Semiotic-Based Construction of a Large Emotional Image Dataset with Neutral Samples

Marco Blanchini, Giovanna Dimitri, Lydia Abady,
Benedetta Tondi, Tarcisio Lancioni, Mauro Barni
University of Siena, Italy
m.blanchini@student.unisi.it

Abstract

Image Visual Sentiment Analysis (VSA) requires the availability of large annotated datasets, whose construction presents many challenges. The necessity of gathering a large amount of labeled images contrasts with the rigorous, but lengthy, process required for manual annotation based on psychovisual experiments, and with the automatic gathering of large amounts of data roughly labeled based on the sentiment analysis of the text accompanying the images, like captions, tweets and tags. An additional limitation is the scarcity of high-quality datasets with a neutral class, which forces the images to be classified into emotions even when the observers show no emotional activation. In this work, we present a scalable methodology rooted in semiotics and art theory for the construction of a 3-class (positive, negative and neutral) VSA dataset, enabling the downloading of a desired quantity of images while maintaining labeling coherence and accuracy. Based on the proposed methodology, we introduce and make publicly available a VSA dataset of over 100,000 images. To validate the quality of the dataset, we used it to train several classifiers and compared their performance with those of classifiers trained on other datasets. The results, we got, show that the classifiers trained on the new dataset provide better performance when tested on independent datasets, including those commonly used for psycho-visual experiments.

1. Introduction

Understanding and studying emotions is a complex endeavor, involving several disciplines, including neuroscience, psychology, and semiotics [15]. More recently, computer science has also played a major role in the study of emotions, leading to the birth of a new interdisciplinary scientific field referred to as affective computing [4]. Vision plays a fundamental role into the formation of emotions [31], for this reason, Visual Sentiment Analysis (VSA) is playing an increasingly important role in affective computing applications [32].

Motivation. VSA research requires the availability of high-quality image datasets labeled with the emotions aroused in the observers. The creation of such datasets poses several challenges. From a cognitive point of view, it is necessary to make sure that the emotion labels are not biased by personal experiences, tastes, or the cultural background of the annotators. The choice of the images to be included in the dataset is also critical, with few general guidelines available mostly rooted in the field of psychology and art theory [29]. From a technical point of view, the increasing use of artificial intelligence techniques for VSA requires the availability of large annotated datasets, containing tens or even hundreds of thousands of images [33]. The creation of large labeled VSA datasets poses additional challenges due to the impossibility of annotating a large number of images by relying on rigorous psycho-visual experiments run under strictly controlled laboratory conditions. At the same time, resorting to online massive labeling campaigns considerably increases the subjectivity and unreliability of the annotations. Usually, automatic VSA datasets construction goes through the collection of massive amounts of images from social networks, and their annotation based on the analysis of the text accompanying them, like captions, tweets, tags, and communication threads. Such a simple approach has several drawbacks, including the lack of control on the images included in the dataset, and a poor quality of the annotations, given that there is no guarantee that the emotion aroused by the images is the same emotion expressed in the accompanying text. Another possibility consists in retrieving the images from the web through text queries with a clear emotional meaning, and use the emotions associated to the queries as image labels. The way the queries are formed, though, is a crucial problem, since no systematic, theoretically sound, approach has been proposed so far.

Another problem is the scarcity of datasets with a neutral

images class, that is a class of images that do not elicit any emotions in the viewers. This lack raises several problems. For example, in the emotional analysis of social media images or advertising images, it is first necessary to determine whether an image generates any emotion, before identifying which specific emotion it evokes. In other cases, the lack of a neutral class may force the annotators to decide for an emotion even when the image does not activate any, or bias automatic emotion recognition tools towards a (possibly random) emotion when a neutrality decision would be the most natural choice. Some examples of images with a neutral content[1] are shown in Figure 1.

Figure 1. Examples of neutral images

Contribution. In this paper, we present a new scalable methodology rooted in semiotics and art theory, for the construction of a 3-class (positive, negative and neutral) emotional dataset, enabling the downloading of any desired quantity of images, while maintaining labeling coherence and accuracy. The proposed approach relies on a theoretical sound methodology to create text queries with a clear emotional characterization, and use them to retrieve the images that will form the dataset. To associate the queries to positive, negative and neutral emotions, we first observe that, with the exception of some iconic cases, subjects are not associated to positive or negative emotions *per se*: images with *kids* may convey both positive and negative emotions, the same goes for images with animals, human artifacts or any other subject. On the contrary, in human language the goal of conveying a positive/negative sentiment or judgement is usually entrusted to adjectives. As a second observation, we exploit the semiotic principle according to which meanings emerge from the contraposition of contrary terms (like *happy-sad*, or *male-female*) [18]. Following these observations, we form the queries used to retrieve (and automatically label) the positive and negative images of the dataset by associating a noun (like *kid, cat, house*) to two contrary adjectives with a clear positive and negative bias (like *happy-sad, beautiful-ugly*). To retrieve the images populating the neutral class, we exploit another semiotic principle, according to which neutrality emerges from the simultaneous contradiction of the contrary terms (*not-sad AND not-happy*) [18]. In fact, due to the way common image retrieval engines work, using queries like *not-happy AND not-sad kid*, would hardly return neutral images, so we rather use language synonymous available in the English language[2]. A possible problem with the approach outlined above, is that the selection of the subjects represented in the image may be arbitrary. To mitigate this problem, we dedicate part of the dataset to artistic images. The reason behind this choice is the capability of artists to intercept the way society represents emotions, and use this understanding to create images with visual stimuli capable of eliciting strong and universal emotions into the observers [13].

