A Multimodal Visual Sentiment Analysis Framework Enhanced With Feature Pyramid Networks

Daniele Galletti[1], Valerio Ponzi[1] and Samuele Russo[2]

[1]Institute for Systems Analysis and Computer Science, Italian National Research Council, Rome, Italy
[2]Neuroimaging Laboratory, IRCCS Santa Lucia Foundation, Rome, Italy

Abstract

Visual Sentiment Analysis aims to understand how images affect people in terms of evoked emotions. This paper presents a complete pipeline for comparing users' emotional responses to images, enabling the analysis of potential discrepancies between machine-inferred and subjective affective states. The proposed framework consists of three main stages. The first stage employs a Convolutional Neural Network (CNN) enhanced with Feature Pyramid Network (FPN) layers to extract multi-scale visual features. Experimental results show that incorporating three additional FPN layers improves performance while introducing only a negligible increase in model complexity. In the second stage, a multimodal approach is adopted, where visual features are integrated with textual features derived from captions generated by an Image Captioning model. This fusion enriches the emotional context by combining visual and linguistic cues. In the final stage, a grounding mechanism is applied to align and merge sentiments from the different modalities into a unified representation. The algorithm's output is then compared with the sentiment expressed by the user, enabling an analysis of the divergence between machine-inferred and human-perceived emotions.

Keywords

Visual Sentiment Analysis, Feature Pyramid Network, Multimodal Evaluation

1. Introduction

Sentiment Analysis is a well-known field in machine learning. The goal of sentiment analysis is to measure how certain topics affect people. The outcomes of this study are very important: having the perception of what the common opinion is, influencing political, economic and social aspects of an entire population [1, 2]. Despite its large use on text corpus and the huge availability of data coming from social platforms, sentiment analysis is still far from achieving always good reliability. The lack of context, the differences between languages and cultures, create, in fact, very important barriers which make sentiment classification a difficult task. Visual Sentiment Analysis (VSA) [3, 4, 5, 6] was born as an additional instrument to understand people's sentiment. It emerged in the last decade, gaining traction with the increasing use of images to express opinions on social media platforms. Images offer an additional channel capable of expressing much more information than text [7]. Images convey both semantic elements (e.g., objects, scenes) and emotional nuances, offering a richer medium than text. For this reason, social media platforms became very popular and VSA, consequently, started to grow. In this work, we present a multimodal sentiment extraction pipeline. This pipeline aims to give a framework to assess how an image is classified in terms of evoked sentiment. The pipeline is built around three main stages. In the first stage, visual features are extracted from the image using an artificial neural network. In the second stage, a neural captioning model generates a description of the image, and in the third stage, the features from the first and the second stage are mixed into a common representation.

The CNN we use in the first stage is a novel architecture that integrates FPN layers into a CNN [8]. This model aims to extract meaningful features at different scales, having the benefits of a CNN for object detection and also exploiting low level features which have proven to be useful for sentiment classification [9]. The model achieves better results when compared to its predecessor [3, 10] and more classical modeling techniques [11, 12, 13, 14]. In the second step, a textual description, coming from the Image Captioning model recently presented by Wang et al. [15], is added to the features extracted in the first step. The description offers an unbiased representation unaffected by the source of the data.

In the last step of the pipeline, a grounding technique is used to merge features coming from visual and textual data. Textual features are converted into a sentiment distribution using the Emotion Sensor dataset [16]. Visual features, which are in another domain of emotions, are similarly converted into the same representation by using an association between the labels of the two representations. This was done since labels in every representation used in this work are meaningful in terms of sentiment content.

The result is then presented to the user. The user's

ICYRIME 2025: 10th International Conference of Yearly Reports on Informatics, Mathematics, and Engineering. Czestochowa, January 14-16, 2025

ponzi@iasi.cnr.it (V. Ponzi)

feedback, in the form of an audio file, is converted to text using the Speech Recognition API [17] and the sentiment is extracted by using the same technique of the third step. The result is also presented to the user, along with the algorithm's result.

