

# High-level context representation for emotion recognition in images

Willams de Lima Costa  
Voxar Labs, Centro de Informática  
Universidade Federal de Pernambuco  
Av. Jorn. Aníbal Fernandes, Recife, Brazil  
[wlc2@cin.ufpe.br](mailto:wlc2@cin.ufpe.br)

Lucas Silva Figueiredo  
Unidade Acadêmica de Belo Jardim  
Universidade Federal Rural de Pernambuco  
PE-166, 100, Belo Jardim, Brazil  
[lsf@cin.ufpe.br](mailto:lsf@cin.ufpe.br)

Estefania Talavera Martínez  
University of Twente  
Drienerlolaan 5, 7522 NB Enschede, Netherlands  
[e.talaveramartinez@utwente.nl](mailto:e.talaveramartinez@utwente.nl)

Veronica Teichrieb  
Voxar Labs, Centro de Informática  
Universidade Federal de Pernambuco  
Av. Jorn. Aníbal Fernandes, Recife, Brazil  
[vt@cin.ufpe.br](mailto:vt@cin.ufpe.br)

## Abstract

*Emotion recognition is the task of classifying perceived emotions in people. Previous works have utilized various nonverbal cues to extract features from images and correlate them to emotions. Of these cues, situational context is particularly crucial in emotion perception since it can directly influence the emotion of a person. In this paper, we propose an approach for high-level context representation extraction from images. The model relies on a single cue and a single encoding stream to correlate this representation with emotions. Our model competes with the state-of-the-art, achieving an mAP of 0.3002 on the EMOTIC dataset while also being capable of execution on consumer-grade hardware at  $\approx 90$  frames per second. Overall, our approach is more efficient than previous models and can be easily deployed to address real-world problems related to emotion recognition.*

## 1. Introduction

Among diverse possible representations of human behavior, emotion recognition is a research topic that gained interest in the last few years, mainly due to the high development of application scenarios that can explore affective computing concepts, such as smart environments and social robotics. By understanding not only the user but also the person, systems can act on top of behavioral knowledge and propose diverse interventions to improve interaction, experience, or even quality of life.

To recognize emotions, systems must first decode emotional signals sent naturally by humans, sometimes even

without intention [4]. These signals are usually referred to as *implicit* and are derived from a form of communication defined as *nonverbal communication*. Research shows that a significant amount of affective information is not communicated verbally but is sent and received naturally through this communication channel, which encompasses facial expressions, body language, speech tonality, and other emotional cues [22, 24]. This ability allows for an unobtrusive experience and a perception of emotion that is not posed.

However, in unrestricted in-the-wild scenarios, our emotions are also influenced by information in the situational context. For example, a person sitting on a beach, enjoying the sun and the sea during their vacation, is more likely to experience positive sentiments such as joy and happiness. In contrast, a person stuck in a traffic jam filled with noise pollution could be inclined towards more negative sentiments such as frustration, anger, or stress. Therefore, environmental stimuli should also be taken into consideration when analyzing emotion.

Researchers have been proposing approaches that take into consideration contextual information for a while now in works such as EMOTIC [11], CAER-Net [14], EmotiCon [20] GLAMOR-Net [13], and EmotionRAM [7], each proposing new approaches on how to leverage context and extract its representations from images on different datasets. We hypothesize that, although working on these low-level representational features could and has led to significant results in the past, generating semantic, high-level descriptions could be more assertive to unseen data, leading to better results in test sets and when deployed to solve real-world problems.

In some scenarios, high-level representations of emotions can be a valuable aid for decision-makers to make in-

formed choices. For example, consider a city planner that needs to decide which public parks in the city need to be improved or renovated first. To make this decision, the planner needs to know how people feel when they are in these spaces, but they may not need to know each individual experience to make this decision. Instead, the high-level representation could be provided as an overview of the emotions associated with that context and how people act towards it. It could be easily compared without needing to act on top of a significant amount of data. Approaches such as this one have several advantages, such as resource-saving.

In this work, we propose an approach for the extraction of high-level context representations of images for the task of emotion recognition. We show in our experiments that these highly-representative descriptions of context are capable of yielding results comparable to the state-of-the-art of emotion recognition on the EMOTIC dataset [12] by itself and could easily be placed into a complete emotion recognition pipeline as a context encoding module to lead to significant improvements in accuracy. We also show that our proposal can perform a fast inference, a desirable feature for low-consumption edge devices that could be deployed in the wild. The contributions of this work are as follows:

- We propose a novel framework that (i) builds a high-level representation of extracted context descriptors from images and (ii) employs Graph Convolutional Networks (GCNs) to classify these representations into emotional categories.
- We benchmark on the well-known EMOTIC dataset [12], achieving a comparable accuracy with the state-of-the-art.
- We discuss how our model’s low computational power requirements make diverse applications possible to solve real-world problems.