To validate the proposed dataset construction methodology, we used it to build, and make publicly available, a 3-class VSA dataset with over 100,000 images. We assessed the quality of the dataset, and indirectly of the dataset construction methodology, by training a deep learning classifier and comparing its performance with those of classifiers trained on other datasets. The results we got show that the classifier trained on the semiotic-based dataset provides better performance when tested on independent datasets, including the accurately labelled datasets commonly used for psycho-visual experiments.

The rest of the paper is organized as follows: in Section 2, we describe the main VSA datasets available today and their limitations. In Section 3, we introduce our new methodology to create VSA datasets. In Section 4, we describe how we used the proposed methodology to build a specific VSA dataset. In Section 5, we describe the experiments we run to validate the dataset. In Section 6, we summarize the main results of our work, discussing its limitations and suggest future directions.

2. State of the art

VSA datasets differ based on their goals, image labeling methods, and the types of emotions used for labeling. The primary distinction lies in the intended use: some datasets are created to train AI networks for emotional content classification, while others are used in psychological research to study emotional reactions to images. In this paper, we focus on the first category.

In terms of emotion classification, datasets either classify images continuously using the Dimensional Emotion

[1] It is possible that in some contexts, or based on the personal experience of the observer, some of these images raise specific emotions, however, they are likely to not elicit any emotions in the majority of the observers.

[2] Although we use English, our methodology can be applied in any language supported by image retrieval engines, allowing dataset construction to adapt to diverse cultural and linguistic contexts [42].

Space (DES) [38], or into discrete categories based on the Categorical Emotion States (CES) [38]. Discrete classification can vary from two classes (positive and negative) to more detailed categories such as Paul Ekman's six basic emotions [12] or Joseph Mikels' eight primary emotions [30]. Mikels' emotions can easily be adapted to positive and negative categories, whereas Ekman's emotions cannot, as they include the emotion of surprise, which cannot directly mapped into a positive or negative category.

A key challenge in constructing VSA datasets is the use of a reliable labeling methodology, which is crucial for creating effective classifiers and conducting psycho-visual studies on human neuro-cognitive responses. Labeling approaches include manual labeling, automatic labeling, and hybrid methods. Manually labelled datasets can be further split into datasets labelled by means of strictly controlled laboratory experiments based on psycho-cognitive theories (hereafter indicated as PC datasets), and datasets labelled by resorting to online interview platforms, with less stringent environment and procedural controls.

A well-known PC dataset is the Affective Picture System (IAPS), with 1200 images categorized into three classes (positive, negative, neutral), and extensively used in studies on emotional reactions. An alternative version, IAPS-a, includes 395 images categorized according to Mikels' 8 emotions [26]. Another popular PC dataset is the Geneva Affective Picture Database (GAPED), which contains 730 images labeled as positive, negative, or neutral [8]. The above datasets are widely recognized for the quality of the annotations, however, their size is by far too small for network training and general AI-oriented applications.

Manually labeled datasets with a size suitable for network training and AI applications algorithms include Emotion6 [34], with 1980 images labeled by 15 annotators based on Ekman's emotions, and FI [47], with 23,308 images each labeled by 5 annotators still based on Ekman's emotions. Unlike datasets based on psychovisual experiments, large VSA datasets, like Emotion6 and FI, rely on uncontrolled online labelling platforms, making them more susceptible to cognitive biases. To mitigate this problem, some datasets are built by using a mixed methodology, combining interviews with emotionally-oriented online queries. Examples include the Flickr dataset with 90,139 images and the Instagram dataset with 65,493 images [23], which have been built by first retrieving images based on emotionally-relevant queries, and then annotating the retrieved images based on 3 interviews for each image. Images for which the query and the interviews did not agree were discarded. Another dataset built by relying on a hybrid procedure is Emoset [44], containing 118,102 images each labelled by 5 annotators based on Mikels' eight emotion categories.

The last category of datasets includes extremely large datasets, built by using a fully automatic image retrieval and labelling methodology. The most popular datasets in this category include VSO [7] and T4S [41]. The Visual Sentiment Ontology (VSO) dataset contains 500,000 images retrieved and labeled[3] using emotionally charged adjective-noun pairs (ANP). The ANP approach builds the queries by identifying emotionally relevant adjectives and pairing them them with nouns. The choice of the pairs is based on usage frequencies. This limits the accuracy and generality of the dataset, since frequent associations with positive or negative adjectives do not necessarily reflect the subject's emotional value. T4S is an extremely large dataset containing 1.5 million images retrieved from from Twitter, and labeled by associating to the images the emotion derived from the analysis of the text accompanying the tweet. While large, fully automatic datasets suffer from lower quality compared to manually labeled ones. Labels derived from text and tags may not accurately reflect the image's emotion, and the datasets often contain heterogeneous images, with random distribution across classes. This heterogeneity makes it difficult to focus on emotionally relevant visual patterns, since the lack of common visual elements across the various classes hampers and effective detection of the patterns eliciting the emotions.

