ArtEmis: Affective Language for Visual Art

Panos Achlioptas[1] panos@cs.stanford.edu
Maks Ovsjanikov[2] maks@lix.polytechnique.fr
Kilichbek Haydarov[3] kilichbek.haydarov@kaust.edu.sa
Mohamed Elhoseiny[3,1] mohamed.elhoseiny@kaust.edu.sa
Leonidas Guibas[1] guibas@cs.stanford.edu

[1]Stanford University
[2]LIX, Ecole Polytechnique, IP Paris
[3]King Abdullah University of Science and Technology (KAUST)

Abstract

We present a novel large-scale dataset and accompanying machine learning models aimed at providing a detailed understanding of the interplay between visual content, its emotional effect, and explanations for the latter in language. In contrast to most existing annotation datasets in computer vision, we focus on the affective experience triggered by visual artworks and ask the annotators to indicate the dominant emotion they feel for a given image and, crucially, to also provide a grounded verbal explanation for their emotion choice. As we demonstrate below, this leads to a rich set of signals for both the objective content and the affective impact of an image, creating associations with abstract concepts (e.g., "freedom" or "love"), or references that go beyond what is directly visible, including visual similes and metaphors, or subjective references to personal experiences. We focus on visual art (e.g., paintings, artistic photographs) as it is a prime example of imagery created to elicit emotional responses from its viewers. Our dataset, termed ArtEmis, contains 455K emotion attributions and explanations from humans, on 80K artworks from WikiArt. Building on this data, we train and demonstrate a series of captioning systems capable of expressing and explaining emotions from visual stimuli. Remarkably, the captions produced by these systems often succeed in reflecting the semantic and abstract content of the image, going well beyond systems trained on existing datasets. The collected dataset and developed methods are available at https://artemisdataset.org.

1. Introduction

Emotions are among the most pervasive aspects of human experience. While emotions are not themselves linguistic constructs, the most robust and permanent access we have to them is through language [44]. In this work, we focus on collecting and analyzing at scale language that *explains* emotions generated by observing visual artworks. Specifically, we seek to better understand the link between the visual properties of an artwork, the possibly subjective affective experience that it produces, and the way such emotions are explained via language. Building on this data and recent machine learning approaches, we also design and test neural-based speakers that aim to emulate human emotional responses to visual art and provide associated explanations.

Why visual art? We focus on visual artworks for two reasons. First and foremost because art is often created with the intent of provoking emotional reactions from its viewers. In the words of Leo Tolstoy, *"art is a human activity consisting in that one human consciously hands on to others feelings they have lived through, and that other people are infected by these feelings, and also experience them"* [55]. Second, artworks, and abstract forms of art in particular, often defy simple explanations and might not have a single, easily-identifiable subject or label. Therefore, an affective response may require a more detailed analysis integrating the image content as well as its effect on the viewer. This is unlike most natural images that are commonly labeled through purely objective content-based labeling mechanisms capturing the objects or actions they include [14, 13]. Instead, by focusing on art, we aim to initiate a more nuanced perceptual image understanding which, downstream, can also be applied to richer understanding of ordinary images.

We begin this effort by introducing a large-scale dataset termed *ArtEmis* [Art Emotions] that associates human emotions with artworks and contains *explanations* in natural language of the rationale behind each triggered emotion.

