An Interpretable Model for the Emotional Response Classification of Photography

Samuel Fields (sef2186)

Intro

Image emotion classification is a topic of increasing interest in the field of computer vision. With advances in machine learning, it is now possible to automatically identify emotions in photos with a high degree of accuracy. However, the inner workings of these algorithms are often opaque, making it difficult for users to understand why a particular emotion was assigned to a given photo. This can make it difficult to trust the results, and can prevent users from gaining insight into the emotional content of their images.

In this paper, I propose an approach to emotion classification for photography that is both reasonably accurate and interpretable. My method leverages a state of the art classification model, a Resnet50 CNN with weights trained on the ImageNet dataset, with a set of linearly connected layers learned on a corpus of images annotated with emotional responses. I then use the LIME (Local Interpretable Model-agnostic Explanations) library to provide locally interpretable explanations of the predictions made by the emotion classification model.

By applying my proposed pipeline to a small corpus of my own personal photography, I show that my model is able to classify some emotional genres of photography reasonably well and is able to provide a strong level of interpretability allowing users to better understand the results of the classification process, and to help to identify meaningful insight of the emotional content their photography.

With an interpretable model for the emotional classification of photography, I hope to empower photographers with the ability to better understand and improve the emotional content of their photographs. By providing explanations of the reasoning behind each emotion prediction, photographers will be able to gain insight into the emotions that are captured in their photos, and will be able to use this information to improve their work and create more powerful and engaging images. Additionally, the interpretability of the model will help to improve the trustworthiness of the emotion classification algorithm, allowing photographers to have confidence in the results and to use the model as a valuable tool in their creative process.

Related work

The emotional classification of images has been an active area of research in recent years, with a growing number of studies exploring various approaches to the problem. Such studies were abel to achieve very impressive results on a wide range of image genres. Zhang et al proposed a multilevel hybrid model that was able to achieve an 86% accuracy in the emotion response classification on a corpus that combined annotated images from sources such as artistic photography, abstract paintings, and digital images shared on social networking sites[1]. As a backbone to their neural speaker model, Achlioptas et al constructed a ResNet50 architecture trained on their Artemis dataset, which consisted of ~80k emotion annotated artworks, that was able to achieve an accuracy of 60.2% when guessing the dominant emotion from 8 possible classes[2].

In the task of interpretable emotional classification of photography, previous work has approached the area from a number of directions. Wang's paper focuses on understanding the affects of aesthic features, such as figure-ground relationship, color pattern, shape, and compisition, on the behaviour of emotion response image classification[3]. However, Wang's research fails to propose an effective pipeline to efficiently provide this understanding at a local level. Another part of Zhang et al's work in their multilevel hybrid model for emotion response classification explored the use of activation maps to explain the classifications produced by their model through the global understanding of the low-level visual features learned by their model. However, Zhang et al's use of low-level visual features does not provide a meaningful level of interpretation, especially at the local level.

As such, I can observe that, although current technology is able to classify the emotional response to a variety of image types, including digital images and art, little work has been made in an attempt to provide the local interpretation of image emotional reponse classifications, especially in efficient and innsightful manner.

---

[1] Zhang, H., Xu, D., Luo, G. et al. Learning multi-level representations for affective image recognition. Neural Comput & Applic 34, 14107–14120, 2022.
[2] Achlioptas, Panos, et al. "Artemis: Affective language for visual art." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[3] Wang, Xiaohui, et al. "Interpretable aesthetic features for affective image classification." 2013 IEEE International Conference on Image Processing. IEEE, 2013.

Methods

The corpus I propose to use is a combination of the Emotion Dataset[4] provided by You et al and the WikiArt Emotions Dataset[5] provided by Saif M. Mohammad. These datasets were used because they are the largest publicly available datasets of images labled by emotion response that I could gain access to. The Emotion Dataset is a collection of 23000 images collected from Flickr and Instagram and labeled by Amazon Mechanical Turk (AMT) workers. Each image is assigned five AMT workers to verify the emotion of each image across eight possible emotion labels: Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sadness; and, only images that have recieved at least three emotion labels from the five assigned annotators are kept. The WikiArt Emotions Dataset is a collection of 4105 artworks, mostly paintings, that were selected from WikiArt.org's collection. Each image was annotated by ten annotators and given labels across twenty possible emotions across three general categories: positive, negative, and other / mixed. Since both of these datasets contained different emotion classes, I decided to aggregate the emotion labels into just three possible categories: positive, negative, and other. The aggregations were decided by each paper's categorization of their emotion labels.

For the task of emotion response classification, I used the ResNet50 architecture for the emotional classification of photography. ResNet50 is a convolutional neural network (CNN) that has been trained on the ImageNet dataset, and has been shown to achieve state-of-the-art performance on a wide range of image classification tasks. One of the key advantages of using the ResNet50 architecture is that it allows us to leverage the vast amounts of data and knowledge that have been used to train the model on the ImageNet dataset. This can provide significant benefits in terms of the accuracy of our emotion classification algorithm, as the model has already been exposed to a wide range of visual patterns and features that are relevant to the task.

In order to further improve the performance of our emotion classification algorithm, I used transfer learning to fine-tune the weights of the ResNet50 model on our own dataset of photos with known emotional reponse. This allows me to adapt the model to the specific characteristics of the dataset, and can improve the accuracy of the model's predictions. To do this, I removed all the linearly connected layers of the original ResNet50 model and replaced it with an initial normalization layer, followed three fully connected linear layers with 512, 128, and 3 total hidden units, respectively. I then trained the new components first and then the whole model on the corpus.

---

[4] You, Quanzeng, et al. "Building a large scale dataset for image emotion recognition: The fine print and the benchmark." Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1. 2016.

