Affection: Learning Affective Explanations for Real-World Visual Data

Panos Achlioptas[1,3] Maks Ovsjanikov[2] Leonidas Guibas[3] Sergey Tulyakov[1]

[1]Snap Inc. [2]LIX, Ecole Polytechnique, IP Paris [3]Stanford University

Abstract

In this work, we explore the space of emotional reactions induced by real-world images. For this, we first introduce a large-scale dataset that contains both categorical emotional reactions and free-form textual explanations for 85,007 publicly available images, analyzed by 6,283 annotators who were asked to indicate and explain how and why *they felt when observing a particular image, with a total of 526,749 responses. Although emotional reactions are subjective and sensitive to context (personal mood, social status, past experiences) – we show that there is significant common ground to capture emotional responses with a large support in the subject population. In light of this observation, we ask the following questions: i) Can we develop neural networks that provide plausible affective responses to real-world visual data explained with language? ii) Can we steer such methods towards producing explanations with varying degrees of pragmatic language, justifying different emotional reactions by grounding them in the visual stimulus? Finally, iii) How to evaluate the performance of such methods for this novel task? In this work, we take the first steps in addressing all of these questions, paving the way for more human-centric and emotionally-aware image analysis systems. Our code and data are publicly available at* `https://affective-explanations.org`.

1. Introduction

A central goal of computer vision has been to gain a semantic *understanding* of visual stimuli [17, 78]. But what exactly do we mean by this understanding? The vast majority of existing image analysis systems focus solely on image *content* [17]. Although models aimed at objective image analysis and captioning have achieved unprecedented success during the past years [70, 73], they largely ignore the more subtle and complex *interactions* that might exist between the image and its potential viewer.

In this work, our primary goal is to take a step toward a more viewer-centered understanding going *beyond* factual image analysis by incorporating the *effect* that an image might have on a viewer. To capture this effect, we argue that emotional responses provide a fundamental link between the visual world and human experience. We thus aim to understand what kinds of emotions a given image can elicit to different viewers and, most importantly, *why?*.

Emotion perception and recognition are influenced by and integrate many factors, from neurophysiological to cultural, from previous subjective experiences to social and even political context [41]. Thus, capturing and potentially reproducing plausible emotional responses to visual stimuli is significantly more challenging than standard image analysis, as it also involves an inherently *subjective* perspective, which is at the core of perception and consciousness [26].

To proceed with the goal of establishing a novel approach to affective analysis of real-world images, we leverage the fact that free-form language provides the simplest access to emotional expressions [60]. Thus, inspired by recent advances in affective captioning of art-works [7], we study emotional responses induced by real-world visual data in conjunction with human-provided *explanations*. This approach links emotions with linguistic constructs, which crucially are easier to curate *at scale* compared to other media (e.g., fMRI scans). Put together, our work expands on the recent effort of Achlioptas *et al.* [7] by considering a visio-linguistic and emotion analysis across a large set of *real-world images*, not only restricted to visual art.

Our main contributions to this end are two-fold: first, we curate a large-scale collection of 526,749 *explanations justifying emotions* experienced at the sight of 85,007 different real-world images selected from five public datasets. The collected explanations are given by 6,283 annotators spanning many different opinions, personalities, and tastes. The resulting dataset, which we term ***Affection***, is very rich in visual and linguistic variations, capturing a wide variety of both the underlying real-world depicted phenomena and their emotional effect. Second, we perform a linguistic and emotion-centric analysis of the dataset and, most importantly, use it to produce deep neural listeners and speakers trained to comprehend, or generate plausible *samples* of visually grounded explanations for emotional reactions to images. Despite the aforementioned subjectivity and thus the more challenging nature of these tasks compared to purely descriptive visio-linguistic tasks (e.g., COCO-based captioning [18]), our methods appear to learn *common* biases of how people react emotionally, e.g., the presence of a shark is much more likely to raise fear than the presence of a peacefully sleeping dog. Such *common sense* expectations are well captured in Affection, which is why we believe even black-box approaches like ours show promising results.

Finally, we explore variants of trained affective neural captioning systems, which allow some control on both the captured emotion and the level of factual visual details that are used when providing an explanation (e.g., 'The sky looks beautiful' to 'The blue colors of the sky and the sea in this sunset make me happy'). Interestingly, we demonstrate that the pragmatic variant demonstrates richer and more diverse language across different images.

In summary, this work introduces new task, termed *Affective Explanation Captioning* (AEC) for real-world images. To tackle AEC we release a new large-scale datased, *Affection*, capturing 526,749 emotional reactions and explanations. We then design a variety of components, including, neural speakers that enable affective captioning, with various degrees of pragmatic and emotional control over their generations. Finally, all our neural speakers show strong performance on emotional Turing tests, where humans find their humans find their generations ∼60%-65% of the time likely to be uttered by other humans supporting rich discriminative references contained in Affection's explanations.

2. Related Works

Emotion representation & learning. Two of the most widely adopted paradigms for representing emotions in existing literature are the discrete *categorical* system [30], and the *continuous* 2D-dimensional Valence-Arousal (VA) model [34, 67]. The former assumes a (typically small) predefined set of affective states, while the latter considers two fundamental dimensions: the emotional arousal (strength or intensity of emotion), and the emotional valence (degree of pleasantness or unpleasantness of the emotion) [35]. Following previous studies [7, 55, 64, 82, 86], we adopt the categorical system of emotion-representation, and use the same set of eight emotion categories: *anger*, *disgust*, *fear*, and *sadness* as negative emotions, and *amusement*, *awe*, *contentment*, and *excitement* as positive ones. In line with previous works ([7, 58, 64]), we treat *awe* as a positive emotion.. While we opt for the categorical system so as to stay closer to relevant existing works, there is still an active debate regarding the nature of emotions and their optimal representation [13, 35].