2. Related Works

Research in Visual Sentiment Analysis has evolved significantly over the past decade, intersecting computer vision, affective computing, and multimodal learning. One of the first paper presented in VSA field was in 2010 [18]. They did positive/negative classification using SIFT features extracted from images mixed with textual metadata associated with the image. Text to sentiment conversion was done using SentiWordNet [19], which was published the same year. The SentiWordNet corpus associates synsets to sentiment polarity. In 2013, Borth et al. [20] created a visual ontology in which sentiments in an image are represented by ANPs (adjective-noun pairs). In 2014 Chen et al. [3] presented DeepSentiBank, a CNN finetuned on the Flickr dataset which classified images into a 1553 (ANP) vector. This vector consists of a meaningful middle-level representation also exploited in this work. More recently, concerning the new CNN structures, Tianrong Raoa et al. [21] used a FRCNN (Faster R-CNN) based on FPN in order to extract the region of interest (RoI) in which sentiment is contained. Other region-based works on VSA were also presented [22]. Concerning recent studies, literature went towards multimodal extraction of features. In 2016 Katsurai and Satoh [23] used both hand-crafted features (SIFT and GIST) and text sentiment analysis on image metadata in order to predict the sentiment polarity. In 2018 Ortis et al. [24] used multimodal classification with visual features, metadata sentiment, and objective extraction of caption which was converted to text. Corchs et al. [25] presented a method that combines visual and textual features by employing an ensemble learning approach. In particular, the authors classified emotions by combining 5 state-of-the-art classifiers trained on visual and textual data. In recent studies, artificial intelligence systems have been successfully applied in real-life environments to assess and react to emotional states, as shown in psychoeducational robotics frameworks (Ponzi et al., 2021 [26]). Additionally, some recent approaches leverage eye-tracking data to infer user attention and emotional engagement with visual stimuli. These methods offer a complementary channel to multimodal sentiment analysis by correlating gaze patterns with affective responses [27, 28, 29, 30].

3. Datasets

In this project we used three different datasets. The sentiment extraction pipeline uses these datasets at different steps.

Flickr Dataset The first dataset used, the Flickr Dataset with CC, was created by Borth et. al. [20]. Images were automatically crawled from Flickr and filtered by their metadata, resulting in 487 256 weakly annotated samples. This dataset represents one of the first and most used dataset ever created for VSA tasks. Each of the 1553 classes is an Adjective-Noun pair (ANP), a mid-level representation for sentiment classification. To build this dataset the authors have crawled Flickr images and extracted textual tags associated with each sample. Most significant tags were then grouped and transformed into a set of pairs of adjective and noun. The pair adjective-noun, called ANP, represents a more emotionally charged concept instead of nouns and adjectives by themselves. Despite its large use this dataset presents some limitations. It is weakly annotated (categorized automatically by metadata posted by users on social networks) and thus subjected to bias. The dataset is also highly unbalanced, the classes in fact present a big variation of samples, going from 23 to 1402 samples per class. We used this dataset in order to finetune the neural network models trained on object detection tasks. Further details are presented in the Implementation and in Result sections.

Emotion Dataset The second dataset, published in 2016 [31] and available on Github, provides 23 308 images manually annotated using the 8 basic emotions presented by Mikels et al. [32]. The team started from 3+ million images weakly labeled; they filtered and annotated images by designing a task in which a group of people is asked to answer simple questions. From the results, they've built the largest manually created dataset up to then. As a motivation for the work, they discussed the predominance, on existing datasets, of images associated with Fear and Sadness emotions (Figure 2). This predominance can result in unbalancing classes, which can prevent an algorithm from working correctly. Emotion Dataset has offered a good benchmark option over the Flickr one since it is more properly classified, less biased, and less unbalanced. An example of images grouped by Mikels emotions is shown in Figure 1. In this work, Emotion Dataset is used to finetune the neural network models by adding a layer that maps ANPs representation (from the Flickr dataset) to Mikels emotions.

Emotion Sensor Dataset The third dataset used is the Full Emotion Sensor dataset [16]. This dataset associates

Figure 1: The emotions category presented by Mikels et. al. [32].

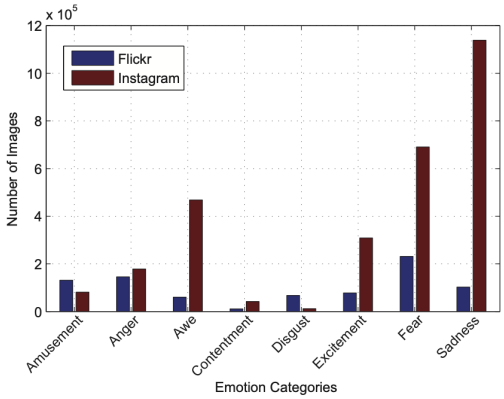

Figure 2: Data distributions of the images downloaded from Flickr and Instagram. Image from [32].