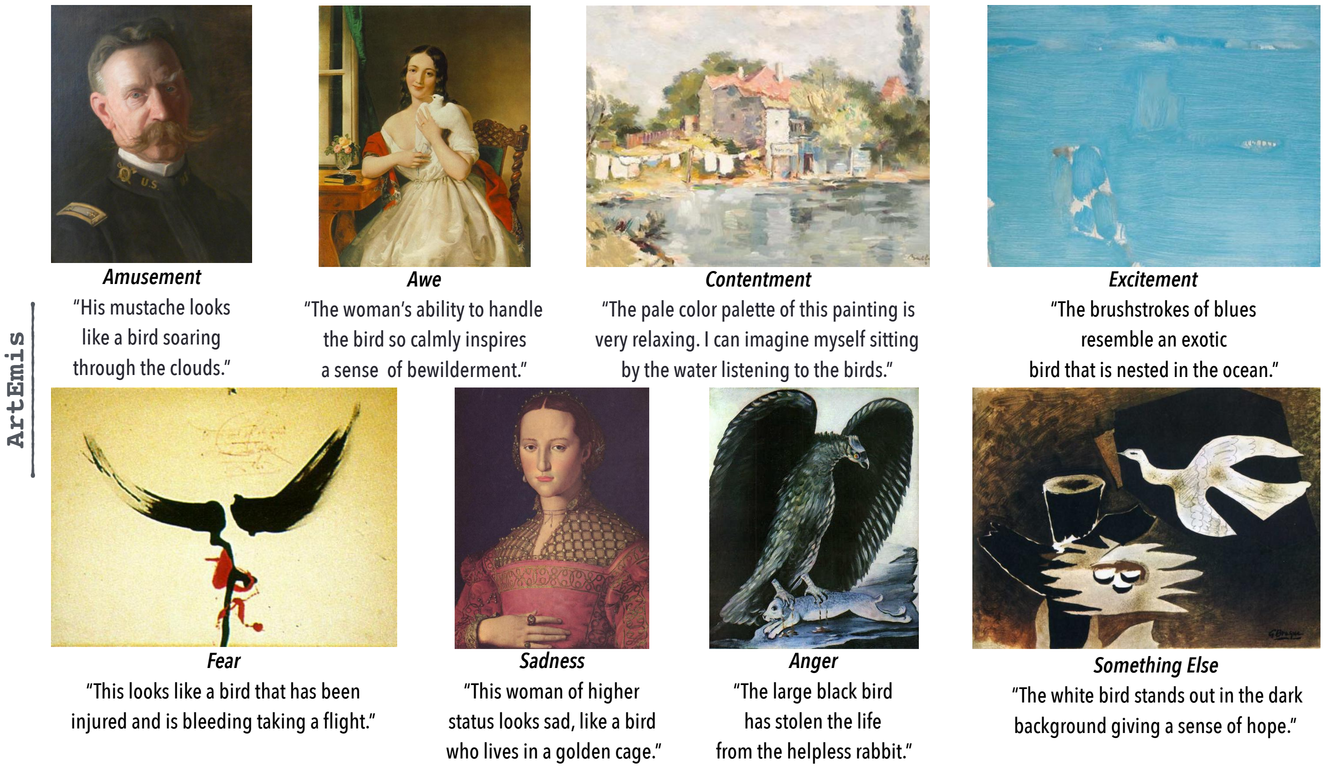

Figure 1. **Examples of affective explanations mentioning the word 'bird'.** In ArtEmis the annotators expose a wide range of abstract semantics and emotional states associated with the concept of a bird when attempting to explain their primary emotion (shown in boldface). The exposed semantics include properties that are not directly visible: *birds can be listened to, they fly, they can bring hope, but also can be sad when they are in 'golden cages'*.

Novelty of ArtEmis. Our dataset is novel as it concerns an underexplored problem in computer vision: the formation of linguistic affective explanations grounded on visual stimuli. Specifically, ArtEmis exposes moods, feelings, personal attitudes, but also abstract concepts like freedom or love, grounded over a wide variety of complex visual stimuli (see Section 3.2). The annotators typically explain and link visual attributes to psychological interpretations e.g., *'her youthful face accentuates her innocence'*, highlight peculiarities of displayed subjects, e.g., *'her neck is too long, this seems unnatural'*; and include imaginative or metaphorical descriptions of objects that do not directly appear in the image but may relate to the subject's experience; *'it reminds me of my grandmother' or 'it looks like blood'* (over 20% of our corpus contains such similes).

Subjectivity of responses. Unlike existing captioning datasets, ArtEmis welcomes the subjective and personal angle that an emotional explanation (in the form of a caption) might have. Even a single person can have a range of emotional reactions to a given stimulus [41, 50, 10, 51] and, as shown in Fig. 2, this is amplified across different annotators. The subjectivity and rich semantic content distinguish ArtEmis from, e.g., the widely used COCO dataset [14]. Figure 1 shows different images from ArtEmis with captions including the word *bird*, where the imaginative and metaphorical nature of ArtEmis is apparent (e.g., 'bird gives hope' and 'life as a caged bird'). Interestingly, despite this phenomenon, as we show later (Section 3.2), (1) there is often substantial agreement among annotators regarding their *dominant* emotional reactions, and (2) our collected explanations are often *pragmatic* – i.e., they also contain references to visual elements present in the image (see Section 3.3).

Difficulty of emotional explanations. There is debate within the neuroscience community on whether human emotions are innate, generated by patterns of neural activity, or learned [53, 4, 8, 9]. There may be intrinsic difficulties with producing emotion explanations in language – thus the task can be challenging for annotators in ways that traditional image captioning is not. Our approach is informed by significant research that argues for the central role of language in capturing and even helping to form emotions [36], including the *Theory of Constructed Emotions* [6, 7] by Lisa Feldman Barrett. Nevertheless, this debate suggests that caution is needed when comparing, under various standard metrics, ArtEmis with other captioning datasets.

Affective neural speakers. To further demonstrate the potential of ArtEmis, we experimented with building a

Awe:	"The peaceful reflections of the moonlight on the water contrast sharply with the cliffs."
Contentment:	"The steep mountains and the moonlight provide safety to the inhabitants of the isolated towns."
Excitement:	"I can imagine the sailors resting this peaceful night, dreaming of new adventures."
Fear:	"The moon casting a light over the dark mountains seems rather ominous."
Something Else:	"The glow of the moon peeking through the clouds gives a spooky, eerie feeling."

Figure 2. **Examples of different emotional reactions for the same stimulus.** The emotions experienced (in bold font) for the shown painting vary across annotators *and* are reasonably justified (next to each emotion, the annotator's explanation is given). We note that 61% of all annotated artworks have at least one positive *and* one negative emotional reaction. See Section 3.2 for details.

number of neural speakers, using deep learning language generation techniques trained on our dataset. The best of our speakers often produce well-grounded affective explanations, respond to abstract visual stimuli, and fare reasonably well in emotional Turing tests (Section 6).

In summary, we make the following key contributions:

- We introduce *ArtEmis*, a large scale dataset of emotional reactions to visual artwork coupled with explanations of these emotions in language (Section 3).
- We show how the collected corpus contains utterances that are significantly more affective, abstract, and rich with metaphors and similes, compared to existing datasets (Sections 3.1-3.2).
- Using *ArtEmis*, we develop machine learning models for dominant emotion prediction from images or text, and neural speakers that can produce plausible grounded emotion explanations (Sections 4 and 6).