Image captioning. There is a rich literature on the topic of image captioning [22, 23, 53, 54, 77, 81] (see a review [73] for more detail). Most existing works concern neural models trained and tested with *descriptive* image-captions using well-established datasets [18, 44, 48, 56, 57, 63, 71]. Furthermore, relevant tasks such as VQA [8], or works that lift image captioning [2, 19], or reference disambiguation in 3D [3, 5, 20, 33, 42, 74], still focus on *descriptive* (object- or scene-centric) language. A notable exception is ArtEmis [7], which introduced a dataset and a series of tools for understanding and emulating the emotional effect of visual artworks. Similarly to that work, we focus on affective captions and develop neural speakers that aim to produce plausible textual utterances to capture the emotional effect of a given image. The key difference of our work compared to [7] is that we focus on *natural images*, not limited to art-works, making our contribution of much broader scope and utility. It is also worth mentioning a connection of our work to recent developments at the intersection of NLP and Causal Representation Learning [31, 68]. Namely, our 'captions' can be viewed as *causal* explanations of an underlying observed phenomenon, that of an emotional reaction.

Pragmatics & discriminative image analysis. CLIP [65] is a prominent discriminative visiolinguistic model that learns to assess the compatibility between a caption and an arbitrary image. In line with recent approaches, we use it alongside our neural speakers to explore the discriminative properties of Affection. In particular, a discriminator like CLIP can provide guidance for generative models such as our neural speakers, by calibrating and prioritizing their sampled productions (text) to increase the final caption-image compatibility [12, 66]. This process of re-ranking the captions to increase their relevance to the depicted image content, i.e., making them *pragmatic* [4, 11, 75] can be, as we show, particularly useful to increase diversity and control the level of visual details expressed in our neural generations. To the best of our knowledge, our work is the first to apply CLIP in this manner for pragmatic captioning. Nevertheless, very recent works appear to use CLIP directly as an auxiliary loss function while training descriptive neural captioning systems [21]. Finally, we mention the work of Bondielli and Passaro [9] who showed that CLIP can be used both in a zero-shot fashion, or by fine-tuning it for emotion-based image classification; and also the work of Wang *et al.* [79] who used CLIP in a zero shot manner to assess the quality and abstract perception of images.

3. Affection Dataset

The ***Affection*** (**Affect**tive Explanat**ion**s) dataset is built on top of images existing in the datasets MS-COCO [18], Emotional-Machines [47], Flickr30k Entities [62], Visual Genome [49], and the images in the image-to-emotion-classification work of Quanzeng *et al.* [64].

In total, we annotate 85,007 unique images corresponding to a curated subset of the 244,172 images contained in the above datasets. Namely, we use *all* images in [64], which have been specifically curated as to evoke emotions among the corresponding annotators of these works. We then use these images and a ResNet-based [36] visual embedding pretrained on ImageNet [27] to find their 3-Nearest-Neighbors in each of the remaining datasets (COCO, Visual-Genome, FlickR30k-Entities) to build Affection.

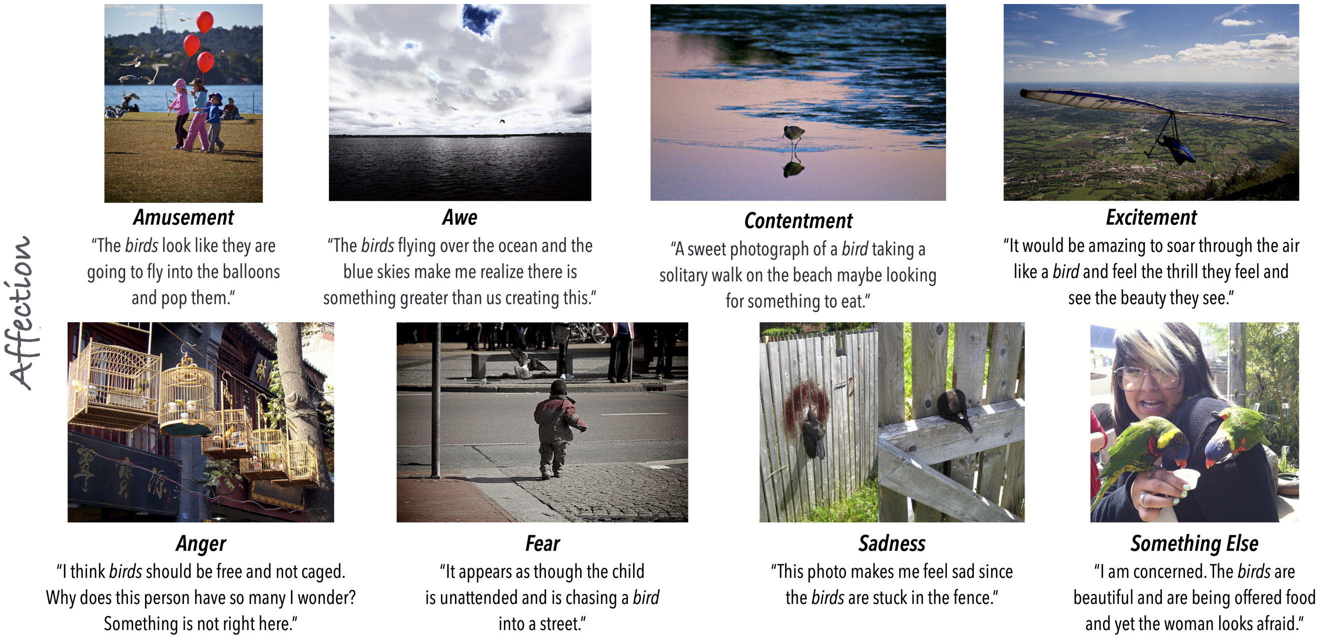

Affection
Amusement
"The *birds* look like they are going to fly into the balloons and pop them."
Awe
"The *birds* flying over the ocean and the blue skies make me realize there is something greater than us creating this."
Contentment
"A sweet photograph of a *bird* taking a solitary walk on the beach maybe looking for something to eat."
Excitement
"It would be amazing to soar through the air like a *bird* and feel the thrill they feel and see the beauty they see."
Anger
"I think *birds* should be free and not caged. Why does this person have so many I wonder? Something is not right here."
Fear
"It appears as though the child is unattended and is chasing a *bird* into a street."
Sadness
"This photo makes me feel sad since the *birds* are stuck in the fence."
Something Else
"I am concerned. The *birds* are beautiful and are being offered food and yet the woman looks afraid."