Another limit of currently available datasets is the scarcity of datasets with a neutral class. Most datasets force neutral images into emotional categories. For instance, datasets built based on Mikels' [44] [47] [48] or Ekman's emotions [24] [34] classify neutral images as evoking one of the basic emotions, even if not present. If we exclude the IAPS and the GAPED datasets, which have a very limited size, the only large datasets containing a neutral class are the Flickr, Instagram [23], and T4s [41]. In all these cases, however, the quality of the neutral class is quite limited, either because the images are retrieved by relying on queries which are not explicitly thought to retrieve neutral images, or due to unreliable labeling of neutral images based on supposedly neutral tweets.

Table 1 summarizes the main publicly available VSA datasets along with their most relevant features.

The dataset construction methodology presented in this paper, combines the simplicity of fully automatic approaches with a high labelling quality, allowing the construction of large datasets suited for general purpose or even context-specific AI application. The presence of a neutral class, makes it possible to use the proposed methodology also within contexts where images do not necessarily arouse positively or negatively biased emotions.

3. Methodology

The target of the VSA dataset construction methodology is the creation of a dataset with images belonging to three

[3] In VSO the labels belong to an extended version of Mikel's emotions [35].

Table 1. State-of-the-art VSA Datasets. The construction of some datasets (including ours) includes a simple manual pruning phase which does not affect the manual or automatic nature of the labeling procedure.

Dataset Name	# Images	Manual Labeling	Automatic Labeling	Emotion Model	# Classes	Neutral class
IAPS [27]	1182	Yes	No	Positive, negative, neutral	3	Yes
GAPED [8]	730	Yes	No	Positive, negative, neutral	3	Yes
Affective Image Classification [29]	1115	Yes	No	Mikels	8	No
Flikr-sentiment [40]	586 000	Yes	No	Positive, negative	2	No
VSO [7]	500 000	Yes	No	Positive, negative	2	No
Twitter I [46]	1269	Yes	No	Positive, negative	2	No
Twitter II [46]	603	Yes	No	Positive, negative	2	No
Emotion6 [34]	1980	Yes	No	Ekman	6	No
FlickrLDL [45]	10 700	Yes	No	Mikels	8	No
TwitterLDL [45]	10 045	Yes	No	Mikels	8	No
Emotic [24]	18 316	Yes	No	Ekman + 20 emotions	26	No
FI [47]	23 308	Yes	No	Mikels	8	No
Flickr [23]	90 139	Yes	No	Positive, negative, neutral	3	Yes
Instagram [23]	65 439	Yes	No	Positive, negative, neutral	3	Yes
T4s [41]	1,5 milion	No	Yes	Positive, negative, neutral	3	Yes
IESN [48]	1 milion	No	Yes	Mikels	8	No
Emoset [44]	118,102	Yes	Yes	Mikels	8	No
Ours	107 117	No	Yes	Positive, negative, neutral	3	Yes

emotional classes: positive, negative and neutral. The proposed methodology permits to automatically create a large dataset of labeled images, suited for the training of deep learning models for the prediction of the polarity of the emotion aroused in the viewers. We use the web as the source of the images due to its vast availability and its capacity of cultural homogenization [20].

The images are retrieved and labeled by relying on a set of basic semiotic principles to construct textual queries with the desired emotion bias. In particular, instead of labeling the images based on the emotions felt by human annotators, we rely on the expressive language forms used to represent specific emotional concepts, and use them to feed an image search engine. The retrieved images, are labeled according to the sentiment expressed by the query, but in the vast majority of cases, there is a strong convergence between the two. The dataset is enriched by the incorporation of artistic images. In this way, we leverage on the capability of artists to understand, exploit and influence the mechanisms whereby visual stimuli raise positive and negative emotions into the viewers. In contrast to other datasets built by relying on text queries to image search engines [7] [48], the way we build the queries ensures that the same subjects are present in every class, but represented with different emotional values. The presence of the same subjects in all the classes allows to use the images in the dataset to extract the visual patterns that are crucial to determine the emotion conveyed by the image, rather than on the subject of the image.

3.1. The semiotic square

Semiotics studies how processes of signification occur and which elements generate them [16]. It examines any kind of expressive forms, including written text and images. In the case of images, the goal of semiotics is to identify the visual forms that allow an image to convey meaning [25].