The rest of the paper is organized as follows: Section 2 provides an overview of works for emotion recognition from images. Our proposed framework for high-level context description is described in Section 3. Later, in Sections 4 and 5, we depict our experimental setup and discuss the obtained results, respectively. Finally, we draw conclusions in Section 6

## 2. Literature review

It is well-known that facial expressions contain very descriptive features related to emotion. According to Barret, faces are to us as words are on a page, and humans can decode them perceptually with some level of ease [2]. Naturally, researchers have been investigating how to extract these features in a task called Facial Expression Recognition (FER). However, some challenges come when evaluating FER in uncontrolled scenarios. The face is, many

times, partially or fully occluded, and techniques that depend solely on this information might not be able to perform well.

Extending these works, researchers have been investigating how adding other nonverbal cues could improve the pipeline. Kosti *et al.* [11] proposed a baseline for this approach, in which both the person and the context in which they are placed would be considered. Therefore, in the case of occlusion, there would be other contributors to perceived emotion. They also published the EMOTIC dataset that serves as a baseline for validating these types of approaches. EMOTIC is still used widely in the literature to validate techniques based on context.

Other techniques have also been proposed, such as CAER-Net [14] and GLAMOR-Net [13], which employ different formats of how to weigh context contributions and, therefore, how important should be the contextual information in each scenario. However, all of these techniques mentioned above have the same limitation by design: the lack of definition of what should be considered as context. For example, the approach proposed by [13] considers detecting the face of a person, completely occluding it with a black rectangle, and using this new image as a representation of context. However, the other body parts are still visible, as are the body parts of other people in the scene, and this image would be fed to a context encoding stream that is designed to extract features from the scene automatically. However, are these encoding streams capable of doing this task without prior knowledge?

Other approaches, such as EmotiCon [20], proposed that it is necessary to use multiple independent and specialist streams to generate representations that can be correlated to emotion, given how context is highly descriptive. Specifically for EmotiCon, as an example, the authors propose the usage of the following context streams: (1) multimodal context, with facial landmarks and body keypoints; (2) situational context, extracted by processing the background image with the person occluded by using a pedestrian tracking method; and (3) socio-dynamic context, which computes proximity features using depth maps.

In a more recent approach, Chen *et al.* [6] proposed models combining different representations from context. For example, they use a deep network for each person on the scene to calculate their social relations between intimate, not intimate, and no relation. They also propose a deep reasoning module for using multiple context representations that are extracted locally and globally and involve scene recognition and body pose estimation, among other modules.

However, humans perceive context differently [1,3]. The literature suggests that humans encode context naturally by using our internal representations of meaning in the image. Therefore, it is not natural for us to take calculated steps to



Raw caption: A bride and groom cutting a wedding cake

Processed caption: bride groom cut wedding cake

(a)



Raw caption: A man riding a wave on a surfboard in the ocean

Processed caption: ride wave surfboard ocean

(b)



Raw caption: A man and a girl sitting at a table with food

Processed caption: sit table food

(c)

Figure 1. Examples of images from EMOTIC with their raw captions and processed captions.

understand context. Instead, our brain automatically classifies these stimuli as positive, neutral, or negative based on our previous knowledge of that information [21]. Our approach differs from the techniques mentioned earlier due to the more straightforward approach for context, in which we try to mimic the context representation of humans based on our best knowledge of the literature on nonverbal communication and behavioral psychology.

### 3. High-level context representation

In this section, we describe our approach for extracting high-level representations from context. Given how humans describe and understand context in images, we propose extracting high-level descriptions of images to correlate them with semantic features. This approach mimics how humans correlate semantic descriptions with emotions to improve interpretability.

#### 3.1. Extraction of image descriptors

**High-level descriptions.** Given an image as input, we first want to extract high-level image descriptions. We employ ExpansionNet-v2 [9], an image captioning model based on the Swin-Transformer architecture [16]. We first traverse through the EMOTIC dataset, and for each sample, we input the image to ExpansionNet-v2 for captioning generation.