Figure 1. **Distinct emotional states and *typical* explanations in Affection related to the entity of "bird."** The explanations capture a wide range of abstract semantics and nuanced associations between entities and the underlying explained emotion (shown in boldface). Note, the semantics include common sense reasoning and cognitive-level understanding of an image, going *beyond* recognizing its visible elements.

Per each image we ask at least 6 annotators to express their emotional reaction and explanation. In this we follow Achlioptas *et al.* [7]. Upon observing an image an annotator is asked first to indicate their dominant emotional reaction by selecting among the eight emotions mentioned in Section 2, or a ninth option, listed as 'something-else'. This option allows one to indicate finer grained emotions not explicitly listed in our UI, or to explain why they might not have any strong emotional reaction to the specific image (in total the annotators used this latter option 7.6% of the time). Then, each annotator is asked to provide a textual explanation for their choice. The explanation should include at least one specific reference to visual elements depicted in the image. See Figure 1 for examples of typical collected annotations.

Our resulting corpus consists of 526,749 emotion indications and corresponding explanations, with the latter using a vocabulary of 41,275 distinct tokens. We note that the 6,283 annotators that worked to built Affection were recruited with Amazon's Mechanical Turk (AMT) services. For more details see our online Supplemental Materials [6].

3.1. Language-oriented Analysis

Richness & diversity. The average length of Affection's explanations is 18.8 words. This is noticeably longer than the average length of utterances of ArtEmis and significantly longer than the captions of many other well-established and descriptive captioning datasets, as shown in Table 1. Moreover, we use NLTK's part-of-speech tagger [14], to analyze Affection's explanations in terms of their average number of contained nouns, pronouns, adjectives, verbs, and adpositions. Across *all* these lexical categories, Affection contains a higher occurrence per caption, implying the use of a rich and complex vocabulary by its annotators. In the Supp. Mat. we also provide statistics of lexical distributions explanations for the same image, which further highlight the diversity and richness of Affection.

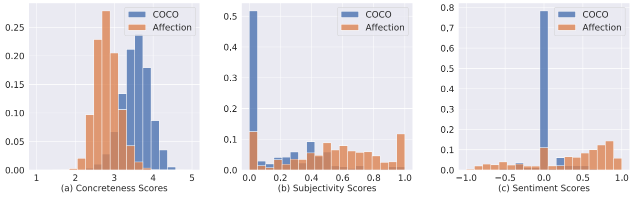


Figure 2. **Characteristic properties of Affection**. Empirical distributions contrasting Affection's explanations to the descriptive captions of COCO [18] along the axes of (a) *Concreteness*, (b) *Subjectivity*, and (c) *Sentiment*.

Dataset	Words	Nouns	Pronouns	Adjectives	Adpositions	Verbs
Affection	**18.8**	**4.5**	**1.3**	**1.8**	**2.2**	**4.0**
ArtEmis [7]	15.9	4.0	0.9	1.6	1.9	3.0
Flickr30k Ent. [84]	12.3	4.2	0.2	1.1	1.9	1.8
COCO [18]	10.5	3.7	0.1	0.8	1.7	1.2
Conceptual Capt. [71]	9.6	3.8	0.2	0.9	1.6	1.1
Google Refexp [56]	8.4	3.0	0.1	1.0	1.2	0.8

Table 1. **Lexical comparison over distinct part-of-speech categories, per *individual captions***. The average occurrences for the shown categories are significantly higher in Affection, indicating that it is comprised of a lexically **richer** and more complex corpus.

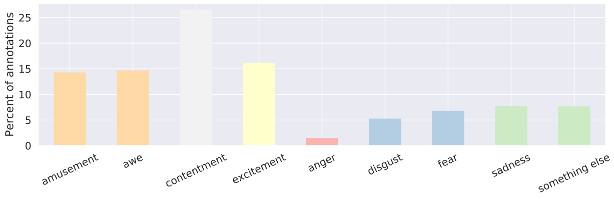


Figure 3. **Empirical distribution of indicated dominant emotion categories accompanying Affection's explanations**. The left-most four bars include *positive* emotions, which are selected 71.3% of the time. *Negative* emotions (5th-8th bars) appear much less frequently (21.1%), while the 'something-else' is preferred 7.6% of the time.

Abstractness & subjectivity. We also measure the degree of abstractness (vs. concreteness) of our corpus, by using the lexicon of Brysbaert *et al.* [16] which provides for thousand word lemmas a scalar value (from 1 to 5) to indicate their concreteness. As an example, the words *bird* and *cage* represent fully concrete (tangible) entities, getting a score of 5, but the words *freedom* and *carefree* are considered more abstract concepts (with scores 2.34 and 1.88, resp.). On average, a uniformly random word of Affection scores 2.82 in concreteness while a random word of COCO does 3.55 (see also Figure 2 (a)). In other words, the annotators of Affection make use of significantly more abstract concepts than in COCO. We also evaluate the degree of subjectivity in Affection's annotations. We use the subjectivity metric provided by TextBlob [52]. As seen in the empirical distribution comparing our explanations to COCO's captions in terms of their subjectivity (Figure 2 (b)), Affection contains significantly more subjective references.

Sentiment analysis. Perhaps as expected by its nature, Affection also contains language that is highly sentimental. To measure this aspect, we use a rule-based sentiment classifier (VADER [40]). VADER classifies 10.5% of Affection's explanations to the neutral sentiment, while it finds a significantly higher fraction (77.4%) of the descriptive COCO-captions to fall in this category. In Figure 2 (c) we present the histogram of VADER's estimated valences for the utterances contained in these two datasets. Valences with small magnitude (closer to 0) indicate a neutral sentiment, while those closer to the extremes (-1, 1) indicate highly negative or positive sentiments.

Comparing Affection to ArtEmis. The two previous analyses contrast our dataset with the descriptive captions of COCO. Another dataset that is more similar in nature with Affection is ArtEmis [7]. We note that Affection's language is similar in terms of abstractness to ArtEmis (average scores of 2.82 vs. 2.81). At the same time, Affection's language is more sentimental (VADER's classifier assigns 10.5% vs. 16.5% of each corpus to the neutral category), and more subjective (average subjectivity scores are 0.53 vs. 0.47, respectively).

