23. You, Q., Luo, J., Jin, H., Yang, J.: Building a large scale dataset for image emotion recognition: The fine print and the benchmark. Proceedings of the AAAI Conference on Artificial Intelligence **30**(1) (Feb 2016)