We then process the raw caption to generate a refined caption. First, we perform the removal of stop words from the caption. Stop words are common words in a language, such as articles, prepositions, and pronouns, but do not have any semantic meaning. Therefore, maintaining these words would only elevate the complexity, given their high frequency in the English language, and by removing them, we have a more representative corpus. We use spaCy<sup>1</sup>, a public library for natural language processing. We also remove

common nouns such as *man*, *woman*, *girl*, and *boy*. We also applied lemmatization to reduce each word to its root form. The remaining words are called *valid words* and will be used in the following steps to generate data representations. Finally, we show examples of images, their original captions, and their processed captions in Figure 1.

**Co-occurrence mining.** The second step involves the generation of co-occurrence matrices that will represent patterns of labels within the dataset, which will be employed in the future through conditional probability. After preprocessing the captions in the dataset, we store this information and count the occurrence of each emotion, and the valid words of each caption, resulting in a matrix  $M_c \in \mathbb{N}^{W \times C}$ , where  $W$  is the number of valid words from the corpus and  $C$  is the number of emotion categories in the dataset. Therefore,  $M_{c_{ij}}$  denotes the number of times that emotion  $C_j$  occurred when the valid word  $W_i$  also occurred. We call this matrix the emotion co-occurrence matrix. Based on the same assumption, we also generate a co-occurrence matrix based on the co-occurrence of valid words. Given a window of size  $s$ , we slide this window to capture the co-occurrence of the valid words, resulting in a matrix  $M_w \in \mathbb{N}^{W \times W}$ . Therefore,  $M_{w_{ij}}$  denotes the number of times that the valid word  $W_i$  appeared together with the valid word  $W_j$ .

**Semantic descriptions.** For each valid word  $W$ , we extract semantic representations that can be correlated with emotion. Given how ExpansionNet-v2 is a model to generate captions in a generic context, extracting the semantic representations of the word will lead to better representations of affective meaning in that caption.

For extracting semantic descriptions, we employ SenticNet [5], a knowledge base for semantics, sentics, and polarity associated with natural language concepts. We query each valid word, and we extract the following attributes: the

<sup>1</sup>Available at <https://spacy.io/>

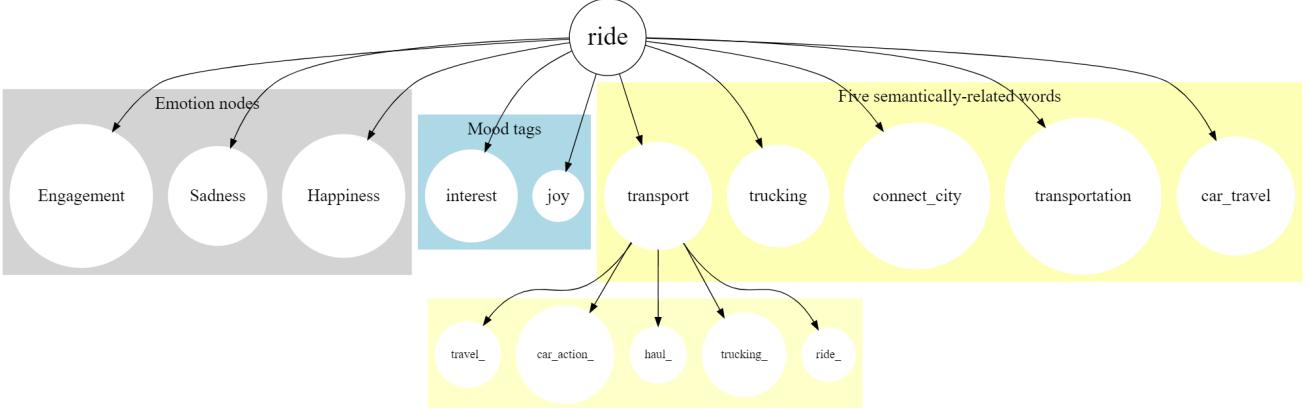


Figure 2. Example of the generated graph. For brevity and visualization, we consider only one valid word in this scenario. For each valid word node, we create nodes for emotions (also reduced for brevity), mood tags, and semantically-related words. For each semantically-related word, we also query SenticNet and extract their semantically-related words, as shown in the "transport" node.

two mood tags associated with the concept; the pleasantness sentic, which represents the perception of pleasantness or unpleasantness of the word; the polarity value, which represents the overall sentiment of the word and finally, the semantically-related concepts.