3.2. Emotion-oriented Analysis.

The annotators of Affection indicated a variety of dominant emotional reactions upon observing different visual stimuli. As seen in Figure 3 positive emotions (see Section 2 for the used convention) were ∼3.4 more likely to occur than negative ones (71.3% vs. 21.1%). Also, the "something-else" category was preferred 7.6% of the time, and it included a large variety of subtler emotional reactions (e.g., curiosity, nostalgia, etc.). Despite the prevalence of positive emotions overall, we note that crucially 50.0% of images were annotated with at *least one positive and one negative emotion*.

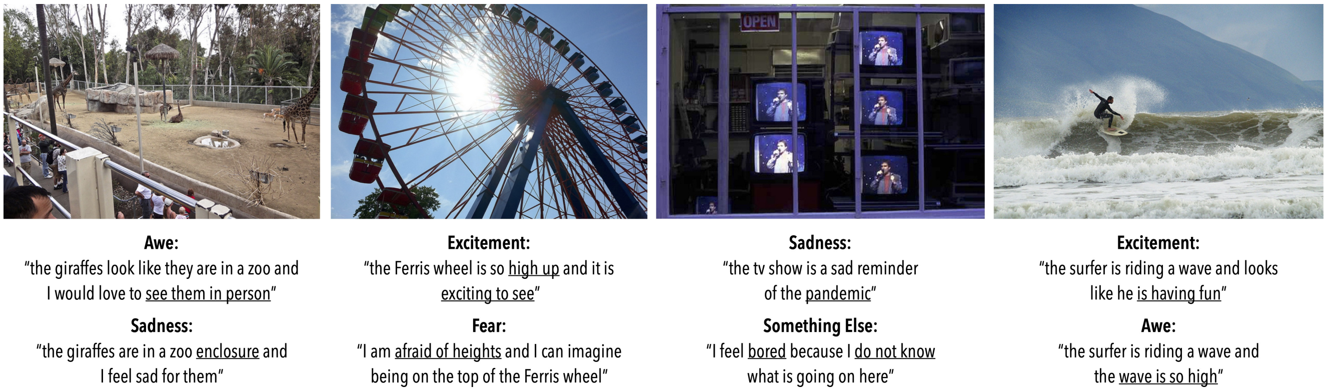
While this result highlights the high degree of subjectivity inherent to our task, 67.5% (57,381) of the annotated images have a strong majority among their annotators who indicated the same fine-grained emotion. First, it is worth noting that this fact establishes Affection as one of the *largest* publicly available image-to-emotion classification datasets, by merely concentrating on these 57,381 images and associating them only with their underlying majority label (i.e., following a similar strategy as the one used in the work of Quanzeng *et al.* [64].). Second, we compare our annotators' agreement with that of ArtEmis. Despite the fact that both datasets use the exact same protocol for annotation purposes, ArtEmis has much *less* agreement among its annotators: specifically, only 45.6% of its annotated artworks attain a strong majority agreement. We hypothesize that this salient (> 20%) difference is related to visual art being in general more ambiguous and thus evoking more polarizing emotions, compared to the more familiar scenarios depicted by real-world images.

4. Affective Tasks in Computer Vision

The properties of Affection allow us to explore a series of computer vision tasks. First, we describe two emotion-oriented classification tasks, necessary for our later experiments. Section 4.2 describes neural listening and speaking tasks, whose primary goal is to evaluate or produce plausible emotion explanations given a visual input. While our methods attain promising results, we emphasize that the value of Affection goes beyond the specific design of the neural listening/captioning systems that we evaluate. We invite the community to build upon our observations and results.

4.1. Basic Classification Tasks

We explore two simple *classification* task. In the first, we predict the emotion class (among the 9 possibilities) of the given *textual utterance* contained in Affection. Similarly to [7], we formulate this problem as a standard 9-way text classification problem. We use an LSTM [39]-based text classifier trained from scratch using the standard cross-entropy loss. We also fine-tune a transformer-based BERT model [29]. In the second problem, we predict the distribution of emotional reactions for a given *image*. To this



Awe:
"the giraffes look like they are in a zoo and I would love to see them in person"

Sadness:
"the giraffes are in a zoo enclosure and I feel sad for them"

Excitement:
"the Ferris wheel is so high up and it is exciting to see"

Fear:
"I am afraid of heights and I can imagine being on the top of the Ferris wheel"

Sadness:
"the tv show is a sad reminder of the pandemic"

Something Else:
"I feel bored because I do not know what is going on here"

Excitement:
"the surfer is riding a wave and looks like he is having fun"

Awe:
"the surfer is riding a wave and the wave is so high"

Figure 4. **Examples of neural speaker generations with the default *emotion-grounded* speaker variant on *unseen* test images** The grounding emotion (shown in boldface) is predicted during inference time by a separately trained image-to-emotion classifier. We ground the speaker's generation with two emotions for each image, corresponding to the most likely (top row) and second most likely (bottom row) predictions. As show in the figure, this variant provides a certain control over the output by aligning it to the requested/input emotion.

end, we fine-tune a ResNet-101 encoder [37] pretrained on ImageNet [27]. As we show below, these two classifiers (i.e., text-to-emotion and image-to-emotion), denoted as $C_{emotion|text}$ and $C_{emotion|image}$ allow us to both evaluate, as well as to control, the emotional content of the trained neural speakers (Sections 5 and 4.2). We note, however, that the two problems mentioned above have independent interest and we explore them in Section 6.

4.2. Neural Listeners and Speakers

Neural comprehension with affective explanations. To test the degree to which the explanations of Affection can be used to identify their underlying described image against random 'distracting' images, we deploy two neural listeners [32, 45]. First, we use Affection to train jointly and from scratch, a transformer-based language encoder and a ResNet-based visual encoder under a self-contrastive criterion. Namely, during training, given a random batch of encoded image-caption pairs we optimize a cross-entropy loss that aligns the two modalities in a joint visual/language embedding space [25, 43, 65]. Second, we deploy a pretrained CLIP model [65], *without* fine-tuning with Affection. Using a non-finetuned version of CLIP allows us to test this popular network *as is* on our data, and to compare its performance against other datasets (e.g., COCO). During inference we input to the listeners a test image along with a set of randomly sampled distracting test images. We also provide its ground-truth explanation. Finally, we output the image with maximal alignment (expressed in logits) to the provided explanation.