Semiotics theorizes that all forms of language, whether verbal or visual, express meaning through a system of opposites [17] [22]. For example, the term *beautiful* derives its meaning from its opposition to the term *ugly*. Similarly, the visual forms that represent a *happy boy* make sense to us because they are contrasted with the visual forms representing a *sad boy*. Given two opposite terms, we can use them to draw the upper side of the so-called *semiotic square* (see Figure 2). By starting from the opposite terms, we can identify the contradictory concepts, which explicitly contradict the terms of the upper side of the square. For instance, in the case of the opposite terms *beautiful-ugly*, the contradictory terms would be *not-beautiful* and *not-ugly*. The upper and lower sides of the square define, respectively, the *complex* and *neutral* axes. The neutral axis, identifies the absence of both the initial opposite terms and hence can be used to create queries characterised by the absence of polarity along the axis identified by the two initial concepts. In our work, we use the neutral axis to retrieve the images that populate the neutral class. The complex axis accounts for constructions wherein the opposite concepts coexist. This is the case, for instance, of images capable of eliciting poignant emotions, wherein positive and negative sentiments coexist together. The lateral sides of the square define complementary relationships (according to which *not-beautiful* is the complement of *ugly* and *not-ugly* the complement of *beautiful*). In this work, we do not explore the opportunities offered by the exploitation of the complex axis and the complementary relatinoships, leaving it for future work.

The semiotic square can be applied to any kind of *language* including text, images [14], and other types of media, such as movies. It is a very useful tool for analyzing emo-

tional and passionate meanings [19], and has found wide applications in psychology [10].

3.2. Semiotic-based construction of a VSA dataset

A leading principle of the dataset construction methodology that we are proposing, which contrasts with most previous efforts, is that positive and negative emotions should not be linked to specific subjects, rather we look for the visual forms and patterns that are responsible for eliciting a positive or negative emotion (the absence of such patterns, then, can be used to define the neutral class). For this reason, the textual queries used to retrieve the images are formed by a list of *noun-adjective* pairs, where the same noun is used to retrieve images belonging to all the classes. It is up to the adjective, then, to polarise the image emotion positively or negatively. The selection of nouns, instead, can be used to characterize the context wherein the positive, negative and neutral classes are defined, or to give a specific cultural flavor to the dataset. The choice of the adjectives is made by resorting to the semiotic principles underlying the semiotic square. Specifically, the positive and negative emotions are associated to a pool of opposite terms with a clear emotional meaning, like *ugly-beautiful* and *happy-sad*. The neutral class deserves particular attention. Following the semiotic square, neutral images should be retrieved by using queries where both the contradictory terms coexist, like *not-ugly AND not-beautiful house*. All the most popular search engines, however, are likely to give a wrong interpretation of queries built in this way, ending up to retrieve images with either beautiful or ugly houses. To get around this problem, we suggest to use synonymous capable to identify the neutral axis concept. For instance, the query a *not-ugly AND not-beautiful house* is replaced by queries like *an ordinary house* or *a normal house*.

The procedure described above requires the definition of a list of subjects to be associated to opposite adjectives and to the neutral axis. The choice of the subjects, however, is quite arbitrary and may invalidate the generality of the dataset. To mitigate this problem, we propose to include in the dataset a subset of artistic images. The reason for doing so is to exploit the capability of artists to capture the expressive forms used in our society to arouse emotions [5], and, in turn, to come out with new expressive forms that are likely to become mainstream [37]. The queries to artistic web sites are used only to retrieve images for the positive and negative class. In fact, it is arguable, that in the great majority of the case, artists are not interested in creating images that do not arouse any emotions. In this case, the query is formed by a single term obtained by the nominalization of the positive and negative adjective used for the general queries: *dirty* $\rightarrow$ *dirtiness*, *pure* $\rightarrow$ *purity*. No constraints are put on the subjects and visual forms used to represent the positive and negative concepts, trusting the artists' capability to use effective and general visual representations of the querying concepts.

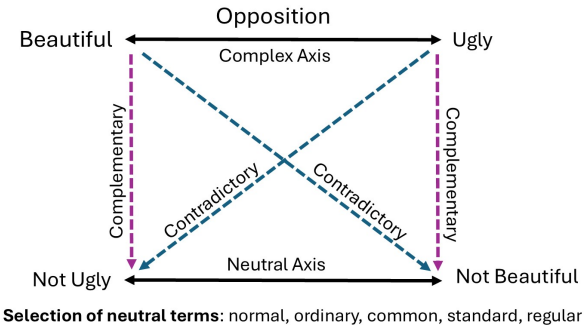

Figure 2. Application of the semiotic square to the pair of opposite adjectives *beautiful* and *ugly* with identification of the corresponding neutral adjectives.

To validate the proposed dataset construction procedure, in Section 4, we show a practical example of its application to actually build a large three-class VSA dataset. Then, in Section 5, we use the dataset to train a pool of neural network classifier capable of classifying a given image as either an emotionally positive, negative or neutral image. By comparing the results we got with those obtained with classifiers trained on other datasets, we see how systematically showing to the classifier images representing the same subjects, with visual forms having different emotional contents, results in better generalization capabilities when the classifier is faced with images coming from other, independent, manually labelled, datasets. A flowchart representation of the method is provided in the supplementary materials.

4. A new VSA image dataset

We now show how we applied the methodology described in Section 3 to construct a VSA dataset containing over 100,000 images[4]. We stress that the dataset described here is only one possible instantiation of the procedure outlined in Section 3. This procedure can be customized in various ways depending on the intended application.