2. Background and related work

Emotion classification. Following previous studies [39, 62, 67, 48], we adopt throughout this work the same discrete set of eight *categorical* emotion states. Concretely, we consider: *anger*, *disgust*, *fear*, and *sadness* as negative emotions, and *amusement*, *awe*, *contentment*, and *excitement* as positive emotions. The four negative emotions are considered universal and basic (as proposed by Ekman in [22]) and have been shown to capture well the discrete emotions of the International Affective Picture System [11]. The four positive emotions are finer grained versions of *happiness* [21]. We note that while *awe* can be associated with a negative state, following previous works ([41, 48]), we treat *awe* as a positive emotion in our analyses.

Deep learning, emotions, and art. Most existing works in Computer Vision treat emotions as an image classification problem, and build systems that try to deduce the main/dominant emotion a given image will elicit [39, 62, 67, 33]. An interesting work linking paintings to textual descriptions of their historical and social intricacies is given in [24]. Also, the work of [30] attempts to make captions for paintings in the prose of Shakespeare using language style transfer. Last, the work of [59] introduces a large scale dataset of artistic imagery with multiple attribute annotations. Unlike these works, we focus on developing machine learning tools for analyzing and generating *explanations* of emotions as evoked by artworks.

Captioning models and data. There is a lot of work and corresponding captioning datasets [64, 31, 54, 34, 40, 47] that focus on different aspects of human cognition. For instance COCO-captions [14] concern descriptions of common objects in natural images, the data of Monroe et al. [42] include discriminative references for 2D monochromatic colors, Achlioptas et al. [1, 2] collects discriminative utterances for 3D objects, etc. There is correspondingly also a large volume on deep-net based captioning approaches [38, 40, 56, 66, 43, 65, 43]. The seminal works of [58, 29] opened this path by capitalizing on advancements done in deep recurrent networks (LSTMs [27]), along with other classic ideas like training with Teacher Forcing [60]. Our neural speakers build on these 'standard' techniques, and ArtEmis adds a new dimension to image-based captioning reflecting emotions.

Sentiment-driven captions. There exists significantly less captioning work concerning sentiments (positive vs. negative emotions). Radford and colleagues [49] discovered that a single unit in recurrent language models trained without sentiment labels, is automatically learning concepts of sentiment; enabling sentiment-oriented manipulation by fixing the sign of that unit. Other early work like SentiCap [46] and follow-ups like [63], provided explicit sentiment supervision to enable sentiment-flavored language generation grounded on real-world images. These studies focus on the visual cues that are responsible for two emotional reactions (positive and negative) and, most importantly, they do not produce emotion-*explaining* language.

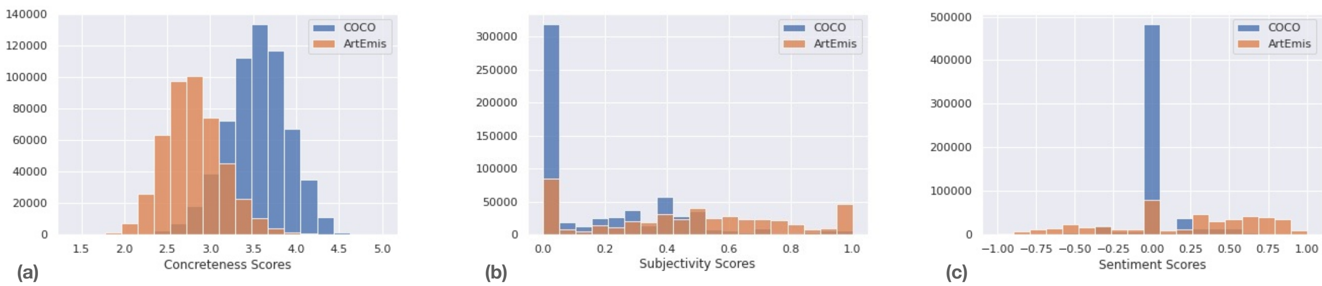

Figure 3. **Key properties of ArtEmis**. Histograms comparing ArtEmis to COCO-captions [14] along the axes of (a) *Concreteness*, (b) *Subjectivity*, and (c) *Sentiment*. ArtEmis has significantly more abstract, subjective and sentimental language than COCO-captions.

3. ArtEmis dataset

The *ArtEmis* dataset is built on top of the publicly available WikiArt[1] dataset which contains 80,031 *unique* and carefully curated artworks from 1,119 artists (as downloaded in 2015). The artworks cover 27 art-styles (abstract, baroque, cubism, impressionism, etc.) and 45 genres (cityscape, landscape, portrait, still life, etc.), constituting a very diverse set of visual stimuli [52]. In ArtEmis we annotated all artworks of WikiArt by asking *at least* 5 annotators per artwork to express their dominant emotional reaction along with an explanation for their response.

Specifically, after observing an artwork, an annotator was asked first to indicate their *dominant* reaction by selecting among the eight emotions mentioned in Section 2, or a ninth option, listed as 'something-else'. This latter option allows the annotators to express emotions not explicitly listed, or to explain why they might not have had *any* strong emotional reaction, e.g., why they felt indifferent to the shown artwork. In all cases, after the first step, the annotator was asked to provide a detailed explanation for their choice in free text that would include specific references to visual elements in the artwork. See Figures 1, 2 for examples of collected annotations.