Default speaker backbones. We evaluate two backbone architectures throughout our neural-speaker studies, including the widely used Show-Attend-and-Tell (SAT) [81] and the recent transformer-based SoTA, GRIT [59]; further details and more speaker model studies are provided in the supplementary materials. In the cases we train a *default* speaker module with either backbone we simply use Affection captions as ground truth annotations, without considering the emotion category labels provided by the annotators.

Emotion grounded speaker. Following ArtEmis [7], we tested speaking variants that also incorporate an additional *emotion* argument. For these variants, during training, in addition to the image, we input to the speaker an MLP-encoded vector representing the *emotion* that the ground-truth explanation justifies. During inference, we replace the ground truth emotion with the most likely predicted emotion by the $C_{emotion|image}$ network described above. Interestingly, we observe that this variant also gives some control over the subjectivity of the response. E.g., are 'risky' activities such as riding a Ferris wheel *preferred*, or are they to be avoided? See Figure 4 for a demonstration of this effect.

Pragmatic variants. As we show in Section 6, the above speakers can generate plausible explanations for a variety of emotions for the underlying visual stimulus. To test the degree to which we can control their ability to include discriminative details of the underlying image, we experiment with their *pragmatic* versions. Specifically, inspired use [32], we augment our neural speakers with the capacity to prioritize sampled explanations, judged by a separately trained 'internal' listener as discriminative (we use a pretrained CLIP in our experiments). In this case, we sample explanations from our speakers but score (i.e., re-rank) them according to:

$$\beta \log(P_L(i,u)) + (1-\beta)\log(P_S(u|i)), \tag{1}$$

where P_L is the listener's probability to associate the image (i) with a given output utterance (u), and P_S is the likelihood of the non-pragmatic speaker version to generate u. The parameter β controls the relative importance of two terms. To make the two terms comparable, we re-scale the probabilities so that on average the two terms have the same magnitude.

*Generations with the **default** neural speaker:*

"the dog is sleeping on the bed and looks very comfortable"

"the house looks like a very peaceful place to be"

"the bird is looking at the camera and it makes me laugh"

"the people are waiting for the bus to arrive"

***Pragmatic** counterparts:*

"this sleeping dog is so cute and adorable it **makes me want to take a nap**"

"the flowers are so bright and colorful and the house looks like a **nice place to live**"

"the bird is standing in the **sand** and it is **cute and funny**"

"this image **makes me feel claustrophobic** because **I do not like crowds** and I would be afraid to be in a crowded airport"

Figure 5. **Effect of boosting the pragmatic content of neural speaker generations via CLIP.** Aside from often correcting the identity of shown objects/actions (right-most image is indeed taken inside an airport), the pragmatic variant tends to use more visual details in its explanations (e.g., *'standing in the sand'*), and perhaps more importantly to expand the explanation to include non-visual but valid associations (e.g., *'take a nap'*, or *'do not like crowds'*).

5. Evaluation

Evaluating a captioning systems is a challenging problem [38, 80, 83]. This is both because the space of possible captions is very large and because the model might suffer from mode collapse [69, 72, 80], producing repetitive and overly simple captions across different images. Both of these problems are exacerbated in Affective Explanation Captioning (AEC) due to its more open-ended and subjective nature. In this work, we highlight these challenges in the context of AEC and explore the efficacy and limitations of a large variety of established metrics in relation to different neural speaking variants, providing several key insights.

Comparing with ground-truth. To evaluate the quality of the output of our neural speakers with respect to the hidden ground-truth annotations of the test images, we first use some of the most established automatic metrics for this purpose: BLEU 1-4 [61], ROUGE-L [50], METEOR [28] and SPICE [10]. These n-gram similarity based metrics, (or semantic-scene-graph-based for SPICE); expect at least *one* of the ground-truth captions to be similar to the corresponding generation. Note that we do not use CIDEr [76], because it requires the output generation to be similar to *all* held-out utterances of an image, which by the nature of Affection is not a well-justified requirement (see [7] for details).

In addition to the above metrics for comparisons with existing captioning approaches and datasets [73], we use the recently proposed CLIPScore and RefClipScore [38], as they have shown improved correlation with human judgement [38]. Specifically, RefClipScore assumes access to ground-truth human annotations which it then compares with the generated caption based on CLIP's association scores, while CLIPScore directly uses CLIP's caption-image compatibility to compare different speakers, making it a ground-truth free metric.

Assessing diversity of productions. To evaluate the susceptibility of different neural speakers to suffer from mode collapse, we consider three metrics. First, we report the average of the *maximum length of the Longest Common Subsequence* (LCS) of the generated captions and (a subsampled version) of all *training* explanations. The smaller the LCS is, the less similar the evaluated captions are from the training data (i.e., less over-fitting occurs). Secondly, we also report for all our neural variants, the percent of *unique* captions they generated across the same set of test images. Last, inspired by the diversity-metric-oriented work of Qingzhong and Antoni [80] and the above-described CLIP's application on caption evaluation [38], we also introduce a new metric, which uses CLIP to detect the lack of caption diversity, and which we dub CLIP-Diversity-Cosine (ClipDivCos)$\in [-1, 1]$. For this metric, we use a pretrained CLIP model to encode *all* generated captions across a set of test images, and report the average pairwise *cosine* of the angles of the embedded vectors. Note that CLIP's textual (and visual) embeddings are optimized to be semantically similar when their angles' cosine is large (+1). Thus, *the smaller* the CLIPDivCos of a collection of vectors is, the more semantically heterogeneous and diverse this collection is expected to be.