We began by selecting a list of 20 nouns, each representing a different subject, chosen among the most common terms in the English language [9], like *city*, *bird*, *car*. For each term, we identified several pairs of adjectives with opposite emotional values, selected from those most commonly associated with the chosen noun. In total, we identified 40 pairs of opposing adjectives, like *beautiful - ugly*, *awful - wonderful*. Eventually, we also identified 15 neutral adjectives to account for the neutral axis of the semi-

[4]The dataset, code, and model of the algorithm trained using the CLIP Image Encoder architecture are available anonymously at the following link: https://mega.nz/folder/ncVxETpI#j1NMDDSGK_chAAmnJt3LbQ

otic square. Some examples of neutral adjectives are *normal*, *ordinary* and *regular*. The number of neutral adjectives is smaller than the number of opposite pairs because many neutral adjectives can describe neutrality across several pairs of opposing adjectives. The list of all nouns and adjectives we used for the queries is provided as supplementary material. For each noun-adjective query, we downloaded between 400 and 600 images for both the positive and negative pairs, and twice that amount for the neutral pairs. To download the images, we used the Google Image search engine, utilizing the advanced search option to locate images associated with consecutive noun-adjective terms. We ensured that only images freely usable for non-commercial purposes were downloaded [1]. Some examples of queries and the images retrieved through them are given in Figure 3.

Regarding artistic images, we downloaded between 400 and 600 artistic images for each noun derived from the nominalization of the positive and negative adjectives. The list of nominalized adjectives used for the queries is provided in the supplementary material. To select the artistic images, we utilized the internal search engines of pexels.com [2] and unsplash.com [3]. These two platforms offer a large collection of artistic images that are freely usable for non-commercial purposes. some examples of artistic images downloaded using the nominalized terms as queries are shown in Figure 3. Other examples of queries and the images retrieved through them mare provided as supplementary material. The dataset comprises only publicly available images that are freely usable for research, even when they depict individuals. Section 3 of the supplementary materials discusses in detail the issues related to consent and copyright, with a focus on the usability of the images.

After downloading, we curated the dataset by visually analyzing all the images to filter images that do not correspond to the query that was used to retrieve them. To avoid introducing a personal bias due to the preferences of the curator, filtering was carried out on the basis of a simple and reproducible procedure. Specifically, each downloaded image was manually checked to confirm the presence of the primary subject as specified by the query. For example, in the case of the query *fresh-apple*, we verified that the image indeed depicts an apple. We also ensured that the images did not contain drawings or synthetic illustrations. We did not apply any filtering to the artistic images. In the end, we collected 107,117 images. The positive class consists of 33,176 images, the neutral class 40,716 images, and the negative class contains 33,176 images. The neutral class is entirely composed of images obtained from neutral noun-adjective queries, while for the positive and negative classes, half the images were obtained from emotionally relevant noun-adjective queries, and half with artistic queries.

5. Dataset validation

In this section, we validate our dataset construction methodology. We do so by training several neural network classifier on the dataset described in Section 4, and showing that the classifiers trained in this way perform better than classifiers trained by relying on other datasets, both in terms of internal coherence, that is, when the classifiers are tested on images belonging to the same dataset used for training, and generalization capability, that is when facing with images belonging to independent datasets.

5.1. Experimental setting

We trained six distinct deep-learning classification networks on different datasets, and tested them both on a left-out subset of the datasets used for training (internal coherence experiments) and on the IAPS and GAPED datasets (cross-dataset experiments). In a first basic set of experiments we considered a 3-class classification task, including positive, negative and neutral images. This forced us to use only datasets providing the neutral class. To further assess the quality of our dataset construction procedure, we also used the dataset described in Section 4 to train two 2-class classifiers, by using only the positive and negative sections of the dataset, and compared the performance we got with those achieved by similar classifiers trained on Emoset, FI and VSO datasets.

Classifiers architectures. For our experiments, we used 6 image classification architectures: ResNet50 [21], ResNeXt101_32x8d [43], Swin-B [28], Swin-T [28], Vision Transformer (ViT_L_32) [11], and a CLIP Image Encoder [36] pre-trained with ViT-L/14.

ResNet50 is a convolutional neural network with 50 layers and residual connections to facilitate training. *ResNeXt101_32x8d* adds a cardinality parameter to the ResNet architecture, increasing the model's representational capacity. The "32x8d" design consists of 32 groups of convolutional layers, each with a cardinality of 8. Within the *Swin Transformer* family, we chose two models: Swin-T, which is characterized by a compact shape and good computational efficiency, and Swin-B, which has a bigger model size than Swin-T, resulting in enhanced expressive capability. *Vision transformers (ViT)* are a class of neural network architectures known for their ability to capture global dependencies in image data through self-attention mechanisms. Our experiments included the ViT_L_32 model, denoting a ViT variant with an "L" designation and an image patch size of 32x32 pixels. Lastly, *CLIP* is a multimodal architecture aiming at learning a joint embedding space for images and text. The CLIP image encoder, trained using a contrastive learning framework, extracts high-level semantic representations from input images. For our task, we leveraged a pretrained CLIP Image Encoder, adding a