In total, we collected **454,684** explanatory utterances and emotional responses. The resulting corpus contains 37,250 distinct words and it includes the explanations of 6,788 annotators who worked in aggregate 11,138 hours to build it. The annotators were recruited via Amazon's Mechanical Turk (AMT) services. In what follows we analyze the key characteristics of ArtEmis, while pointing the interested reader to the Supplemental Material [3] for further details.

3.1. Linguistic analysis

Richness & diversity. The average length of the captions of ArtEmis is 15.9 words which is significantly longer than the average length of captions of many existing captioning datasets as shown in Table 1. In the same table, we also show results of analyzing ArtEmis in terms of the average number of nouns, pronouns, adjectives, verbs, and adpositions. ArtEmis has a higher occurrence per caption for each of these categories compared to many existing datasets, indicating that our annotations use rich natural language in connection to the artwork and the emotion they explain. This fact becomes even more pronounced when we look at *unique*, say adjectives, that are used to explain the reactions to the same artwork among different annotators (Table 2). In other words, besides being linguistically rich, the collected explanations are also highly *diverse*.

[1] https://www.wikiart.org/

Sentiment analysis. In addition to being rich and diverse, ArtEmis also contains language that is affective. We use a rule-based sentiment analyzer (VADER [28]) to demonstrate this. The analyzer assigns only 16.5% of ArtEmis to the neutral sentiment, while for COCO-captions it assigns 77.4%. Figure 3 (c) shows the histogram of VADER's estimated valences of sentimentality for the two datasets. Absolute values closer to 0 indicate neutral sentiment.

Dataset	Words	Nouns	Pronouns	Adjectives	Adpositions	Verbs
ArtEmis	15.9	4.0	0.9	1.6	1.9	3.0
COCO Captions [14]	10.5	3.7	0.1	0.8	1.7	1.2
Conceptual Capt. [54]	9.6	3.8	0.2	0.9	1.6	1.1
Flickr30k Ent. [64]	12.3	4.2	0.2	1.1	1.9	1.8
Google Refexp [40]	8.4	3.0	0.1	1.0	1.2	0.8

Table 1. **Richness of individual captions** of ArtEmis vs. previous works. We highlight the richness of captions as units and thus show word counts averaged over *individual captions*.

Dataset	Nouns	Pronouns	Adjectives	Adpositions	Verbs
ArtEmis	18.7 (3.4)	3.1 (0.6)	8.3 (1.5)	6.5 (1.2)	13.4 (2.4)
COCO Captions [14]	10.8 (2.2)	0.6 (0.1)	3.3 (0.7)	4.5 (0.9)	4.5 (0.9)
Conceptual Capt. [54]	3.8 (3.8)	0.2 (0.2)	0.9 (0.9)	1.6 (1.6)	1.1 (1.1)
Flickr30k Ent. [64]	12.9 (2.6)	0.8 (0.2)	4.0 (0.8)	4.9 (1.0)	6.4 (1.3)
Google Refexp [40]	7.8 (2.2)	0.4 (0.1)	2.8 (0.8)	2.9 (0.8)	2.3 (0.6)

Table 2. **Diversity of captions per image** of ArtEmis vs. previous works. Shown are *unique* word counts for various parts-of-speech averaged over *individual images*. To account for discrepancies in the number of captions individual images have, we also include the correspondingly normalized averages inside parentheses.

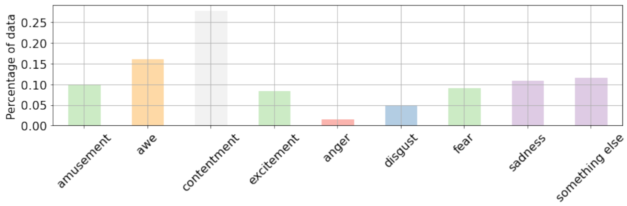

Figure 4. **Histogram of emotions captured in ArtEmis** . Positive emotions occur significantly more often than negative emotions (four left-most bars contain 62.0% of all responses vs. 5th-8th bars contain 26.3%). The annotators use a non-listed emotion ('something-else' category) 11.7% of the time.

3.2. Emotion-centric analysis.

In Figure 4 we present the histogram over the nine options that the users selected, across all collected annotations. We remark that positive emotions are chosen significantly more often than negative ones, while the "something-else" option was selected 11.7%. Interestingly, 61% of artworks have been annotated with at least one positive and one negative emotion simultaneously (this percent is 79% if we treat something-else as a third emotion category). While this result highlights the high degree of subjectivity w.r.t. the emotional reactions an artwork might trigger, we also note that that there is significant agreement among the annotators w.r.t. the elicited emotions. Namely, 45.6% (36,534) of the paintings have a strong majority among their annotators who indicated the same fine-grained emotion.