Specializing to affective explanations. The last axis of evaluation relates to *affective explanations*, and for which relevant metrics where introduced in ArtEmis [7]. Concretely, first, we estimate the fraction of a speaker's productions that contain *metaphors and similes*. We do this by tagging generations that include a small set of manually curated phrases. The estimated fraction of *ground-truth* Affection explanations that have such metaphorical-like content is 19.7% – setting the 'ideal' expected percentage for our neural speakers. Secondly, we use the *emotional-alignment* metric introduced in ArtEmis [7]. For this, we use the trained

$C_{emotion|text}$ classifier to predict the most likely emotion for each generated utterance. Namely, for test images where the human-indicated emotions formed a strong majority among their annotators, we report the percent of their corresponding neural based captions where the $\arg\max(C_{emotion|caption})$ is equal to the ground truth majority-chosen emotion.

Human-based evaluation. All previously defined metrics can be automatically computed at scale, and aim to be a proxy for the more expensive and precise human-based quality assessment [24, 38, 46]. As a final quality check, we deploy user-based studies that emulate an *emotional Turing test* [7], assessing how likely it is for third-party human observers to decide that the synthetic captions were made by other humans, instead of being produced by neural speakers.

6. Experimental Results

For the experiments described in this section we train neural networks by using an 85%-5%-10% train/val/test split of Affection, making sure that the splits have no overlap in terms of their underlying images.

Preliminary experiments: emotion classification from text or images. As shown by previous studies [7] predicting the *fine-grained* emotion supported by an affective explanation is much harder than binary or ternary text-based sentiment classification [15, 85]. By using the neural-based text predictors described in Section 4.1 we found that an LSTM-based classifier attains 69.8% average accuracy on the same test split used for our neural-speakers (52,188 explanations). A BERT-based classifier achieved an improved accuracy of 72.5% when fine-tuned on this task. Interestingly, these two models, when trained with ArtEmis [7] generalized more poorly (63.3% and 64.8%, respectively) – suggesting that the explanations of Affection are more indicative of the the emotion they support, compared to those of ArtEmis. Importantly, these classifiers failed *gracefully* and were mostly confused among subclasses of the same, positive or negative emotion, achieving 94.0%, 95.5% accuracy in this binary classification task, respectively.

We evaluated the *image*-to-emotion classification problem described in Section 4.1 on the subset of 5,672 test images for which our annotators indicated a *unique strong majority* emotion. On this set, a fine-tuned ResNet-101 encoder predicts the fine-grained emotion correctly 59.1% of the time. Notice that the emotion label distribution of Affection is highly imbalanced, e.g., *anger* attains a strong majority in (0.35%) cases, compared to *contentment* (34.9%). This fact makes our particular prediction problem harder than usual but also open to possibly more specialized solutions such as using a focal loss [1, 51]. Last, if we convert the ResNet predictions to binary sentiments, the ResNet predicts the ground-truth sentiment 88.5% of the time correctly, also indicating a graceful failure mode.

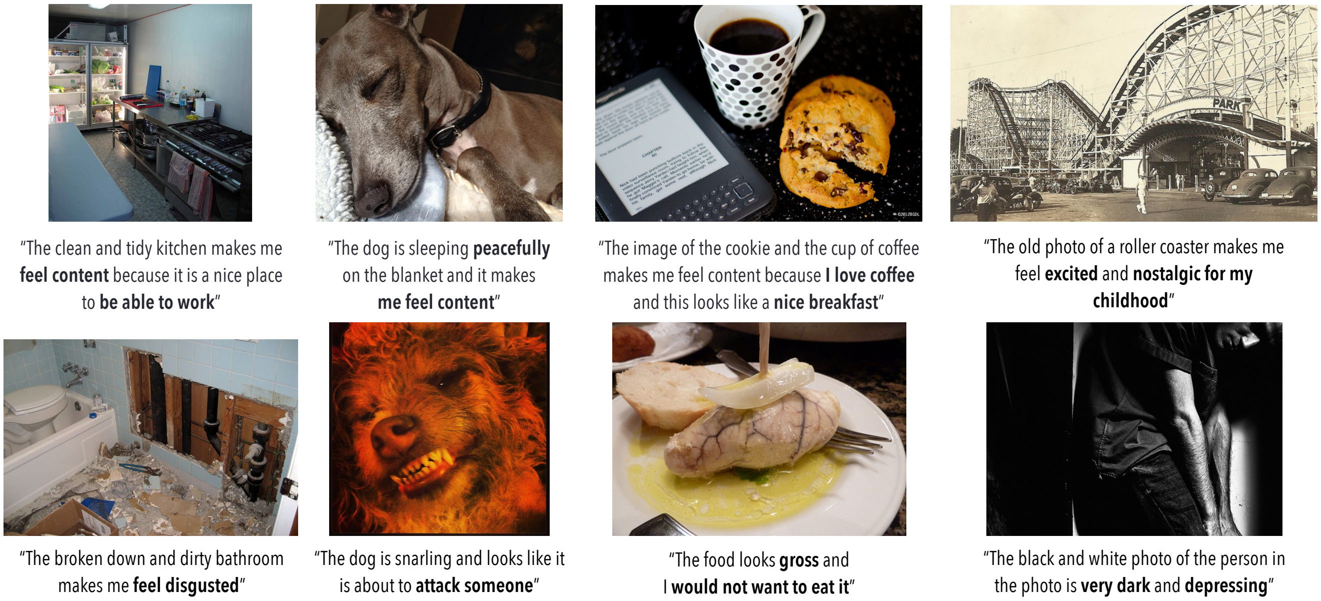
Neural Comprehension of Affective Explanations. Next, we explore the extent to which the textual explanations in Affection refer to discriminative visual elements of their underlying images, to enable their *identification* among arbitrary images, with the help of two 'neural listeners' described in Section 4.2. Specifically, we use a pretrained CLIP model with 400M parameters (version `ViT-B/32`) and couple all ground-truth image-caption/explanation pairs of a dataset with a varying number of randomly chosen images from the same dataset. We then retrieve for each given caption the image with the largest (cosine-based) similarity. We note that even with as many as *ten distracting images* the retrieval of CLIP accuracy remains strong, at 89.7% accuracy on Affection (compared to 96.5% on COCO). This suggests that Affection's explanations contain significant amounts of 'objective' and discriminative grounding details to enable excellent identification of an image from its underlying explanation. Additional results on this experiment are provided in the Supp. Mat. [6].