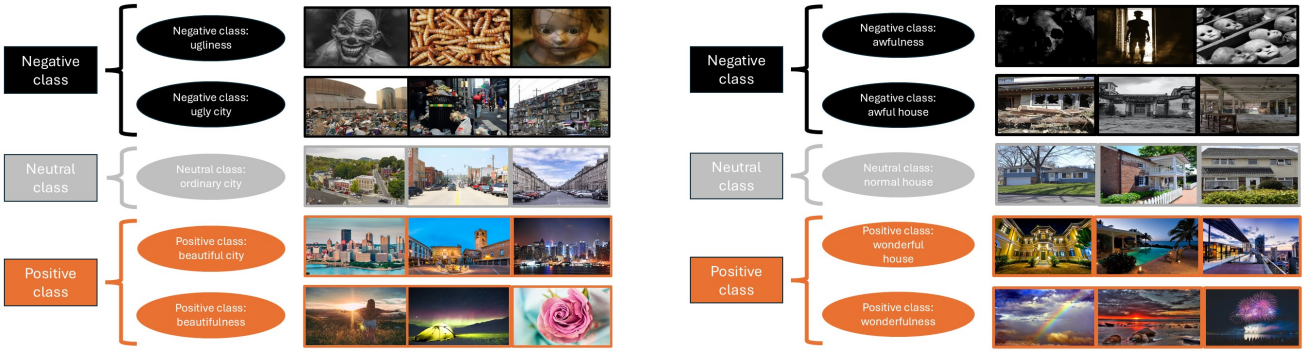

Figure 3. Examples of images retrieved by using semiotic and artistic queries. On the left: examples of images downloaded by pairing the subject *city* with the opposite adjectives *beautiful* and *ugly*, the neutral term *ordinary*, and by nominalizing the opposite adjectives into the terms *beautifulness* and *ugliness*. On the right: examples of images obtained with the noun *house*, the adjectives *wonderful*, *awful*, and *normal* and the nominalized terms *wonderfulness* and *awfulness*.

fully connected layer after the encoder.

Training. Each dataset used for training was split into training, validation, and test subsets sets with a ratio of 0.7:0.15:0.15. To enhance dataset diversity during training, various augmentations were used, including JPEG compression with random quality between 60 and 95, as well as vertical and horizontal flips. We avoided augmentations affecting colors, as it could impact the sentiment evoked by the images. Except for the CLIP Image Encoder, all the architectures were initialized with pre-trained weights from the ImageNet model. For the CLIP encoder, the weights of the encoder were frozen, and only the final fully connected layer was trained. Each model underwent 100 epochs of training, incorporating early stopping to ensure optimal performance. We employed an Adam optimizer with an adaptive learning rate controlled by a linear scheduler, initialized at 0.0001, and a batch size of 64. We resized all images to 224×224 for all networks except for the CLIP image encoder where we resized the input to a 336×336 size.

Datasets. For the internal coherence experiments, in addition to our dataset, we trained the 6 networks described above on the Flickr, Instagram and T4S datasets, which are the only large enough datasets featuring a neutral class. In this way, we obtained 24 classifiers, which were tested on the test subsets of the datasets used for training. Then we selected the best performing architectures, namely the CLIP image encoder and the Swin-t transformer, and tested them on IAPS and GAPED PC datasets, as independent cross-dataset tests of the performance of the classifiers.

2-class tests. In order to assess the validity of our dataset regardless of the presence of the neutral class, we repeated the cross-dataset tests by training the CLIP and Swin-t classfiers only on the positive and negative class of our dataset and on the positive and negative classes of Emoset, FI and VSO datasets. For Emoset and FI, the positive and negative classes were constructed by mapping the 8 classes of Mickel's system into positive (Joy, Trust, Surprise, Anticipation) and negative (Fear, Sadness, Disgust, Anger) emotions [30] [44]. We selected Emoset for excellent internal accuracy, while FI was chosen as the top-performing manually labeled dataset [47]. As to VSO, we chose it because it provides the best performance among two-class automatically labeled datasets and is labeled using a query system based on the association of subjects with emotionally relevant adjectives, but without a systematic method to build the queries. The comparison with VSO is particularly telling, since it shows that our semiotic-based method enables automatic labeling matching or surpassing the performance obtained by relying on manually labeled datasets.

5.2. Experimental results

In Table 2, we present the internal classification accuracy of the classifiers, described in Section 5.1, each evaluated on the test subsection of the dataset used for training. The results show that the CLIP Image Encoder consistently achieves the highest accuracy (at 88% for our dataset), while ViT_L_32 performs the worst. The results in the table indicate that the classifiers trained on the other datasets provides a noticeably lower internal accuracy compared to the models trained on our dataset. The experiments also show that for all the datasets, including ours, better performance are obtained with pre-trained architectures compared to non-pre-trained ones.