Idiosyncrasies of language use. We also explore the degree to which ArtEmis contains language that is abstract vs. concrete, subjective vs. objective, and estimate the extent to which annotators use similes and metaphors in their explanations. For measuring the abstractness or concreteness, we use the lexicon in Brysbaert et al. [12] which provides for 40,000 word lemmas a rating from 1 to 5 reflecting their concreteness. For instance, *banana* and *bagel* are maximally concrete/tangible objects, getting a score of 5, but *love* and *psyche* are quite abstract (with scores 2.07 and 1.34, resp.). A random word of ArtEmis has 2.81 concreteness while a random word of COCO has 3.55 (p-val significant, see Figure 3 (a)). In other words, ArtEmis contains on average references to more abstract concepts. Next, to measure the extent to which ArtEmis makes subjective language usage, we apply the rule-based algorithm provided by TextBlob [37] which estimates how subjective a sentence is by providing a scalar value in $[0, 1]$. E.g., *'The painting is red'* is considered a maximally objective utterance (scores 1), while *'The painting is nice'*, is maximally subjective (scores 0). We show the resulting distribution of these estimates in Figure 3 (b). Last, we curated a list of lemmas that suggest the use of similes with high probability (e.g., 'is like', 'looks like', 'reminds me of'). Such expressions appear on 20.9% of our corpus and, as shown later, are also successfully adopted by our neural-speakers.

3.3. Maturity, reasonableness & specificity.

Finally, we investigated the unique aspects of ArtEmis by conducting three separate user studies. Specifically we aim to understand: a) what is the *emotional and cognitive maturity* required by someone to express a random ArtEmis explanation?, b) how *reasonable* a human listener finds a random ArtEmis explanation, even when they would not use it to describe their own reaction?, and last, c) to what extent the collected explanations can be used to *distinguish* one artwork from another? We pose the first question to Turkers in a binary (yes/no) form, by showing to them a randomly chosen artwork and its accompanying explanation and asking them if this explanation requires emotional maturity higher than that of a typical 4-year old. The answer for 1K utterances was 'yes' **76.6%** of the time. Repeating the same experiment with the COCO dataset, the answer was positive significantly less (**34.5%**). For the second question, we conducted an experiment driven by the question "Do you think this is a realistic and reasonable emotional response that could have been given by someone for this image?". We elaborate on the results in Supp. Mat.; in summary, **97.5%** of the utterances were considered appropriate. To answer the final question, we presented Turkers with one piece of art coupled with one of its accompanying explanations, and placed it next to two random artworks, side by side and in random order. We asked Turkers to guess the 'referred' piece of art in the given explanation. The Turkers succeeded in predicting the 'target' painting **94.7%** of the time in a total of 1K trials.

4. Neural methods

4.1. Auxiliary classification tasks

Before we present the neural speakers we introduce two auxiliary *classification* problems and corresponding neural-based solutions. First, we pose the problem of predicting the emotion explained with a given textual explanation of ArtEmis. This is a classic 9-way text classification problem admitting standard solutions. In our implementations we use cross-entropy-based optimization applied to an LSTM text classifier trained from scratch, and also consider fine-tuning to this task a pretrained BERT model [20].

Second, we pose the problem of predicting the expected distribution of emotional reactions, given an artwork. To address this problem we fine-tune a ResNet-32 encoder [26] pretrained on ImageNet [18] by minimizing the KL-divergence between its output and the empirical user distributions of ArtEmis. Having access to these two classifiers, which we denote as $C_{emotion|text}$ and $C_{emotion|image}$ respectively, is useful for our neural speakers as we can use them to evaluate, and also, steer, the emotional content of their output (Sections 5 and 4.2). Of course, these two problems have also intrinsic value and we explore them in detail in Section 6.

4.2. Affective neural speakers

Baseline with ANPs. In order to illustrate the importance of having an emotion-explanation-oriented dataset like ArtEmis for building affective neural speakers; we borrow ideas from previous works [63, 46] and create a baseline speaker that does not make any (substantial) use of ArtEmis. Instead, and similar to what was done for the baseline presented in [46], we first train a neural speaker with the COCO-caption dataset and then we inject *sentiment* to its generated captions by adding to them appropriately chosen adjectives. Specifically we use the intersection of Adjective Noun Pairs (ANPs) between ArtEmis and the ANPs of [46] (resulting in 1,177 ANPs, with known positive and negative sentiment) and capitalize on the $C_{emotion|image}$ to decide what sentiment we want to emulate. If the $C_{emotion|image}$ is maximized by one of the four positive emotion-classes of ArtEmis, we inject the adjective corresponding to the *most frequent* (per ArtEmis) positive ANP, to a randomly selected noun of the caption. If the maximizer is negative, we use the corresponding ANP with negative sentiment; last, we resolve the something-else maximizers ($<$10%) by fair coin-flipping among the two sentiments. We note that since we apply this speaker to ArtEmis images and there is significant visual domain gap between COCO and WikiArt, we fine-tune the neural-speaker on a small-scale and separately collected (by us) dataset with *objective* captions for 5,000 wikiArt paintings. We stress that this new dataset was collected following the AMT protocol used to build COCO-captions, i.e., asking only for objective (not affective) descriptions of the main objects, colors etc. present in an artwork.

Basic ArtEmis speakers. We experiment with two popular backbone architectures when designing neural speakers trained on ArtEmis: the classic Show-Attend-Tell (SAT) approach [61], which combines an image encoder with a word/image attentive LSTM; and the recent line of work of top-down, bottom-up meshed-memory transformers (M^2) [15], which replaces the recurrent units with transformer units and similarly to Andersen et al. [5] relies on separately computed object-bounding-box detections (computed using Faster R-CNN [25]). We also include a much simpler baseline that uses ArtEmis: for a testing image we find its nearest *visual* neighbor in the training set (using features from a ResNet-32 pretrained on ImageNet) and output one of the latter's human ground-truth captions at random.