Neural-based <u>A</u>ffective <u>E</u>xplanation <u>C</u>aptioning (AEC). Finally, we deploy the speakers described in Section 4.2 to give the first neural-based solution to our core problem of AEC. We report here results with the SAT backbone (denoted as 'default'), as well as its emotionally-grounded and pragmatic variants mentioned above. Similar, findings with the GRIT-based baselines are also reported in the Supp.

Table 2 reports machine-based evaluation metrics and provides several important insights for each variant. Namely, in this table, we observe that the standard n-gram-based metrics of the first group: BLEU-1-4,..., and SPICE, are *slightly* improved if we use the default or pragmatic models, which do not explicitly use emotion for grounding. Given that the held-out explanations typically justify a large variety of emotions for each image, biasing the generation with a single specific (guessed) emotion, as the emotion-grounded variants do, might be too restrictive. On the other hand, for the subset of test images with a ground-truth strong-emotional majority, where we evaluate the emotional-alignment score; we see a noticeable improvement when using the emotion-grounded variants. Interestingly, these variants also fare better regarding the number of similes they produce by better approaching Affection's ground-truth average of 19.7%.

Regarding RefClipScore and ClipScore, the pragmatic variants *significantly* outperform their non-pragmatic counterparts. Given that we use CLIP to re-rank and select their generations, this result might be somewhat expected. However, as we will show next (emotional Turing test), these variants also fare better in our human-based evaluation. Also, equally important, for *all* diversity-oriented metrics (third group of metrics in Table 2), the pragmatic variants fare best. For a qualitative demonstration of pragmatic inference's effect see Figure 5. We show curated generations from the (emotion-grounded) pragmatic variant in Figure 6.

Emotional Turing test. We also evaluate how likely our neural speaking variants' output generations can be per-



"The clean and tidy kitchen makes me **feel content** because it is a nice place to **be able to work**"

"The dog is sleeping **peacefully** on the blanket and it makes **me feel content**"

"The image of the cookie and the cup of coffee makes me feel content because **I love coffee** and this looks like a **nice breakfast**"

"The old photo of a roller coaster makes me feel **excited** and **nostalgic for my childhood**"

"The broken down and dirty bathroom makes me **feel disgusted**"

"The dog is snarling and looks like it is about to **attack someone**"

"The food looks **gross** and **I would not want to eat it**"

"The black and white photo of the person in the photo is **very dark** and **depressing**"

Figure 6. **Curated examples of neural speaker generations on unseen images from the emotion-grounded, pragmatic speaker variant.** The top row includes generations that reflect a positive sentiment, while the bottom row showcases generations grounded on similar visual subjects (object classes) e.g., another dog, food item, etc., that give rise to negative emotions. Remarkably, this neural speaker appears to take into account the underlying fine-grained visual differences to properly modulate its output, providing strong **explanatory power** behind the emotional reactions. Note, also, how the explanations can include purely human-centric semantics (*'nostalgic of my childhood'*, *'love coffee'*), and use explicit **psychological** assessments (*'feel content/excited/disgusted'*, *'is depressing'*).

	Speaker Variants			
Metrics	Default	Emo-Grounded	Default (Pragmatic)	Emo-Grounded (Pragmatic)
BLEU-1,2 (↑)	**64.4**, **38.3**	63.1, 36.9	64.3, 38.0	63.4, 37.0
BLEU-3,4 (↑)	**22.2**, **13.2**	20.9, 12.0	21.8, 12.8	20.9, 11.9
METEOR (↑)	14.9	14.4	**15.1**	14.8
ROUGE-L (↑)	30.8	30.5	**31.0**	30.8
SPICE (↑)	7.4	7.2	**8.0**	7.7
CLIPScore (↑)	66.7	66.8	**69.2**	**69.2**
RefCLIPScore (↑)	75.0	75.0	**76.3**	**76.3**
Unique-Productions (↑)	78.7	80.7	82.9	**83.7**
Max-LCS (↓)	70.4	70.4	68.6	**68.4**
ClipDivCos (↓)	73.1	72.8	**69.8**	70.2
Similes (↓)	42.8	36.3	40.0	**34.5**
Emo-Alignment (↑)	48.1	55.2	48.2	**55.9**

Table 2. **Neural speaker machine-based evaluations**. The Default models use for grounding only the underlying image, while the Emo-Grounded variants also input an emotion-label. Pragmatic variants use CLIP to calibrate the score of sampled productions before selecting the final proposal.

ceived as if they were made by humans. For this, we form a random sample of 500 *test* images and accompany each image with both of their human-made explanations and a generation made by a neural speaker. We do this by considering all four speaking variants to obtain 2,000 image-caption samples. We then ask AMT annotators who have never seen these sampled images to select one among four options: (a) *both* explanations seem to have been made by humans justifying their emotional reaction to the shown image; (b) *none* of the explanations are likely to have been made by humans for that purpose, or (c) (and (d)) to select the explanation that seems more likely to have been made by a human. We observe that for all neural speakers, in more than 40% of the cases, both utterances were thought of as human-made. Moreover in a significant fraction of the answers (Default variant: 15.6%, Emo-grounded: 18.2%, Default-Pragmatic: 19.7%, Emo-Grounded Pragmatic: 19.0%), the neural-based generations were deemed more likely than the human-made ones. These results highlight both the complexity of the AEC problem as well as the promising overall quality of our neural speaker solutions, enabled by the Affection dataset.

7. Conclusion and Future Vision

Humans react to stimuli beyond the literal content of an image as they respond to the story behind it. More broadly the image evokes emotions relating to human experience in general and the viewer's experience in particular. In this work we have shown how linguistic explanations of affective responses can express and illuminate this larger context, leading to a more comprehensive understanding of how image content and its elements affect human emotion. Using our new Affection dataset, we have demonstrated that neural speakers can produce generations that mimic well these human responses. In addition to highlighting the importance and utility of exploring the affective dimension of image understanding, we believe that this study can stimulate work in many interesting novel directions.

Acknowledgements. This work was supported by a Vannevar Bush Faculty Fellowship, the ERC Starting Grant No. 758800 (EXPROTEA) and the ANR AI Chair AIGRETTE.