Table 3 reports the results of the classifiers with the best internal accuracy, namely CLIP image encoder and Swin-t, on two independent datasets, namely, IAPS and GAPED. Even in this case, the best results are obtained by the CLIP encoder trained on our dataset. The same applies to the Swin-t transformer, eventually demonstrating the capability of the dataset construction procedure we have presented to

Architecture	Flickr	Instagram	T4s	Ours
CLIP Image Encoder	75.24%	71.24%	59%	**88.64%**
Swin-b	69.48%	68.57%	52.32%	**86.61%**
Swin-t	70.29%	68.77%	52.55%	**86.44%**
ResNet50	69.79%	68.45%	52.1%	**84.18%**
ResNeXt101	68.96%	64.23%	51.61%	**84.81%**
ViT_L_32	66.98%	62.61%	49.14%	**78.35%**

Table 2. Internal classification accuracy of the various classifiers tested on the test subsection of the dataset used for training.

Test Set	CL. Flickr	CL. Instagram	CL. T4s	CL. ours
IAPS	75.15%	78.17%	56.43%	**84.41%**
GAPED	76.08%	77.22%	50.39%	**83.53%**
	Swin-t Flickr	**Swin-t Instagram**	**Swin-t T4s**	**Swin-t ours**
IAPS	64.17%	63.23%	49.12%	**69.70%**
GAPED	65.49%	62.10%	53.20%	**68.86%**

Table 3. Cross-dataset accuracy of the CLIP Image Encoder (CL.) and Swin-t classifiers, trained on Flickr, Instagram, T4S, and our datasets, and tested on the IAPS and GAPED datasets.

Test Set	CL. Emoset	CL. FI	CL. VSO	CL. ours
IAPS 2 Cls.	93.76%	90.28%	84.27%	**95.65%**
GAPED 2 Cls.	91.19%	86.57%	82.29%	**92.21%**
	Swin-t Emoset	**Swin-t FI**	**Swin-t VSO**	**Swin-t ours**
IAPS 2 Cls.	75.62%	69.38%	72.57%	**80.23%**
GAPED 2 Cls.	73.43%	56.00%	66.00%	**77.23%**

Table 4. Cross-dataset accuracy of the CLIP Image Encoder (CL.) and Swin-t classifiers, trained on the positive and negative classes of Emoset, FI, VSO and our dataset, and tested on the positive and negative classes of IAPS and GAPED datasets.

expose the common patterns linked to positive, negative and neutral (lack of) emotions. Lastly, Table 4 shows the accuracy of the CLIP image encoder and the Swin-t transformer trained to distinguish only positive and negative images of the IAPS and GAPED datasets, trained as explained in Section 5.1. Once again, the classifiers trained on our dataset provides superior performance compared to those trained on the other datasets, including VSO, that has been built using subject-adjective queries similar to ours. It is also noticebale that it does not apply a systematic method, as proposed in our approach. Noticeably, the classifier trained on our dataset outperforms that trained on the manually labeled FI dataset. Our dataset ensures better performance also with respect to Emoset, which is the best available large dataset built up to now, consisting of 100,000 images selected using basic emotions as queries and labeled by five annotators. The slight superiority of our dataset on the psychological datasets emphasizes how our method can produce VSA datasets with accurate labeling and emotional image classification capabilities that match or even surpass the best state-of-the-art manually labeled datasets.

6. Concluding remarks

We have proposed a scalable and easily customizable methodology to build a VSA dataset without resorting to expensive and time-consuming interviews. The methodology relies on semiotic and art-theory principles and is expressly thought to label images according to the polarity (positive or negative) of the sentiment they evoke into the viewers. It also allows to build a neutral class which may be particularly useful, for example, to understand the role of neutral images in online communications, or to separate emotional from neutral content. Instead of focusing on the emotions elicited in the observers, our methodology focuses on the visual forms used to express and represent emotions. As a byproduct of our research, we have built and made publicly available a new VSA dataset with over 100,000 images, split into positive, negative and neutral classes. The proposed methodology allows to create VSA datasets of any size and tailor them to specific needs. For example, it can facilitate the creation of a dataset to understand the emotions evoked by photos of houses, which could be useful for a real estate agency, or it could be targeted at photos related to an armed conflict to study the use of emotional images in war propaganda.

VSA faces inherent challenges due to the subjective nature of emotional perceptions and the strong connection between an image emotional meaning and the context wherein it is used. We have tried to mitigate these problems by making sure that the dataset includes representations of the same subjects with different emotional content, so to ease discerning the visual characteristics that determine the emotion conveyed by the images. Yet, subjectivity, cultural factors and contextualization remain an issue. For example, a neutral photo of a city might convey a positive meaning in a war-related context, implying that the city was not damaged. Similarly, a photo of a child with a neutral expression might generally be seen as positive image because it depicts a positive subject. However, when placed next to images of happy or sad children, it might take on a neutral meaning by comparison. The relationship between image meaning and context has been widely studied in semiotics [6] and can be addressed by a proper choice of the terms used to retrieve the images of the dataset. On this line, future work may focus on using semiotic principles to generate context-, cultural-specific VSA datasets.