Except for the valid word not existing on the SenticNet knowledge base, we use WordNet [18, 19] to search for synonyms. The advantage of this approach is that the words in WordNet are grouped using synsets, which are sets of synonyms with similar concepts or meanings. Therefore, by querying a word in WordNet, its synonym will have a significant relationship and most surely have the same meaning. For each possible synonym, we rank the list according to the similarity with the valid word and iterate through the list, selecting the first synonym present in SenticNet. In the rare case that the valid word is not present on SenticNet and neither are its synonyms, we drop the valid word from the caption and proceed to the next step without it.

**Graph generation.** With this prior knowledge (e.g., co-occurrence and semantic representations), we can capture relationships between valid words and emotions and also between themselves. Given how they are a particularly effective method of describing structured data, we choose to model these representations using graphs. Although some of the knowledge is learned prior, the definition and construction of graphs are done as needed and in real-time. This allows this technique to generate representations from unseen data.

We use Deep Graph Library<sup>2</sup>, a framework-agnostic library for generating and manipulating graphs. We start by constructing an empty graph  $G = (V, E)$ , in which  $V$  is a set of nodes and  $E$  is a set of edges. In this case,

$V = E = \{\emptyset\}$ . For each valid word  $W$ , we start by adding a new node  $V_{W,i}$  to the before empty set of nodes  $V$  of the graph. We use GloVe [23] to fetch the valid word embedding and use this representation as the feature  $X \in \mathbb{R}^{50}$  for node  $V_{W,i}$ . If the valid word is absent on GloVe, we randomly sample this embedding from a uniform distribution [-0.01, 0.01]. We save this representation for future use in case this valid word reappears.

Next, we add a node  $V_C$  for each emotion category  $C$  in the dataset. For EMOTIC, since we have 26 possible emotions, we add 26 nodes and place edges  $e = (V_W, V_{C,i})$  between the valid word and each emotion. We define the weight  $w_e$  according to the equation below:

$$w_e = P(C_i|W) = \frac{M_{C_{W,i}}}{\sum M_{C_W}}, \quad (1)$$

where the edge weight  $w_e$  between the valid word node  $V_W$  and the  $i_{th}$  emotional category  $C_i$  is  $P(C_i|W)$ , which is given by the co-occurrence between the valid word  $W$  and emotion category  $C_i$  divided by the sum of the co-occurrence between the valid word  $W$  and all possible emotions  $C$ , extracted from the co-occurrence matrix  $M$ . Next, we add nodes related to the sentic semantic description of the word. First, we add two nodes relative to the mood tags extracted from SenticNet, and we set the weights of the edge between the valid word node and them to be the pleasantness value of the word. Next, we add five nodes relative to the five semantically-related words available from querying SenticNet and add edges using the polarity value as weight.

For each of the five semantically-related words to the valid word, we also query SenticNet and extract their five semantically-related words. The polarity value of the word in the first level gives the edge between these connections. We hypothesize that by adding another level of semantic

<sup>2</sup>Available at [www.dgl.ai](http://www.dgl.ai)

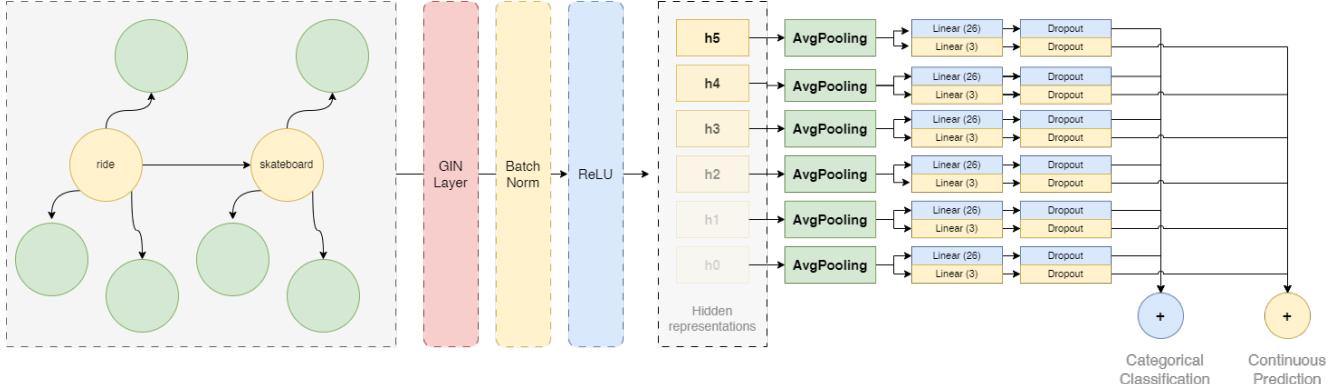


Figure 3. Our proposed architecture for high-level context representation. Given an input graph that is generated using previously learned knowledge and image captioning, we use an adaptation of Graph Isomorphism Network (GIN) [25] to classify it among a set of emotions. Given how EMOTIC also has annotations using a continuous model, we adapt the pipeline to generate two predictions, which are considered to calculate the loss.

relationships, we will be able to extract even deeper representations of context.