[5] Mohammad, Saif, and Svetlana Kiritchenko. "Wikiart emotions: An annotated dataset of emotions evoked by art." Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018.

As mentioned previously, we used LIME for the locally interpretable emotional classification of our images. LIME works by perturbing ann input image in various ways and measuring the effect on the model's predictions. This allows us to identify the specific visual features and boundaries that are most important for the model's decision, and quantify its impact on the predicted classification of the image. Specifically, per a given input image, I was able to produce a heatmap that described and quantified what regions positively or negatively contributed to the model's predicted emotion response and by how much. This provides a clear and intuitive explanation of the reasoning behind each emotion prediction.

Experiment

Model Performance

On a test set with a stratified sampling of the emotion classes, the model was able to achieve an accuracy of 73%, which is state-of-the art for a ResNet architecture. A breakdown of the performance of the model across the different datasets and class labels is shown below.

WikiArt Emotions	Precision	Recall	F1-Score	Support
Negative	0.30	0.48	0.37	164
Other	0.68	0.61	0.64	244
Positive	0.72	0.59	0.65	412
Weighted Average	0.62	0.57	0.59	820
Accuracy	0.57			

Image Emotions	Precision	Recall	F1-Score	Support
Negative	0.63	0.71	0.67	2264
Other	0.00	0.00	0..00	0
Positive	0.85	0.77	0.81	4471
Weighted Average	0.78	0.75	0.76	6735
Accuracy	0.75			

WikiArt Emotions + Image Emotions	Precision	Recall	F1-Score	Support
Negative	0.60	0.70	0.64	2428
Other	0.44	0.61	0.51	244
Positive	0.84	0.76	0.80	4883
Weighted Average	0.63	0.69	0.65	7555
Accuracy	0.73			

I then constructed and annotated a corpus of my own photography. It contained 17 total images: 4 “Negative” labled, 4 “Other” labled, and 8 “Positive” labled. On this unseen corpus, the model obtained a 56% accuracy. A breakdown of the performance is shown below.

Photography	Precision	Recall	F1-Score	Support
Negative	0.44	1.00	0.62	4
Other	0.00	0.00	0.00	4
Positive	0.71	0.62	0.67	8
Weighted Average	0.39	0.56	0.49	16
Accuracy	0.56			

Although my photography corpus is a very small sample to generalize from, we can still make the observation that the model performs much worse on this corpus than it does on the corpus it was trained on. We can also make the observation that the model is not able classify any of the “Other” labled photography in this corpus, but it does perform reasonably well on the “Negative” and “Positive” labled photographs.

Image Emotional Response Classification Interpretation

I selected a correctly classified “Negative” image from my photography corpus to exemplify a possible practice of emotional response classification interpretation. The image selected (Figure 1) could be categorized in the genre of a low-light portrait photograph. As the photographer and

annotator of this image, what I initally considered to be the “negative” aspects of this photograph were the low-light background, the blank expression of the subject (infered behind the mask being worn), and the coldness of the concrete below the subject and of the grey bars on the side.

Figure 1: “Negative” emotional reponse, low-light potrait photograph

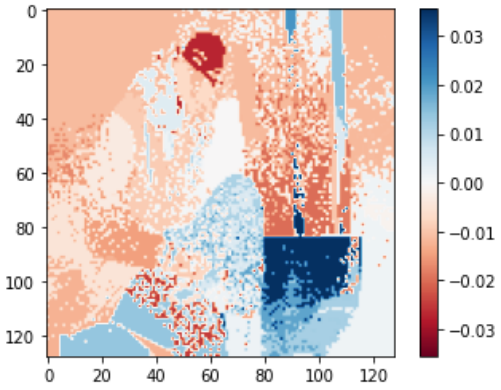

Figure 2: LIME heatmap of Figure 1

Observing the LIME heatmap of the image (Figure 2), in which the blue areas of the image convey areas of the image that contributed towards the “Negative” classification of the image and the red areas of the image are areas that contributed against scaled by the intensity of the blue / red hue, we can make the following observations:

1. The model interpreted the dark background as a slightly “not Negative” aspect of the image.
2. The subjects face was interpreted as strongly “not Negative.”
3. The concrete below the subject, the grey bars, and the subjects legs were the primary “Negative” aspects of the image.

Essentially, from this practice I would conclude that the model interpreted the monotone features of the image, the grey and grey-ish portions, as the contributing “Negative” features of this photograph. While I would argue that these are not the only contributing features, I do find it incredibly interesting and insightful that the monotone color of the subjects jeans were a contributing feature. Although I did not consider the color of the jeans to be a meaningful aspect in the emotional response of the image, upon a guided reflection provoked by Figure 2, I did find myself in agreement with the model.

In this practice, we can see that, although the model is not entirely correct in its interpretation of the image since the observations provoked by the LIME heatmap stand partially in opposition to my intentions and opinions as the photographer, the practice of interrogating and interpreting the LIME heatmap provided a novel insight of the impact of monotone colors had on the negative

efficacy of the image and, thus, improved my understanding of the emotional response evoked by my photograph.

Conclusion

In this paper, I presented a new approach to the interpretable emotional classification of photography. My method combines the use of the ResNet50 architecture with transfer learning, and the LIME algorithm for interpretability, to achieve a strong performance on a corpus of images with known emotional content, although there is still ample room for development for performance on artistic photography.

My work represents an improvement in the field of interpretable emotional classification for photography. I have shown that, by providing a highly accurate and interpretable approach to this problem, there is potential to help photographers better understand and gain new insight of the emotional content of their photographs. Overall, my approach has the potential to enable a wide range of applications in the field of photography, and can hopefully help photographers to create more emotionally impactful images.