An important yet unexplored aspect is the role of ambivalent images [39], which evoke both positive and negative emotions. In semiotics, these images align with the complex axis of the semiotic square, where opposing elements coexist. Understanding their role in visual communication presents unique challenges that require further study. An other possible extension regards the generalization of our procedure to more fine-grained labels.

References

[1] Google images. https://images.google.com/. Accessed: 29/7/2023 - 30/3/2024.

[2] Pexels.com. https://pexels.com/. Accessed: 29/7/2023 - 30/3/2024.

[3] Unsplash.com. https://unsplash.com/. Accessed: 29/7/2023 - 30/3/2024.

[4] Sitara Afzal, Haseeb Ali Khan, Imran Ullah Khan, Md. Jalil Piran, and Jong-Weon Lee. A comprehensive survey on affective computing; challenges, trends, applications, and future directions. *CoRR*, abs/2305.07665, 2023.

[5] Rudolf Arnheim. *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press, Berkeley, CA, revised edition, 1974 edition, 1954.

[6] Roland Barthes. Rhetoric of the image. *Visual culture: the reader*, 4:33–40, 1964.

[7] Damian Borth, R. Ji, Tao Chen, Thomas M. Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[8] Elise S. Dan-Glauser and Klaus R. Scherer. The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43:468–477, 2011.

[9] Mark Davies and Dee Gardner. *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates and Thematic Lists*. Routledge, 1st edition, 2010.

[10] Roberto De Luca Picione and Jaan Valsiner. Psychological functions of semiotic borders in sense-making: Liminality of narrative processes. *European Journal of Psychology*, 13(3):532–547, Aug 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[12] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6:169–200, 1992.

[13] C. Fiedler. *On Judging Works of Visual Art*. University of California Press, 1957.

[14] J.M. Floch. *Visual Identities*. Continuum studies in semiotics. Bloomsbury Academic, 2000.

[15] Elaine Fox. Perspectives from affective science on understanding the nature of emotion. *Brain and Neuroscience Advances*, 2, 2018. PMID: 32166161.

[16] A.J. Greimas, P. Perron, F. Collins, and F. Jameson. *On Meaning: Selected Writings in Semiotic Theory*. Theory and history of literature. University of Minnesota Press, 1987.

[17] Algirdas Julien Greimas. *Du Sens*. Editions du Seuil, Paris,, 1970.

[18] Algirdas Julien Greimas. *Structural Semantics: An Attempt at a Method*. University of Nebraska Press, Lincoln, 1983.

[19] Algirdas Julien Greimas and Jacques Fontanille. *The Semiotics of Passions: From States of Affairs to States of Feeling*. University of Minnesota Press, 1993.

[20] John Hartley, Indrek Ibrus, and Maarja Ojamaa. *On the Digital Semiosphere: Culture, Media and Science for the Anthropocene*. Bloomsbury Academic, New York, 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[22] Louis Hjelmslev and Francis J. Whitfield. *Prolegomena to a Theory of Language*. University of Wisconsin Press, 1961.

[23] Marie Katsurai and Shin'ichi Satoh. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2837–2841, 2016.

[24] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Ágata Lapedriza. Emotic: Emotions in context dataset. pages 2309–2317, 07 2017.

[25] G.R. Kress, T. Van Leeuwen, and T. van Leeuwen. *Reading Images: The Grammar of Visual Design*. Routledge, 1996.

[26] Peter J. Lang. International affective picture system (iaps) : Technical manual and affective ratings. 1995.

[27] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, 1999.

[28] 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021.

[29] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[30] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005.

[31] Upali Nanda, Xi Zhu, and Ben H. Jansen. Image and emotion: From outcomes to brain behavior. *HERD: Health Environments Research & Design Journal*, 5(4):40–59, 2012.

[32] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. A survey on visual sentiment analysis. *IET Image Process.*, 14:1440–1456, 2020.

[33] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. Survey on visual sentiment analysis. *IET Image Processing*, 14(8), 2020.

[34] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, 2015.

[35] Robert Plutchik and Henry Kellerman. *Theories of emotion*, volume 1. Academic press, 2013.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[37] Fernande Saint-Martin. *Semiotics of Visual Language*. Indiana University Press, Bloomington, IN, 1987.

[38] Klaus R. Scherer. Dimensional and categorical approaches to emotion: Theory and research. *Emotion Review*, 1(3):292–300, 2009.

[39] Iris Schneider, Lotte Veenstra, Frenk van Harreveld, Norbert Schwarz, and Sander Koole. Let's not be indifferent about neutrality: Neutral ratings in the iaps mask mixed affective responses. *Emotion*, 03 2016.

[40] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon S. Hare. Analyzing and predicting sentiment of images on the social web. *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[41] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317, 2017.

[42] Anna Wierzbicka. *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press, Cambridge, UK, 1999.

[43] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995. IEEE Computer Society, 2017.

[44] Jingyuan Yang, Qiruin Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20326–20337, 2023.

[45] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[46] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 381–388. AAAI Press, 2015.

[47] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI Conference on Artificial Intelligence*, 2016.

[48] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer. Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(1s), jan 2019.