the most used 23 730 words coming from the internet to a distribution over 7 emotions. The dataset, whose preview at the current time is not available anymore [16], was created by collecting thousands of sentences from blogs and online posts. The authors then labeled manually and automatically the sentences using 7 emotions and calculated naive Bayes to classify words. The 7 emotions correspond to an extended version of the 6 Ekman basic emotions [33], by adding a neutral emotion in case of an equal distribution over the other 6. This dataset, made for NLP tasks, is used in this work in order to convert different representations of sentiment into a common one. The outcome of the algorithm will be a distribution over these 7 emotions.

4. Emotion Representation

There are several ways to represent a sentiment. Different psychological studies have led to different ways of representing human feeling in terms of basic emotions. In order to create and categorize data under some classes, both psychological studies and data analysis were performed. The most popular model in the literature is Plutchik's Wheel of Emotions [34]. This model defines 8 basic emotions with 3 valences each, resulting in 24 total classes. In this work, we used three different representations. The first, used in Flickr dataset with CC [20], is the ANP representation. It consists of pairs of adjectives and nouns which are meaningful in terms of the emotion's content. The second representation was introduced by Mikels et al. [32] and used in Emotion dataset [31]. It defines 8 classes of emotions as the results from an analysis on the IASP dataset. The third method was presented by Ekman et al. [33]. They found 6 basic emotions by categorizing facial expressions of individuals subjected to a test, which involved 10 different cultures. This representation was used in the Emotion Sensor dataset [16] by adding one additional neutral sentiment.

In this work, we tackle the problem of having different emotion representations by using a grounding technique that transforms all representations into one. Such technique assumes that there exists an association among different sentiment spaces since all the representations cover the same emotional content. The common representational model is chosen to be the extended Ekman representation, used in the Emotion Sensor dataset. Using this dataset, we convert the other two representations into a distribution over 7 basic emotions.

The conversion between Mikels' representation and Ekman's was performed using Mikels' labels. The labels are directly mapped into a distribution by the Sensor dataset. Some labels are common to both representations; thus, the output distribution presents a big predominance of that emotion (example shown in Figure 3). Some other labels can give problems connected to their distribution. The Sensor dataset can present, in fact, some non-coherent distribution due to the poor quality of data and the nature of the dataset. This is reflected in the conversion as shown in Figure 4.

The ANP representation is converted into the distribution over 7 emotions using the same technique. Each of the words of the pairs corresponds to one distribution; the output is the sum over the two distributions.

5. Models

5.1. Visual Sentiment Extraction

In this work, we use a Convolutional Neural Network to extract visual features from an image. The proposed CNN is a modification of a popular architecture for object detection [8]. We trained the architecture and tested it on the Flickr dataset as done by Chen et al. [3]. Aside from this, we've created a new architecture by introducing 3 Feature Pyramid layers. These layers extract low-level

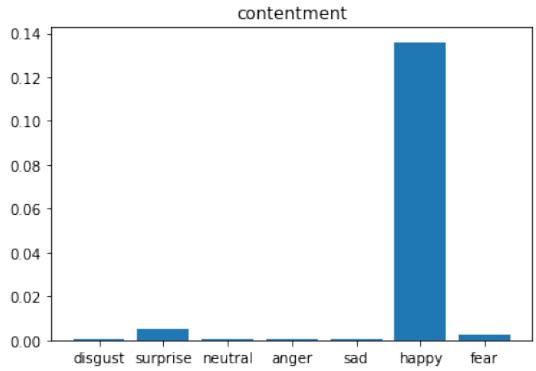

Figure 3: Example of Mikel label 'contentment' conversion into extended Ekman distribution.

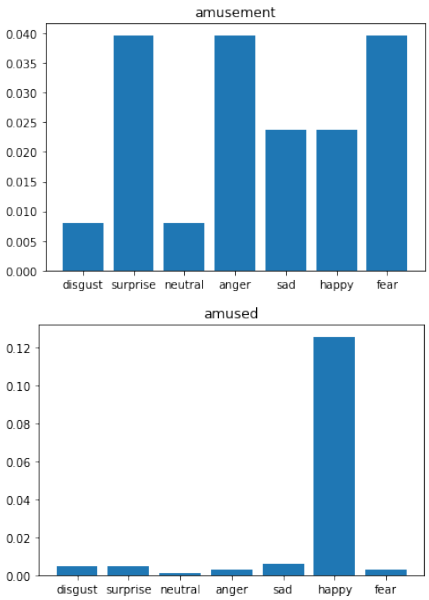

Figure 4: Extended Ekman distribution over 'amusement' word on top, and 'amused' word on the bottom.

features, complementing the high-level representations produced by the final convolutional layers. The backbone uses an AlexNet architecture common to both models. The architecture includes five convolutional layers, interleaved with max-pooling operations—except between the third and fourth layers. The last part consists of 3 densely connected layers with Dropout used at training time.