We perform the same process for each valid word  $W$  in the caption. After the nodes for all valid words are created, we add edges between these nodes. The weight of each edge is given by the co-occurrence of the words  $M_{W_{i,j}}$ , divided by the total number of times the word appeared. Finally, we show an example of the graph with reduced information for brevity in [Figure 2](#).

### 3.2. Deep GCN for Emotion Recognition

Given the construction of graphs to represent context, we use a deep graph convolutional neural network for graph classification and, consequently, for emotion recognition. Given a set of graphs  $G_1, \dots, G_N$  and a set of emotion categories  $C \in \mathbb{R}^{26}$ , we aim to classify each graph according to an emotional category. For this task, we propose adapting Graph Isomorphism Network (GIN) [25], chosen due to its simple architecture, which could lead to reasonable inference rates in low-energy, low-consumption devices.

First, given a graph as input, we store this graph's features directly in the hidden representations stack as  $h_0$ . After this, we loop through a GIN convolutional block containing a GIN layer, batch normalization, and ReLU. We iterate over this block five times in this approach, generating representations  $h_1$  to  $h_5$ . Finally, we iterate through the hidden representations, average pooling these features and reducing their dimensionality. In parallel, we keep a stream for the categorical classification, which outputs classification labels  $C$ , and another stream for continuous predictions for a VAD model. The VAD model [17] is a common approach for emotion recognition which represents emotions in a three-dimensional space, where each dimension corresponds to a different aspect of the emotional experience.

The VAD model is composed of Valence, Arousal, and

Dominance. The Valence (V) axis refers to how pleasant or unpleasant emotion is to the individual experiencing it. Arousal (A) represents the level of energy associated with the emotion; high arousal, for example, is related to more energetic, excited emotions, while lower arousal is to more calm emotions. Finally, Dominance (D) reflects the degree of control or power the person feels over this emotional state. Researchers have found that different emotions tend to cluster in regions of the VAD space, reflecting their shared aspects and underlying representation. Therefore, each emotion can be represented as a linear combination of the three axes [10].

Therefore, we learn categorical labels and continuous values during training. We define our loss as a weighted combination of the individual losses of each output. Given a prediction  $\hat{y} = (\hat{y}_{cat}, \hat{y}_{cont})$  in which  $\hat{y}_{cat} \in \mathbb{R}^C$  and  $\hat{y}_{cont} \in \mathbb{R}^3$ , we define the loss in this prediction as  $L = \lambda_{cat} L_{cat} + \lambda_{cont} L_{cont}$ , where  $L_{cat}$  and  $L_{cont}$  represents the loss of each individual prediction. For  $L_{cat}$  implement a weighted euclidean loss as used in EMOTIC [12], which is defined as follows:

$$L_{2_{cat}}(\hat{y}_{cat}) = \sum_{i=1}^{26} w_i (\hat{y}_{cat,i} - y_{cat,i})^2, \quad (2)$$

in which  $\hat{y}_{cat,i}$  is the prediction for the  $i_{th}$  category and  $y_{cat,i}$  is its ground-truth label. The weight  $w_i$  is defined as  $w_i = \frac{1}{\ln(c+p_i)}$ , where  $p_i$  is the probability of the  $i_{th}$  category and  $c$  is a parameter to control the range of valid values. We also employ a L2 loss for  $L_{cont}$ , defined as:

$$L_{2_{cont}}(\hat{y}_{cont}) = \sum_{j=1}^3 (\hat{y}_{cont,j} - y_{cont,j})^2. \quad (3)$$

Finally, we train our model on the EMOTIC dataset using the abovementioned features. We use the default Py-

Torch data loader and implement access to the list of graphs to feed the model during execution. We show an overview of our model on [Figure 3](#).

## 4. Experiments

**Dataset.** We perform our experiments on the Emotions in Context dataset (EMOTIC) [12], using the 2019 version, to allow direct comparison with the state-of-the-art. This dataset is focused on samples of people in their unposed environment and, therefore, comprises images with real, felt emotions and expressions. Another focus of this dataset is to have images with context visibility and variability, meaning that the images are framed with contextual information and that various contexts are present among the samples. The authors also gathered samples from other well-established datasets, such as COCO [15] and ADE20K [29], extending these images with emotion annotations.