Emotion grounded speaker. We additionally tested neural speakers that make use of the emotion classifier, i.e., $C_{emotion|image}$. At training time, in addition to grounding the a neural-speaker with the visual stimulus and applying teacher forcing with the captions of ArtEmis, we further provide at each time step a feature (extracted via a fully-connected layer) of the emotion-label in that particular training example. This extra signal promotes the *decoupling* of the emotion conveyed by the linguistic generation, from the underlying image. In other words, this variant allows us to independently set the emotion we wish to explain for a given image. At inference time (to keep things fair) we deploy first the $C_{emotion|image}$ over the test artwork, and use the output maximizing emotion, to first ground and then sample the generation of this variant.

Details. To ensure a meaningful comparison between neural speakers, we use the same image-encoders, learning-rate schedules, LSTM hidden-dimensions, etc. across all of them. When training with ArtEmis we use an $[85\%, 5\%, 10\%]$ train-validation-test data split and do model-selection (optimal epoch) according to the model that minimizes the negative-log-likelihood on the validation split. For the ANP baseline, we use the Karpathy splits [29] to train the same (SAT) backbone network we used elsewhere. When *sampling* a neural speaker, we keep the test generation with the highest log-likelihood resulting from a greedy beam-search with beam size of 5 and a soft-max temperature of 0.3. An exception to the above (uniform) experimental protocol was made for the speakers trained with Meshed Transformers. In this case we used the publicly available implementation [16] with minimal adaptation.

5. Evaluation

In this section we describe the evaluation protocol we follow to quantitatively compare our trained neural networks. First, for the auxiliary classification problems we report the average attained accuracy per method. Second, for the evaluation of the neural speakers we use three categories of metrics that assess different aspects of their quality. To measure the extent to which our generations are linguistically similar to held-out ground-truth human captions, we use various popular machine-based metrics: e.g., BLEU 1-4 [45], ROUGE-L [35], METEOR [19].

We highlight that CIDEr-D [57] which requires a generation to be semantically close to *all* human-annotations of an artwork, is not a metric well-suited for ArtEmis, due to the large diversity and inherent subjectivity of our dataset. We also evaluate the *novelty* of the captions of our neural speakers; here we report the average maximum length of the longest common subsequence for a generation and (a subsampled version) of all training utterances. The smaller this metric is, the farther away one can assume that the generations are from the training data [23]. We also report the fast to compute number (fraction) of *unique* generations made over an input set of images. The third axis of evaluation concerns two unique properties of ArtEmis and affective explanations in particular. First, we report the fraction of a speaker's productions that contain similes, i.e., generations that have lemmas like 'thinking of', 'looks like' etc. This fraction is a proxy for how often a neural speaker chooses to utter metaphorical-like content. Secondly, by tapping on the $C_{emotion|text}$, we can compute which emotion is most likely explained by the generated utterance; this estimate allows us to measure the extent to which the deduced emotion is 'aligned' with some ground-truth. Specifically, for test artworks where the emotion annotations form a strong majority, we define the *emotional-alignment* as the percent of the grounded generations where the $\arg\max(C_{emotion|generation})$ agrees to the emotion made by the majority.

The above metrics are algorithmic, i.e., they do not involve direct *human judgement*, which is regarded as the golden standard for quality assessment [17, 32] of synthetic captions. The discrepancy between machine and human-based evaluations can be exacerbated in a dataset with subjective and affective components like ArtEmis. To address this, we evaluate ArtEmis-trained basic and emotion-grounded speaker variants, via user studies that emulate a Turing test; i.e., they assess the extent to which the synthetic captions can be 'confused' as being made by humans.

6. Experimental results

Estimating emotion from text or images alone. We found experimentally that predicting the fine-grained emotion explained in ArtEmis data is a difficult task (see examples where both humans and machines fail in Table 3). In a small-scale study with experts (authors of this paper), humans could infer the explained emotion from the text alone 61.2% accurately (in 500 trials). Interestingly, the neural networks of Section 4.1 attained 63.3% and 64.8% (LSTM, BERT respectively) on the entire test split used by the neural-speakers (39,850 utterances). Crucially, both humans and neural-nets failed gracefully in their predictions and most confusion happened among subclasses of the same, positive or negative category (we include confusion matrices in the Supp. Mat.). For instance, if we binarize the predictions made on the 9-way problem and the ground-truth labels into positive vs. negative emotion sentiment (ignoring the something-else class); the experts, the LSTM-based, and the BERT-based models, guess correctly 85.9%, 89.4%, 91.5% of the time, respectively.

Since we train our image classifiers to predict a distribution of emotions, we select the maximizer of their output and compare it with the 'dominant' emotion of the test images for which the emotion distribution is unimodal with a mode covering more than 50% of the mass (38.1% of the split). The attained accuracy for this sub-population is 60.2%.

ArtEmis Utterance	Guess	GT
"The scene reminds me of a perfect summer day."	Contentment (H)	Awe
"This looks like me when I don't want to get out of bed on Monday morning."	Something-Else (M)	Amusement
"A proper mourning scene, and the mood is fitting."	Sadness (H)	Contentment

Table 3. **Examples showcasing the difficulty of emotion-deduction from text.** The first two examples' interpretation depends highly on personal experience (first & middle row). The third example uses language that is emotionally subtle. (H): human-guess, (M): neural-net guess, GT: ground-truth.

