Affection: Learning Affective Explanations for Real-World Visual Data

Panos Achlioptas\textsuperscript{1,3} Maks Ovsjanikov\textsuperscript{2} Leonidas Guibas\textsuperscript{3} Sergey Tulyakov\textsuperscript{1}

\textsuperscript{1}Snap Inc. \textsuperscript{2}LIX, Ecole Polytechnique, IP Paris \textsuperscript{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 caption-
ing [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 Affection Explanation Captioning (AEC) for real-world images. To tackle AEC we release a new large-scale dataset, 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 captions highly related to their experiences (7, 58, 64, 86), 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 Affection Explanation Captioning (AEC) for real-world images. To tackle AEC we release a new large-scale dataset, 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 captions highly related to their experiences (7, 58, 64, 86), 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 Affection Explanation Captioning (AEC) for real-world images. To tackle AEC we release a new large-scale dataset, 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 captions highly related to their experiences (7, 58, 64, 86), 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 Affection Explanation Captioning (AEC) for real-world images. To tackle AEC we release a new large-scale dataset, 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 captions highly related to their experiences (7, 58, 64, 86), 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.
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.

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

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.
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
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 $C_{\text{emotion|text}}$ and $C_{\text{emotion|image}}$ 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 pre-trained 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 $C_{\text{emotion|image}}$ 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:

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

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 $\beta$ 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.

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.
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) ∈ [−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 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
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 \(C_{\text{emotion|caption}}\) 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 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 Affective Explanation Captioning (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 broken down and dirty bathroom makes me feel disgusted.

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

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.

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

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

The broken down and dirty bathroom makes me feel disgusted.

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

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’).

Table 2. Neural speaker machine-based evaluations. The Default model uses 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.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Speaker Variants</th>
<th>Default</th>
<th>Emo-Grounded (Pragmatic)</th>
<th>Emo-Grounded (Pragmatic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1.2 (↑)</td>
<td>Default</td>
<td>64.4, 38.3</td>
<td>63.1, 36.9</td>
<td>64.3, 38.0</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>63.1, 36.9</td>
<td>64.3, 38.0</td>
<td>63.4, 37.0</td>
</tr>
<tr>
<td>BLEU-3.4 (↑)</td>
<td>Default</td>
<td>22.2, 13.2</td>
<td>20.9, 12.0</td>
<td>21.8, 12.8</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>20.9, 12.0</td>
<td>21.8, 12.8</td>
<td>20.9, 11.9</td>
</tr>
<tr>
<td>METEOR (↑)</td>
<td>Default</td>
<td>7.4</td>
<td>14.4</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>14.4</td>
<td>15.1</td>
<td>14.8</td>
</tr>
<tr>
<td>ROUGE-L (↑)</td>
<td>Default</td>
<td>30.8</td>
<td>30.5</td>
<td>31.0</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>30.5</td>
<td>30.8</td>
<td>30.8</td>
</tr>
<tr>
<td>SPICE (↑)</td>
<td>Default</td>
<td>7.4</td>
<td>7.2</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>7.2</td>
<td>7.7</td>
<td>7.7</td>
</tr>
<tr>
<td>CLIPScore (↑)</td>
<td>Default</td>
<td>66.7</td>
<td>66.8</td>
<td>69.2</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>66.8</td>
<td>69.2</td>
<td>69.2</td>
</tr>
<tr>
<td>RetCLIPScore (↑)</td>
<td>Default</td>
<td>75.0</td>
<td>75.0</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>75.0</td>
<td>76.3</td>
<td>76.3</td>
</tr>
<tr>
<td>Unique-Productions (↑)</td>
<td>Default</td>
<td>78.7</td>
<td>80.7</td>
<td>82.9</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>80.7</td>
<td>82.9</td>
<td>83.7</td>
</tr>
<tr>
<td>Max-LCS (↑)</td>
<td>Default</td>
<td>70.4</td>
<td>70.4</td>
<td>68.6</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>70.4</td>
<td>68.6</td>
<td>68.4</td>
</tr>
<tr>
<td>ClipDivCos (↑)</td>
<td>Default</td>
<td>73.1</td>
<td>72.8</td>
<td>69.8</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>72.8</td>
<td>69.8</td>
<td>70.2</td>
</tr>
<tr>
<td>Similes (↑)</td>
<td>Default</td>
<td>42.8</td>
<td>36.3</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>36.3</td>
<td>34.5</td>
<td>34.5</td>
</tr>
<tr>
<td>Emo-Alignment (↑)</td>
<td>Default</td>
<td>48.1</td>
<td>55.2</td>
<td>48.2</td>
</tr>
<tr>
<td></td>
<td>Emo-Grounded</td>
<td>55.2</td>
<td>55.9</td>
<td>55.9</td>
</tr>
</tbody>
</table>

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.

Received 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 neutral 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


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