EMOTIC was annotated using the Amazon Mechanical Turk service. The images on the training set were annotated only once, while the test and validation sets were annotated by three and five annotators, respectively. According to the authors, the dataset contains 23,571 images with 34,420 annotated subjects, 66% males and 34% females, 10% children, 7% teenagers, and 83% adults. Differently from what the literature on behavioral psychology suggests [8], EMOTIC employs 26 emotional categories proposed based on word connections and inter-dependence of words, forming, therefore, word groupings. However, the dataset lacks the *Neutral* category, which can harm the annotation process.

**Comparison with the state-of-the-art.** We compare our results with other techniques of state-of-the-art, namely EMOTIC [12], Zhang’s work [28], EmotiCon [20], DRM [6], LEKG [6], which are two variations of Chen’s method [6], and Yang’s work [26]. However, as we later describe in Table 1, these techniques are often built on top of a combination of multiple nonverbal cues.

**Validation metrics.** Besides the quantitative evaluation using the mean average precision (mAP) metric, as is done in the current literature [6, 11, 12, 20, 28], we present some examples to perform a brief qualitative evaluation of the predictions.

**Implementation details.** We train our model from scratch, learning the parameters using Adadelta [27]. After an empirical comparison of multiple values on the validation set, the batch size is set to 16. We use a learning rate of 0.001 and a weight decay of 0.0004.

Regarding the experimentation environment, we train and validate our model on a desktop computer running

| Technique     | mAP             | # of nonverbal cues                                                                 |
|---------------|-----------------|-------------------------------------------------------------------------------------|
| EMOTIC [12]   | 27.38<br>20.47* | 2 (body and context)                                                                |
| Zhang [28]    | 28.42           | 1 (context with two streams)                                                        |
| EmotiCon [20] | 35.48<br>26.87* | 4 (face description, body pose, context and depth mapping)                          |
| DRM [6]       | 26.48           | 5 (three body descriptions and two context descriptions)                            |
| LEKG [6]      | 29.47           | 2 (scene recognition and global context)                                            |
| Yang [26]     | <b>37.73</b>    | 5 (face landmarks, body pose, context, relationships, and agent-object interaction) |
| <b>Ours</b>   | <b>30.02</b>    | 1 (single-stream context)                                                           |

\* Chen’s reproduction of the work, as reported in [6]

Table 1. Quantitative evaluation of our approach compared with state-of-the-art models on EMOTIC dataset.

Ubuntu 20.04 LTS with an Intel i7-4790K with 32 GB of RAM and an NVIDIA RTX 2080 Ti with 12GB of VRAM. For training and experimenting with our model, we use PyTorch 1.12 with CUDA 11.3 and CuDNN 8.3.2. For the experimentations regarding inference time, we also compare it with a consumer-grade notebook with Windows 10 Pro, 16 GB of RAM, and an NVIDIA GeForce GTX 1060M with 6GB of VRAM.

## 5. Results and discussion

We compare our results with different approaches in Table 1. Our proposed method outperforms other graph-related methods, such as the work by Zhang, Liang, and Ma [28], and also DRM and LEKG, which are two variations of Chen *et al.* [6]. We did not compare our method with Chen’s TEKG model because it has a local approach that could not be extended to real-world problems by itself. Although EmotiCon [20] reported a higher mAP than our method by 5.46 mAP, according to Chen *et al.* [6], their result of 35.48 mAP is not reproducible, reporting a score of 26.87 mAP, in which our method can perform by 3.15 mAP. Additionally, EmotiCon uses four nonverbal cues, while we only employ one. Yang *et al.* [26] reports the highest mAP in this dataset, with a result of 37.73 mAP, which is 7.71 mAP higher than our result, by also employing five nonverbal cues. Therefore, we demonstrate that our model is competitive with the state-of-the-art, even with just one cue. While Zhang, Liang, and Ma [28] also use only one cue, they process it at two levels and can be considered two contextual cues.