5.2. Feature Pyramid Network Model

Feature Pyramid Network was presented by Lin et al. [35], it combines low-resolution features with high-resolution features permitting the network to sensitize at different scales. As remarked in the original paper of FPN, a deep ConvNet computes a feature hierarchy, layer by layer, which produces feature maps of different spatial resolutions but introduces large semantic gaps caused by different depths. On the contrary, FPN produces predictions that are independently made on each level. In this way, the network maintains the semantic meaning of low-level features combined with the one coming from high-level features. The FPN functionality is composed of two pathways, the bottom-up pathway and the top-down pathway. The bottom-up pathway is the main pathway for feature extraction; each layer has a feed-forward connection to the next layer and a lateral connection going to the respective layer of the top-down pathway. The top-down pathway consists of the same layers as in the bottom-up pathway but reversed in connections. This time features go from the smallest layer of the pyramid to the biggest. Features are upscaled and summed to the lateral connections of the bottom-up part. Every top-down layer has its own lateral connection, which is the output of the pyramid. The implemented Feature Pyramid layers are three; we used the network's main feedforward branch as the bottom-up pathway. Parallelly, a top-down pathway creates three predictions which are fed into three smaller branches and merged by averaging with the final features extracted from AlexNet. The aim of introducing an FPN inside the model is to show how a small FPN with not many parameters (5M+ new parameters, AlexNet has over 61M parameters) can improve performances on VSA. This is justified by the need for different scales of meaningful features. As literature has shown, the sentiment contained in an image can arise not uniquely from objects in the image, but from color distributions, lighting, and other factors whose information can be lost in a deep CNN. Previous works remark the fact that a multimodal approach mixing low-level handcrafted features with high-level features extracted from the CNN brings the algorithm to better performances. The goal of this model is to prove that using an FPN structure one can achieve better performances without losing the benefits of having a fast and unique model. The FPN model was compared with the DeepSentiBank model [3] achieving better performances. Further information is shown in the Result section.

6. Data Bias

Data bias is a problem still present in Sentiment Analysis. It is connected to the different cultures, languages, and contexts in which different people live. Most of the

datasets for VSA are, in fact, crawled from the internet and automatically annotated from metadata. This way of proceeding can disadvantage the algorithm's performance, since the images can be wrongly labeled. Training a model by providing more input channels has been shown to be an effective way of tackling the bias problem [9]. Despite this, there is no manually annotated dataset that provides both image and text channels. Text is instead available in large, weakly labeled datasets crawled from the internet.

Some works tried to solve the bias problem by extracting objective features from data. These features do not come from the same source as the training data, but they are generated from the elaboration of another Machine Learning algorithm. The final features come from different joint algorithms' results. This approach has recently been revealed to be very effective [25], [24]. In this work, we adopt a similar approach to the one used by Ortis et al. [24]. We used an Image Captioning model to generate an objective description of the image. We then convert the caption to a sentiment distribution using the Emotion Sensor dataset [16]. The image captioning model used was recently presented in Wang et al. [15]. Once the caption is generated, relevant keywords are extracted for sentiment mapping. In order to filter keywords inside the phrase, we filtered English stopwords provided by the nltk corpus [36] and used the nltk POS tagger [37] and WordNet [38] to lemmatize the words if a correspondence is not found inside the Emotion Sensor dataset. As we will show in the Result section, the extraction of a neutral description is effective, but is nothing without a good (and unbiased) conversion into the sentiment distribution.

7. User's input

The user's input represents the second input to the system. The audio is converted into text using the Speech Recognition API for Python [17] and converted into sentiment distribution using the Emotion Sensor dataset [16]. This distribution is then presented to the user along with the result from the pipeline.