Neural speakers. In Table 4 we report the machine-induced metrics described in Section 5. First, we observe that on metrics that measure the linguistic similarity to held-out utterances (BLEU, METEOR, etc.) our speakers fare noticeably worse as compared to how these architectures fare when trained and tested with objective datasets like COCO-captions; e.g., BLEU-1 with SOTA [15] is 82.0. This is expected given the analysis of Section 3 that shows how ArtEmis is a significantly more diverse and subjective dataset. Second, there is a noticeable difference in all metrics in favor of the four models trained with ArtEmis (denoted as Basic or Grounded) against the simpler baselines that do not. This implies that we cannot simply reproduce ArtEmis with ANP injection on objective data. It further demonstrates how even among similar images the annotations can be widely different, limiting the Nearest-Neighbor (NN) performance. Third, on the emotion-alignment metric the emotion-grounded variants fare significantly better than their non-grounded version. These variants also produce a percent of similes closer to the ground-truth's percentage of 20.9. However, as seen by the Longest Common Subsequence (denoted as LCS) and the fraction of unique generations these variants also tend to create less novel captions.

Qualitative results of the emotion-grounded SAT speaker are shown in Figure 5. As seen in Figure 5 this speaker can create pragmatic explanations that can include visual analogies, or nuanced associations in support of the grounding emotion. More examples, including typical failure cases and generations from other variants, are provided in the Supplemental Material.

Turing test. For our last experiment, we performed a user study taking the form of a Turing Test deployed in AMT. First, we use a neural-speaker to make one explanation for a test artwork and couple it with a randomly chosen ground-truth for the same stimulus. Next, we show to a user the two utterances in text, along with the artwork, and ask them to make a multiple choice among 4 options. These were to indicate either that one utterance was more likely than the other as being made by a human explaining their emotional reaction; or, to indicate that both (or

Figure 5. **Examples of a neural speaker productions on *unseen* artworks.** The produced explanations reflect a variety of dominant emotional-responses (shown above each utterance in bold font). The top row shows examples where the deduced grounding emotion was positive; the bottom row shows three examples where the deduced emotion was negative and an example from the something-else category. Remarkably, the neural speaker can produce pragmatic explanations that include **visual analogies**: *looks like it is floating, like it is boiling fire*, and **nuanced** explanations of affect: *calm and meditative, pain and suffering*. Examples sampled from the SAT-based variant.

none) were likely made by a human. We deploy this experiment with 500 artworks, and repeat it separately for the basic and the emotion-grounded (SAT) speakers. Encouragingly, **50.3%** of the time the users signaled that the utterances of the emotion-grounded speaker were on-par with the human groundtruth (20.6%, were selected as the more human-like of the pair, and 29.7% scored a tie). Furthermore, this variant also achieved significantly better results than the basic speaker, which surpassed or tied to the human annotations 40% of the time (16.3% with a win and and 23.7% as a tie). To explain this differential, we hypothesize that grounding with the *most likely* emotion steered the better-performing variant to create more commonplace explanations which thus were harder to discriminate as non-human plausible.

metric	NN	ANP	M^2(Basic/Grounded)	SAT (Basic/Grounded)
BLEU-1	0.364	0.396	0.507 / 0.511	**0.536** / 0.520
BLEU-2	0.139	0.134	0.282 / 0.282	**0.290** / 0.280
BLEU-3	0.054	0.042	**0.159** / 0.154	0.155 / 0.146
BLEU-4	0.022	0.014	**0.095** / 0.090	0.087 / 0.079
METEOR	0.102	0.088	0.140 / 0.137	**0.142** / 0.134
ROUGE-L	0.210	0.202	0.280 / 0.286	**0.297** / 0.294
max-LCS	7.513	**6.299**	8.286 / 8.141	7.955 / 7.632
Unique-fraction	**0.960**	0.730	0.250 / 0.230	0.480 / 0.460
Emo-Alignment	0.327	0.406	0.410 / **0.521**	0.406 / 0.519
Similes-fraction	**0.200**	0.001	0.709 / 0.437	0.481 / 0.268

Table 4. **Neural speaker machine-based evaluations**. NN: Nearest Neighbor baseline, ANP: baseline-with-injected sentiments, M^2: Meshed Transformer, SAT: Show-Attend-Tell. The Basic models use for grounding only the underlying image, while the Grounded variants also input an emotion-label.

7. Conclusion

Human cognition has a strong affective component that has been relatively undeveloped in AI systems. Language that explains emotions generated at the sight of a visual stimulus gives us a way to analyze how image content is related to affect, enabling learning that can lead to agents emulating human emotional responses through data-driven approaches. In this paper, we take the first step in this direction through: (1) the release of the ArtEmis dataset that focuses on linguistic explanations for affective responses triggered by visual artworks with abundant emotion-provoking content; and (2) a demonstration of neural speakers that can express emotions and provide associated explanations. The ability to deal computationally with images' emotional attributes opens an exciting new direction in human-computer communication and interaction.