The number of nonverbal cues employed is directly related to the inference time of the model, a question that we wanted to tackle. Emotion recognition models should be easily deployed on edge when thinking about real-world

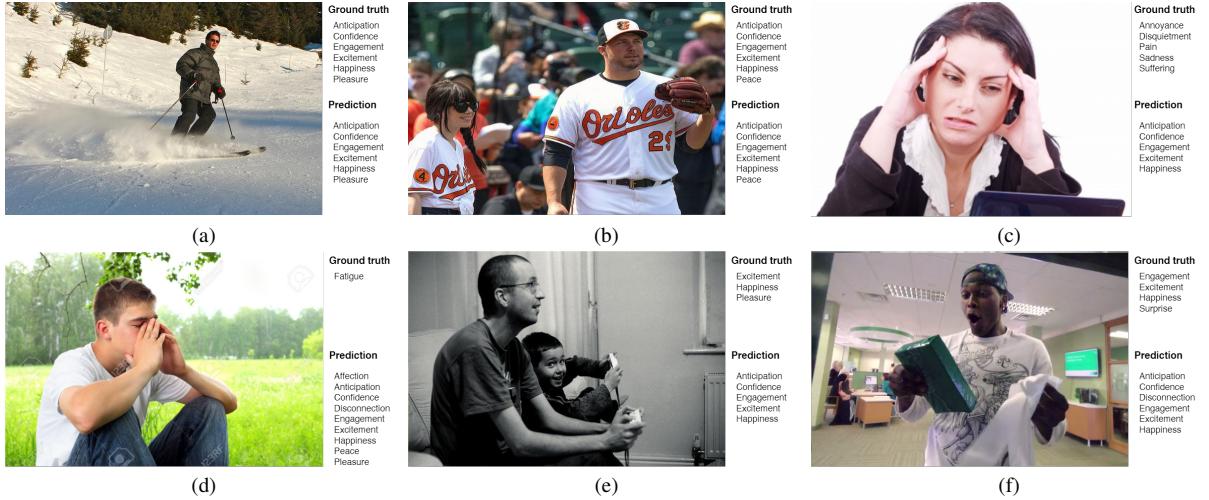


Figure 4. Qualitative results of our model on the EMOTIC dataset. For each image, we have the ground-truth emotion as annotated in the dataset and the prediction of the network.

| Method                | mAP    |
|-----------------------|--------|
| GIN [25] (AvgPooling) | 0.3002 |
| GIN [25] (SumPooling) | 0.1715 |
| Simple GCN            | 0.2505 |

Table 2. Ablation study of the proposed method.

situations. This would allow for multiple data capture and processing points without increasing spending too much or requiring high energy consumption. These are two current barriers imposed when deploying deep learning models in the wild. However, in cases where multiple cues are needed, our model could act as a context encoding stream, even with a convolutional neural network, to extract descriptions and contribute to the overall perception of emotion.

We conducted ablation studies on our model to assess its performance under different configurations. Our findings, present in Table 2, indicate that using sum pooling instead of the current pooling approach for GIN results in inferior performance. Furthermore, we also compared the performance of our model with that of using a simple GCN consisting of two GCN blocks, ReLU activations, and a classifier. Our model outperformed this simple GCN, which resulted in a lower mAP.

We also evaluate our model with a qualitative analysis, as shown on Figure 4. For each model on the EMOTIC test set, we feed this image to the proposed pipeline, generating categorical predictions for the image. Although the model can also generate results using the VAD model, it is difficult for humans to understand and compare these values since this is unnatural for us. Therefore, we choose to use only categorical values for the qualitative evaluation. This

experiment shows, as expected that our method looks for cues in context for emotion prediction. In Figure 4a and Figure 4b, the model could predict all categories present on the ground truth. In this case, given how the context is representative, our model can act well and give an overview of the emotion in that scene. In opposite cases, such as Figure 4c, the context is not representative, and the network cannot predict any correct emotion class. In Figure 4d, the context is related to a set of emotions, for example, positive emotions, but the perceived emotion of the person is actually negative. In this example, when looking at the person, we can perceive an emotion related to tiredness, which is confirmed by the ground truth *Fatigue*. However, since the model does not look at face or body language from context, it perceives the wrong emotion. Finally, for Figure 4e and Figure 4f, the context is very generic, but the model can extract cues from it and classify correctly, at least on some level.

Finally, we test our model on different environments to assess the computational power required and the inference time. We execute the entire testing pipeline by setting a batch size of 1 for individual predictions and evaluate on both environments described in Section 4. We store each individual prediction into a list and then compute the minimum and average values. The minimum value indicates the sample in which the inference was faster, while the average value indicates the average inference time for the model. In a moderate deep learning machine, the inference of our model took 4.0264ms as a minimum, and 4.1546ms on average, leading to  $\approx 248$  fps and  $\approx 240$  fps, respectively. For a consumer-grade notebook, the inference of our model took 8.9597ms as a minimum, and 10.3898ms on average, leading to  $\approx 111$  fps and  $\approx 96$  fps, respectively. Finally, on

the same consumer-grade notebook without using CUDA, the inference of the model took 13.0021ms as a minimum and 17.7365ms on average, leading to  $\approx 77$  fps and  $\approx 56$  fps, respectively, on an Intel Core i7-7700HQ @ 2.80GHz CPU.