8. Implementation details

The captioning model was used in inference mode, it wasn't used at training time for speed limitations. The Flickr Dataset with CC [20] was resized before feeding the algorithm since originally it was 60 GB large, unfeasible to use in the settings described above. The resized dimension is 9 GB. Concerning the Emotion Dataset a filtering step was adopted since some images presented placeholders to indicate their unavailability. We removed them by using a hashing comparison which measures similarities between images. 1357 faulty images were found in the dataset, which in total remained with 21 951 samples. The Emotion Sensor dataset presented some lack of words useful in order to convert ANPs to sentiment distributions. These words, when converted, are replaced by their synonyms, provided by [39] and [40] English dictionaries. The synonyms, manually annotated, were organized in a file. The user's input is provided in audio file format.

9. Results

Results shown here are relative to the benchmark computed on the datasets presented above.

The first result is relative to the ANP classification task using the Flickr dataset. The dataset was split before training into 3 subsets: training, evaluation, and test set. Since the dataset is very unbalanced, we've created the test set such that at least a number of samples remained in the training set. In this way, classes with few samples are guaranteed to have at least a certain number of images in the training set. The minimum number was chosen to be 14. The two models involved are the DeepSentiBank and the FPN model, both share the same backbone (AlexNet) pre-trained on the ImageNet task. The metrics used to evaluate the model are the top 1, top 3, and top 10 accuracy, the same used in the DeepSentiBank paper [3]. The training was done using a Stochastic Gradient Descent optimizer with learning rate parameter set to 1e-3, weight decay to 5e-4, and momentum to 0.9. The learning rate was shrunk by a factor of 10 every 20 epochs. The batch size was 16 samples. Both models were trained for 40 epochs. Table 1 shows the best performances achieved by both models. As shown in Table 1 FPN model achieves +1% better performance in the three metrics with respect to the DeepSentiBank model. The low-level features extracted by the FPN layers contribute additional, complementary information that improves classification performance. The second result consists of the evaluation of the FPN model and the DeepSentiBank model on the Emotion Dataset. Both models were trained with a Stochastic Gradient Descent optimizer with a learning rate of 1e-3, a batch size of 16, and trained for 20 epochs. Model weights were initialized from the training on the Flickr dataset. The results presented in Table 9 are measured on test data. The results here confirm the previous statement about the FPN model. In this case, using a more balanced and unbiased dataset, the FPN reaches almost a +3% F1 score more than the DeepSentiBank model, confirming its potential in VSA tasks. The result of the FPN model training the last layer only reaches a comparable score with respect to the base finetuned DeepSentiBank model. This outcome is likely due to the fact that by training the last layer

Table 1
1553 ANP classification performances using top-k metrics.

model	top 1	top 3	top 10
deepsentibank	0.0747	0.1196	0.1948
fpn	0.0833	0.1332	0.2094

only, it creates a mapping between ANP representations and Mikels representations but with no improvement of feature extraction capabilities of the entire model. The last evaluation in Table 9 was done without finetuning the FPN with the Emotion Dataset but only using the conversion technique from ANP to Mikels representation. Mikels labels are converted into Ekman distributions as well as predicted ANP pairs. In order to evaluate this model, the absolute distances between the two sets of distributions were calculated, and the class with the minimum distance value was considered the one predicted. The results show the weakness of this method. Converting labels and ANPs can degrade accuracy. The Emotion Sensor dataset presents many issues of non-coherent distributions, which affect the accuracy of the conversion. As said before, these problems are connected to the nature of the dataset. The manual conversion is also used in the next results.

Table 9 shows the results of the evaluation on the same model as in Table 9 with multimodal evaluation. In this evaluation, also text features from the Image Captioning model are included. The captioning model by itself (with no visual features) reaches 0.21% precision, the fpn with manual conversion reaches comparable results. While comparing this result with the one without the captioning model (Table 9) it performs +6% better in F1 score. This outcome is due to the presence of the captioning model, which by itself achieves the same performance as with the FPN with manual conversion. On the other hand, the two models finetuned on the Emotion Dataset reach slightly lower scores with respect to their version without the captioning model. This is probably due to the conversion that was done in order to produce results in the same sentiment space by the Emotion Sensor dataset.

9.1. Emotion Conversion Results

In this section we present some examples concerning the conversion of the ANP to Ekman and of the Mikels to Ekman representations. The content of this section gives additional material which justifies the results above. Some examples of wrong ANP conversion are shown in Figure 5.