Acknowledgements. This work is funded by a Vannevar Bush Faculty Fellowship, a KAUST BAS/1/1685-01-01, a CRG-2017-3426, the ERC Starting Grant No. 758800 (EXPROTEA), the ANR AI Chair AIGRETTE, and gifts from the Adobe, Amazon AWS, Autodesk, and Snap Corp.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. *European Conference on Computer Vision (ECCV)*, 2020. 3

[2] 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. 3

[3] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. *Supplementary Material for ArtEmis: Affective Language for Visual Art*, (accessed January 1st, 2021). Available at https://artemisdataset.org/materials/artemis_supplemental.pdf, version 1.0. 4

[4] Ralph Adolphs. How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences. *Social cognitive and affective neuroscience*, 2017. 2

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[6] Lisa Feldman Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 2006. 2

[7] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 2017. 2

[8] Lisa Feldman Barrett, Kristen A Lindquist, Eliza Bliss-Moreau, Seth Duncan, Maria Gendron, Jennifer Mize, and Lauren Brennan. Of mice and men: Natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. *Perspectives on Psychological Science*, 2007. 2

[9] Lisa Feldman Barrett and Ajay B Satpute. Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience letters*, 2019. 2

[10] M. M. Bradley, M. K. Greenwald, M.C. Petry, and P. J. Lang. Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1992. 2

[11] Margaret M. Bradley and Peter J. Lang. The international affective picture system (iaps) in the study of emotion and attention. *Series in affective science. Handbook of emotion elicitation and assessment*, 2007. 3

[12] Marc Brysbaert, Amy Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 2014. 5

[13] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[14] 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. *CoRR*, abs/1504.00325, 2015. 1, 2, 3, 4

[15] 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. 6, 7

[16] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. *Meshed-Memory Transformer for Image Captioning*, (accessed September 1st, 2020). Available at https://github.com/aimagelab/meshed-memory-transformer. 6

[17] 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

[18] 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. 5

[19] 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

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 5

[21] Ed Diener, Christie Napa Scollon, and Richard E Lucas. The evolving concept of subjective well-being: The multifaceted nature of happiness. *Advances in Cell Aging & Gerontology*, 2003. 3

[22] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 1992. 3

[23] Angela Fan, Mike Lewis, and Yann Dauphin. Strategies for structuring story generation. *CoRR*, abs/1902.01109, 2019. 6

[24] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. *CoRR*, abs/1810.09617, 2018. 3

[25] Ross Girshick. Fast r-cnn. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[26] 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

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 3

[28] 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

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 6

[30] Prerna Kashyap, Samrat Phatale, and Iddo Drori. Prose for a painting. *CoRR*, abs/1910.03634, 2019. 3

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and L. Tamara Berg. Referitgame: Referring to objects in pho-

tographs of natural scenes. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 3

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

[33] 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. 3

[34] 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*, 2017. 3

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

[36] Kristen A Lindquist, Jennifer K MacCormack, and Holly Shablack. The role of language in emotion: Predictions from psychological constructionism. *Frontiers in psychology*, 2015. 2

[37] Steven Loria. *TextBlob*, (accessed November 16, 2020). Available at `https://textblob.readthedocs.io/en/dev/`. 5

[38] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[39] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, 2010. 3

[40] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Murphy Kevin. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283, 2016. 3, 4

[41] 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, 3

[42] Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *CoRR*, abs/1703.10186, 2017. 3

[43] K. Varun Nagaraja, I. Vlad Morariu, and Davis S. Larry. Modeling context between objects for referring expression understanding. *European Conference on Computer Vision (ECCV)*, 2016. 3

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

[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 2002. 6

[46] Alexander Mathews Patrick, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. *AAAI*, 2016. 3, 6

[47] 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. 3

[48] You Quanzeng, Luo Jiebo, Jin Hailin, and Yang Jianchao. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *CoRR*, abs/1605.02677, 2016. 3

[49] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444, 2017. 3

[50] David Rubin and Jennifer Talarico. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory (Hove, England)*, 2009. 2

[51] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980. 2

[52] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *CoRR*, abs/1505.00855, 2015. 4

[53] Alexander J Shackman and Tor D Wager. The emotional brain: Fundamental questions and strategies for future research. *Neuroscience letters*, 2019. 2

[54] 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. 3, 4

[55] Leo Tolstoy. *What is Art?* London: Penguin, 1995 [1897]. 1

[56] Ramakrishna Vedanta, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. *CoRR*, abs/1701.02870, 2017. 3

[57] 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

[58] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2015. 3

[59] J. Michael Wilber, Chen Fang, H. Jin, Aaron Hertzmann, J. Collomosse, and J. Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. *International Conference on Computer Vision (ICCV)*, 2017. 3

[60] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1989. 3

[61] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015. 6

[62] 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. 3

[63] Quanzeng You, Hailin Jin, and Luo Jiebo. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *CoRR*, abs/1801.10121, 2018. 3, 6

[64] 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*, 2014. 3, 4

[65] Licheng Yu, Zhe Lin, Xiaohui Shen, Yangm Jimei, Xin Lu, Mohit Bansal, and L. Tamara Berg. Mattnet: Modular attention network for referring expression comprehension. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[66] Licheng Yu, Patrick Poirson, Shan Yang, C. Alexander Berg, and L. Tamara Berg. Modeling context in referring expressions. *European Conference on Computer Vision (ECCV)*, 2016. 3

[67] 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. 3