We do not compare our inference time with the other techniques we evaluated above since neither has official open-source implementations. However, we may infer that models such as EmotiCon, which uses various other deep learning models to extract and process specific cues, would take longer than ours for execution.

## 6. Conclusion

In this work, we present an approach for emotion recognition based on contextual cues of the image. Our proposed approach uses image captioning to generate high-level descriptions of the image. Then, together with data mining and semantic analysis of the words, we generate and employ graphs that are high-level representations of the context. Finally, we employ graph classification approaches to classify the graphs among a set of emotional categories. Although other authors are basing their approach on multi-modality and various descriptions of context, this simple approach is effective and comparable to the more robust models. Finally, we show that our technique is fast and requires low computational power, allowing for easy and cheap deployment in multiple scenarios.

For future works, we plan on combining these high-level representations of context into a multi-cue model that also works with faces and body language to tackle the cases in which the context did not match the perceived emotion of people on the scene.

## Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

## References

- [1] Lisa Feldman Barrett and Elizabeth A Kensinger. Context is routinely encoded during emotion perception. *Psychological science*, 21(4):595–599, 2010. [2](#)
- [2] Lisa Feldman Barrett, Kristen A Lindquist, and Maria Gendron. Language as context for the perception of emotion. *Trends in cognitive sciences*, 11(8):327–332, 2007. [2](#)
- [3] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011. [2](#)
- [4] Ross Buck. Nonverbal communication: Spontaneous and symbolic aspects. *American behavioral scientist*, 31(3):341–354, 1988. [1](#)
- [5] Erik Cambria and Amir Hussain. *Sentic Computing: a common-sense-based framework for concept-level sentiment analysis*. Springer Cham, 2015. [3](#)
- [6] Jing Chen, Tao Yang, Ziqiang Huang, Kejun Wang, Meichen Liu, and Chunyan Lyu. Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition. *Applied Intelligence*, 53(4):4201–4217, 2023. [2, 6](#)
- [7] Willams Costa, David Macêdo, Cleber Zanchettin, Estefanía Talavera, Lucas Silva Figueiredo, and Veronica Teichrieb. A fast multiple cue fusing approach for human emotion recognition. *SSRN preprint 4255748*, 2022. [1](#)
- [8] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. [6](#)
- [9] Jia Cheng Hu, Roberto Caviglioli, and Alessandro Capotondi. Expansionnet v2: Block static expansion in fast end to end training for image captioning. *arXiv preprint arXiv:2208.06551*, 2022. [3](#)
- [10] Agata Kołakowska, Wioleta Szwoch, and Mariusz Szwoch. A review of emotion recognition methods based on data acquired via smartphone sensors. *Sensors*, 20(21):6367, 2020. [5](#)
- [11] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1667–1675, 2017. [1, 2, 6](#)
- [12] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2019. [2, 5, 6](#)
- [13] Nhat Le, Khanh Nguyen, Anh Nguyen, and Bac Le. Global-local attention for emotion recognition. *Neural Computing and Applications*, pages 1–15, 2021. [1, 2](#)
- [14] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. *IEEE International Conference on Computer Vision (ICCV)*, pages 10143–10152, 2019. [1, 2](#)
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. [6](#)
- [16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10012–10022, 2021. [3](#)
- [17] Albert Mehrabian. Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies. 1980. [5](#)
- [18] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. [4](#)
- [19] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998. [4](#)
- [20] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege’s principle. June 2020. [1, 2, 6](#)

- [21] M Carmen Pastor, Margaret M Bradley, Andreas Löw, Francesco Versace, Javier Moltó, and Peter J Lang. Affective picture perception: emotion, context, and the late positive potential. *Brain research*, 1189:145–151, 2008. 3
- [22] Dipika S Patel. Body language: An effective communication tool. *IUP Journal of English Studies*, page 2, 2014. 1
- [23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 4
- [24] Philipp V Rouast, Marc TP Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, 2019. 1
- [25] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 5, 7
- [26] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. Emotion recognition for multiple context awareness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 144–162. Springer, 2022. 6
- [27] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 6
- [28] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019. 6
- [29] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 6