As depicted in the figures, representations may fall into outlier values. We can see that 'fluffy hair' ANP is associated with Fear as the predominant sentiment. This is because the Emotion Sensor dataset presents 'fluffy' as a fearful word. The same happens for 'illegal war' which results to be a happy ANP according to the Emotion Sensor dataset. The presence of such outliers in the Emotion Sensor Dataset can cause a wrong sentiment classification.

The Mikels conversion is less affected by this kind of outlier, having fewer classes. The 8 classes are almost all classified in a balanced way. The only class which is clearly not classified correctly is the 'amusement' class. As shown in the Emotion Representation section, the Emotion Sensor dataset in fact associates the 'amusement' word with a distribution which is not correct.

The issue of conversion affects also the captioning model, but no NLP evaluation test was done in this project.

9.2. Full Pipeline results

The full functionality of the pipeline is shown in Figure 6. We've chosen to not merge the output from the text and image extractions. The results are shown to the user as graphs that indicate the likelihood of belonging to a certain sentiment.

10. Conclusion

In this work we presented a pipeline that aims to be a systematic evaluation of a multimodal pipeline for automated sentiment inference from visual data. The full sentiment pipeline uses text, visual, and audio information in order to present a final result to the user. In this paper we focused the attention more on the Visual Sentiment task while leaving the other aspects to already developed algorithms.

We have demonstrated the effectiveness of FPN layers in the VSA task. Thanks to these layers, the network gains even more advantage using the Emotion dataset [31]. The FPN model has shown its improvement even by adding simple branches on the main backbone. The model used in this work was an old state-of-the-art network. It was used to have a direct comparison with the original paper in which ANPs were introduced. Many SOTA CNNs can be used for the same task. Future works could prove the performance gain by introducing FPN layers also in these novel structures.

This work attempts to address the multiple representation problem of sentiment by using an easy technique of conversion. We leveraged the Emotion Sensor dataset in order to extract a distribution associated with each word. The technique presented inaccuracies due to the need to have more solid bases on the dataset used. Further development could rely on more structured datasets, so that the last step performances can be improved.

Model	Precision	Recall	F1 score
Finetuned deepsentibank	0.4473	0.4356	0.4353
Finetuned fpn	0.5030	0.4756	0.4813
Fpn last layer only	0.4377	0.4283	0.4303
Fpn with manual conversion	0.1785	0.1579	0.1008

Table 2
Performance comparison of models on the Emotion Dataset using Mikels' 8 emotion classes. Metrics include precision, recall, and F1-score. Results show that the FPN-based model outperforms the baseline DeepSentiBank model.

Model (+ Caption)	Precision	Recall	F1 score
Finetuned deepsentibank	0.4474	0.4207	0.4238
Finetuned fpn	0.4851	0.4524	0.4571
Fpn with manual conversion	0.2150	0.1844	0.1695
Caption only	0.2147	0.1842	0.1693

Table 3
Performance of multimodal models combining visual and caption-based textual features on the Emotion Dataset (Mikels' 8-class scheme). The table includes results for models with and without fine-tuning, as well as the caption-only baseline. The FPN model benefits significantly from the addition of caption features.

Another promising direction for future work is the Sentiment Analysis on audio data. Future development of this kind can bring important improvements to pipeline stages.

The VSA problem is still far from being solved. Exploiting multimodality is the key to reach further results. We have seen, although, that with the actual settings, data bias and unavailability of unique representations can make VSA as well as Sentiment Analysis a very difficult task.

11. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's

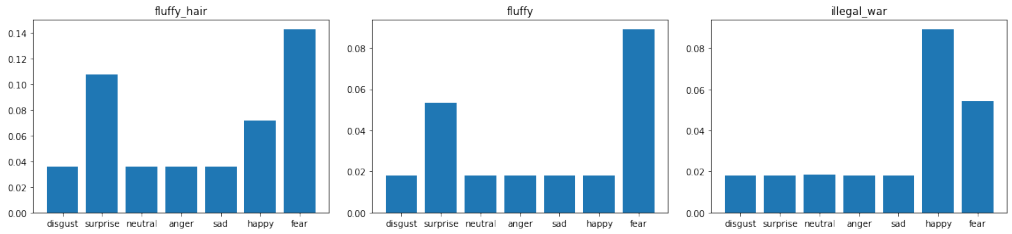

Figure 5: 'fluffy_hair' ANP wrongly associated using the conversion to extended Ekman. In the second image 'fluffy 'word sentiment distribution. In the third image 'illegal_war' ANP wrongly converted.