References

[1] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Long tail visual relationship recognition with hubless regularized relmix. In *International Conference on Computer Vision (ICCV)*, 2021. 7

[2] Ahmed Abdelreheem, Kyle Olszewski, Hsin-Ying Lee, Peter Wonka, and Panos Achlioptas. ScanEnts3D: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes. *Computing Research Repository (CoRR)*, abs/2212.06250, 2022. 2

[3] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[4] Panos Achlioptas, Judy Fan, Robert XD Hawkins, Noah D Goodman, and Leonidas J. Guibas. ShapeGlot: Learning language for shape differentiation. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[5] Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. ChangeIt3D: Language-assisted 3d shape edits and deformations. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[6] Panos Achlioptas, Maks Ovsjanikov, Leonidas Guibas, and Sergey Tulyakov. *Affection: Learning Affective Explanations for Real World Visual Data (Supplementary Material)*, (accessed October 2022). Available at `https://affective-explanations.org/materials/affection_supp_mat.pdf`. 3, 7

[7] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. ArtEmis: Affective language for visual art. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 4, 5, 6, 7

[8] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. *Computing Research Repository (CoRR)*, abs/1505.00468, 2015. 2

[9] Bondielli Alessandro and Passaro Lucia C. Leveraging CLIP for image emotion recognition. In *NL4AI*, 2021. 2

[10] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. *European Conference on Computer Vision (ECCV)*, 2016. 6

[11] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2

[12] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with latent language. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2018. 2

[13] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017. 2

[14] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 3

[15] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 2021. 7

[16] Marc Brysbaert, Amy Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 2014. 4

[17] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 2021. 1

[18] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and Lawrence C. Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *Computing Research Repository (CoRR)*, abs/1504.00325, 2015. 2, 3

[19] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in rgb-d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[20] Z. Dave Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. *Computing Research Repository (CoRR)*, abs/1912.08830, 2019. 2

[21] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In *Findings of NAACL*, 2022. 2

[22] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[23] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[25] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[26] Stanislas Dehaene. *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin, 2014. 1

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 5

[28] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. 6

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository (CoRR)*, abs/1810.04805, 2018. 4

[30] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 1992. 2

[31] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Computing Research Repository (CoRR)*, abs/2109.00725, 2021. 2

[32] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 2016. 5

[33] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3D: A minimally contrastive benchmark for grounding spatial relations in 3D. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[34] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2011. 2

[35] Stephan Hamann. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, 2012. 2

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computing Research Repository (CoRR)*, abs/1512.03385, 2015. 2

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[38] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 6, 7

[39] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 4

[40] C.J. Hutto and Eric E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. eighth international conference on weblogs and social media. *ICWSM*, 2014. 4

[41] Assia Jaillard and Thomas Zeffiro. *Phylogeny of Neurological Disorders/Anatomy and Disorders of Basic Emotion in Stroke*, chapter 29. Elsevier, 2020. 1

[42] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[43] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021. 5

[44] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and L. Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2

[45] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 5

[46] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. *CoRR*, abs/1612.07600, 2016. 7

[47] H. Kim, Y. Kim, S. J. Kim, and I. Lee. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 2018. 2

[48] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, and et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 2017. 2

[49] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 2017. 2

[50] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004. 6

[51] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision (ICCV)*, 2017. 7

[52] Steven Loria. *TextBlob*, (accessed September 13, 2022). Available at `https://textblob.readthedocs.io/en/dev/`. 4

[53] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[54] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning. *Computing Research Repository (CoRR)*, abs/2201.12723, 2022. 2

[55] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, 2010. 2

[56] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Murphy Kevin. Generation and comprehension of unambiguous object descriptions. *Computing Research Repository (CoRR)*, abs/1511.02283, 2016. 2, 3

[57] Willi Menapace, Aliaksandr Siarohin, Stéphane Lathuilière, Panos Achlioptas, Vladislav Golyanik, Elisa Ricci, and Sergey Tulyakov. Plotting behind the scenes: Towards learnable game engines. *Computing Research Repository (CoRR)*, abs/2303.13472, 2023. 2

[58] 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*, 2005. 2

[59] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. GRIT: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[60] Andrew Ortony, Gerald L. Clore, and Collins Allan. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988. 1

[61] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002. 6

[62] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[63] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[64] You Quanzeng, Luo Jiebo, Jin Hailin, and Yang Jianchao. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *Computing Research Repository (CoRR)*, abs/1605.02677, 2016. 2, 4

[65] 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. *Computing Research Repository (CoRR)*, abs/2103.00020, 2021. 2, 5

[66] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *Computing Research Repository (CoRR)*, abs/2102.12092, 2021. 2

[67] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 1980. 2

[68] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *Computing Research Repository (CoRR)*, abs/2102.11107, 2021. 2

[69] Paul Hongsuck Seo, Piyush Sharma, Tomer Levinboim, Bohyung Han, and Radu Soricut. Reinforcing an image caption generator using off-line human feedback. In *AAAI Conference on Artificial Intelligence*, 2020. 6

[70] Himanshu Sharma. A survey on image captioning datasets and evaluation metrics. *IOP Conference Series: Materials Science and Engineering*, 2021. 1

[71] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 2, 3

[72] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *International Conference on Computer Vision (ICCV)*, 2017. 6

[73] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 1, 2, 6

[74] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. *Computing Research Repository (CoRR)*, abs/2107.12514, 2021. 2

[75] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[76] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[77] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *Computing Research Repository (CoRR)*, abs/1411.4555, 2015. 2

[78] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, and Diego Andina. Deep learning for computer vision: A brief review. *Intell. Neuroscience*, 2018. 1

[79] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. *arXiv preprint arXiv:2207.12396*, 2022. 2

[80] Qingzhong Wang and Antoni Chan. Describing like humans: on diversity in image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[81] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015. 2, 5

[82] Victoria Yanulevskaya, Jan Gemert, Katharina Roth, Ann-Katrin Schild, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In *IEEE International Conference on Image Processing*, 2008. 2

[83] Yanzhi Yi, Hangyu Deng, and Jinglu Hu. Improving image captioning evaluation by considering inter references variance. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 6

[84] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2014. 3

[85] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018. 7

[86] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM International Conference on Multimedia*, 2014. 2