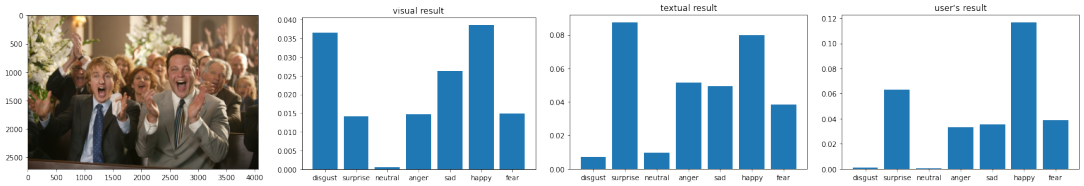

Figure 6: The output of the pipeline. The image in input on the left. The viual features result in the second image. The third image contains the caption automatically generated from the image. The user's input in audio format, transcribed as 'I feel happy', gives the result shown in the forth image. Image could be subject to copyright.

content.

References

[1] O. Oyebode, R. Orji, Social Media and Sentiment Analysis: The Nigeria Presidential Election 2019, in: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0140–0146. ISSN: 2644-3163.

[2] N. Brandizzi, A. Fanti, R. Gallotta, S. Russo, L. Iocchi, D. Nardi, C. Napoli, Unsupervised pose estimation by means of an innovative vision transformer, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 13589 LNAI, 2023, p. 3 – 20. doi:10.1007/978-3-031-23480-4_1.

[3] T. Chen, D. Borth, T. Darrell, S.-F. Chang, DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks, Technical Report arXiv:1410.8586, arXiv, 2014. ArXiv:1410.8586 [cs] type: article.

[4] I. E. Tibermacine, A. Tibermacine, W. Guettala, C. Napoli, S. Russo, Enhancing sentiment analysis on seed-iv dataset with vision transformers: A comparative study, in: Proceedings of the 2023 11th international conference on information technology: IoT and smart city, 2023, pp. 238–246.

[5] C. Randieri, A. Pollina, A. Puglisi, C. Napoli, Smart glove: A cost-effective and intuitive interface for advanced drone control, Drones 9 (2025). doi:10.3390/drones9020109.

[6] E. Iacobelli, S. Russo, C. Napoli, A machine learning based real-time application for engagement detection, in: CEUR Workshop Proceedings, volume 3695, 2023, p. 75 – 84.

[7] C. Napoli, C. Napoli, V. Ponzi, A. Puglisi, S. Russo, I. E. Tibermacine, Exploiting robots as healthcare resources for epidemics management and support caregivers, in: CEUR Workshop Proceedings, volume 3686, 2024, p. 1 – 10.

[8] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems, volume 25, Curran Associates, Inc., 2012.

[9] V. Ponzi, S. Russo, A. Wajda, C. Napoli, A comparative study of machine learning approaches for autism detection in children from imaging data, in: CEUR Workshop Proceedings, volume 3398, 2022, p. 9 – 15.

[10] C. Randieri, A. Pollina, A. Puglisi, C. Napoli, Smart glove: A cost-effective and intuitive interface for advanced drone control, Drones 9 (2025). doi:10.3390/drones9020109.

[11] G. Zimatore, C. Serantoni, M. C. Gallotta, L. Guidetti, G. Maulucci, M. De Spirito, Automatic detection of aerobic threshold through recurrence quantification analysis of heart rate time series, International Journal of Environmental Research and Public Health 20 (2023). doi:10.3390/ijerph20031998.

[12] M. C. Gallotta, G. Zimatore, L. Falcioni, S. Migliaccio, M. Lanza, F. Schena, V. Biino, M. Giuriato, M. Bellafiore, A. Palma, et al., Influence of geographical area and living setting on children's weight status, motor coordination, and physical activity, Frontiers in pediatrics 9 (2022) 794284.

[13] G. Zimatore, M. Cavagnaro, Recurrence analysis of otoacoustic emissions, Understanding Complex Systems (2015) 253 – 278. doi:10.1007/978-3-319-07155-8_8.

[14] M. C. Gallotta, V. Bonavolontà, G. Zimatore, S. Iazzoni, L. Guidetti, C. Baldari, Effects of open (racket) and closed (running) skill sports practice on children's attentional performance, Open Sports Sciences Journal 13 (2020) 105 – 113. doi:10.2174/1875399X02013010105.

[15] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, Technical Report arXiv:2202.03052, arXiv, 2022. ArXiv:2202.03052 [cs] type: article.

[16] J. Bil, Full Emotions Sensor Dataset Containing Top 23 730 English Words Classified Statistically Into 7 Basic Emotions, 2022.

[17] A. Zhang (Uberi), SpeechRecognition: Library for performing speech recognition, with support for several engines and APIs, online and offline., 2017.

[18] S. Siersdorfer, E. Minack, F. Deng, J. Hare, Analyzing and predicting sentiment of images on the social web, in: Proceedings of the 18th ACM international conference on Multimedia, MM '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 715–718.

[19] A. Esuli, F. Sebastiani, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006.

[20] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: Proceedings of the 21st ACM international conference on Multimedia, MM '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 223–232.

[21] T. Rao, X. Li, H. Zhang, M. Xu, Multi-level Region-based Convolutional Neural Network for Image Emotion Classification, Neurocomputing 333 (2019).
[22] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, L. Wang, Visual Sentiment Prediction Based on Automatic Discovery of Affective Regions, IEEE Transactions on Multimedia 20 (2018) 2513–2525. Conference Name: IEEE Transactions on Multimedia.
[23] M. Katsurai, S. Satoh, Image sentiment analysis using latent correlations among visual, textual, and sentiment views, 2016.
[24] S. Russo, F. Fiani, C. Napoli, Remote eye movement desensitization and reprocessing treatment of long-covid- and post-covid-related traumatic disorders: An innovative approach, Brain Sciences 14 (2024). doi:10.3390/brainsci14121212.
[25] S. Corchs, E. Fersini, F. Gasparini, Ensemble learning on visual and textual data for social image emotion classification, Int. J. Mach. Learn. & Cyber. 10 (2019) 2057–2070.
[26] V. Ponzi, S. Russo, V. Bianco, C. Napoli, W. Agata, et al., Psychoeducative social robots for an healthier lifestyle using artificial intelligence: a case-study, in: CEUR Workshop Proceedings, volume 3118, CEUR-WS, 2021, pp. 26–33.
[27] E. Iacobelli, V. Ponzi, S. Russo, C. Napoli, Eye-tracking system with low-end hardware: development and evaluation, Information 14 (2023) 644.
[28] E. Iacobelli, D. Pelella, V. Ponzi, S. Russo, C. Napoli, et al., A fast and accessible neural network based eye-tracking system for real-time psychometric and hci applications, in: CEUR WORKSHOP PROCEEDINGS, volume 3870, CEUR-WS, 2024, pp. 32–41.
[29] C. Napoli, V. Ponzi, A. Puglisi, S. Russo, I. Tibermacine, et al., Exploiting robots as healthcare resources for epidemics management and support caregivers, in: CEUR Workshop Proceedings, volume 3686, CEUR-WS, 2024, pp. 1–10.
[30] N. Boutarfaia, S. Russo, A. Tibermacine, I. E. Tibermacine, Deep learning for eeg-based motor imagery classification: Towards enhanced human-machine interaction and assistive robotics, in: CEUR Workshop Proceedings, volume 3695, 2023, p. 68 – 74.
[31] Q. You, J. Luo, H. Jin, J. Yang, Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark, Technical Report arXiv:1605.02677, arXiv, 2016. ArXiv:1605.02677 [cs] type: article.
[32] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, P. A. Reuter-Lorenz, Emotional category data on images from the international affective picture system, Behavior Research Methods 37 (2005) 626–630.
[33] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, Universals and cultural differences in the judgments of facial expressions of emotion, J Pers Soc Psychol 53 (1987) 712–717.
[34] R. Plutchik, Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION, in: R. Plutchik, H. Kellerman (Eds.), Theories of Emotion, Academic Press, 1980, pp. 3–33.
[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature Pyramid Networks for Object Detection, Technical Report arXiv:1612.03144, arXiv, 2017. ArXiv:1612.03144 [cs] type: article.
[36] NLTK :: Natural Language Toolkit, 2022.
[37] S. Loria, textblob-aptagger, 2022. Original-date: 2013-09-18T20:03:40Z.
[38] G. A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (1995) 39–41.
[39] Oxford Learner's Dictionaries | Find definitions, translations, and grammar explanations at Oxford Learner's Dictionaries, 2022.
[40] Thesaurus.com - The world's favorite online thesaurus!, 2022.