AFFECTION: LEARNING AFFECTIVE EXPLANATIONS FOR REAL-WORLD VISUAL DATA
Supplemental Material

Panos Achlioptas[1] * panos@snap.com
Maks Ovsjanikov[2] maks@lix.polytechnique.fr
Leonidas Guibas[3] guibas@cs.stanford.edu
Sergey Tulyakov[1] stulyakov@snap.com

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

https://affective-explanations.org
October 9th 2022

A Details on Building Affection

We build Affection by annotating images existing in the following **five** datasets: MS-COCO [CFL+15], Visual-Genome [KZG+17], Flickr30k Entities [PWC+15], Emotional-Machines [KKKL18] and the images considered in the work of Quanzeng *et al.* [QJHJ16]. Specifically, we begin by annotating with affective responses *all* images in the latter two emotion-oriented works. We then proceed by using the images in Quanzeng *et al.* to find for each one of them its three *nearest-neighbors* in the image collections of MS-COCO, Visual-Genome and Flickr30k Entities, respectively. We include and annotate with affective responses the found neighbors, resulting in covering additionally 22,770 images from MS-COCO, 13,202 from Flickr30k Entities, and 16,437 from Visual-Genome.

To implement the nearest neighbor search we use the 512D embedding space formed by the output weights of the final convolutional layer of a ResNet-32 [HZRS15], pre-trained on ImageNet [DDS+09]. Before running the search algorithm, we apply an average pooling to the 7×7 spatial dimensions of the ResNet layer (forming a $1 \times 1 \times 512$ embedding vector per image).

Secondary details. For the Visual Genome, it is worth noting that we restrict the nearest-neighbor search on its 56,506 images (out of 108,077) that are *not* included in COCO [CFL+15], or Flickr30k Ent. [PWC+15], to potentially discover a larger number of *unique* neighbors across the individual datasets. As a final step, upon aggregating all relevant images from all corresponding (five) datasets, we use "fdups" [Lop22] to discover and remove possible duplicates among them. For the final version of Affection, we remove 198 duplicates found in this manner.

B Analyzing Properties of Affection

In this Section, we briefly include some supplementary analysis, similar in spirit to the one presented in Section 3 of the main paper [AOGT22].

First, we analyze how some of the key linguistic properties discussed in that main paper, are manifested in the annotations collected for each of the five underlying image datasets used to build Affection. Namely, we report the average attained scores (computed with the same methods described in the main paper) for the properties of *concreteness*, *subjectivity* and use of *sentimental* language. As seen in Figure 1, the annotations collected based on grounding visual stimuli found in the emotion-oriented datasets of Emotional-Machines [KKKL18] and of Quanzeng *et al.*, result on average in *only slightly* more abstract, subjective and sentimental affective explanations (language), compared to the subset of images of the remaining datasets. In other words, it appears that w.r.t. these key characteristics of Affection's annotations, the images used across *all* underlying datasets do not result in any significant discrepancy among the responses they evoked.

*Corresponding author.

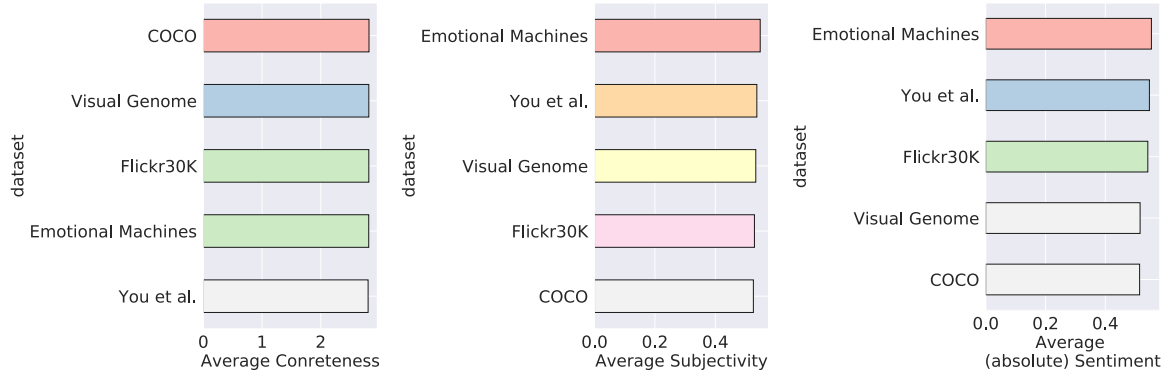

Figure 1: **Measuring key properties of Affection across its underlying image datasets**. Histograms comparing Affection in each of its underlying image datasets along the axes of (a) *Concreteness*, (b) *Subjectivity*, and (c) *Sentiment*.

Second, for the above described propertied, we also compare Affection to ArtEmis [AOH+21] (Figure 2). As mentioned in the main paper, Affection and ArtEmis are similar in terms of their average concreteness scores (average scores of 2.82 vs. 2.81), but Affection contains significantly more subjective and sentimental annotations (see histograms (b) and (c) of Figure 2).

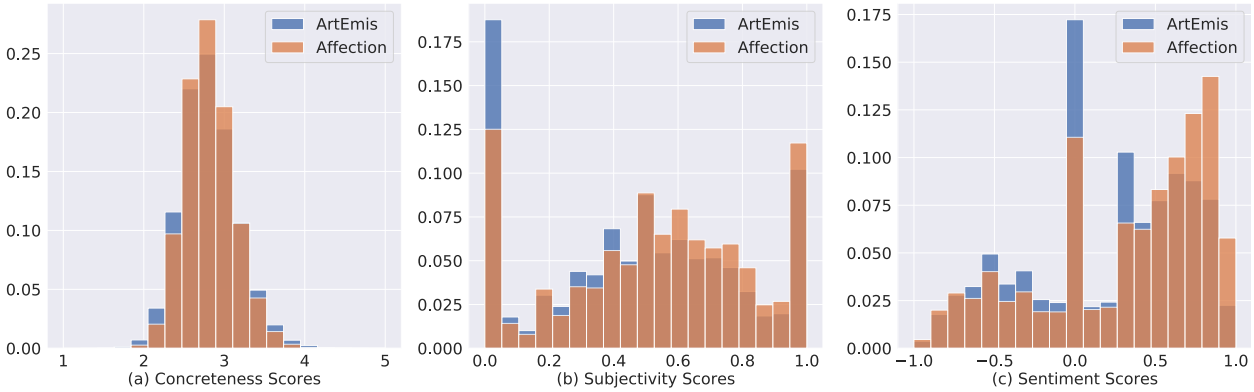

Figure 2: **Comparing Affection to ArtEmis** along the axes of (a) *Concreteness*, (b) *Subjectivity*, and (c) *Sentiment*. The histograms presented here are analogous to those contrasting Affection to COCO in Figure 2 of the main paper.

C Fine-grained Emotion Classification from a Single Modality

As stated in Section 6 of the main paper, the auxiliary emotion classifiers trained with Affection fail gracefully, as in, they primarily confuse fine-grained emotion classes of the same (positive or negative) sentiment. Here, we include the actual confusion matrices for the ResNet101-based image-2-emotion classifier (Figure 4) and the LSTM-based text-2-emotion classifier (Figure 3). We note that for the image-2-emotion classifier we use during training and inference only images for which there is a strong majority among the annotators w.r.t. the emotions they indicated. Crucially, as stated in the main paper, the underlying distribution of emotions when considering only such images is highly imbalanced (see Figure 7).

Figure 3: **Confusion matrix for an LSTM-based text2emotion 9-way classifier trained and tested with Affection's explanations**.

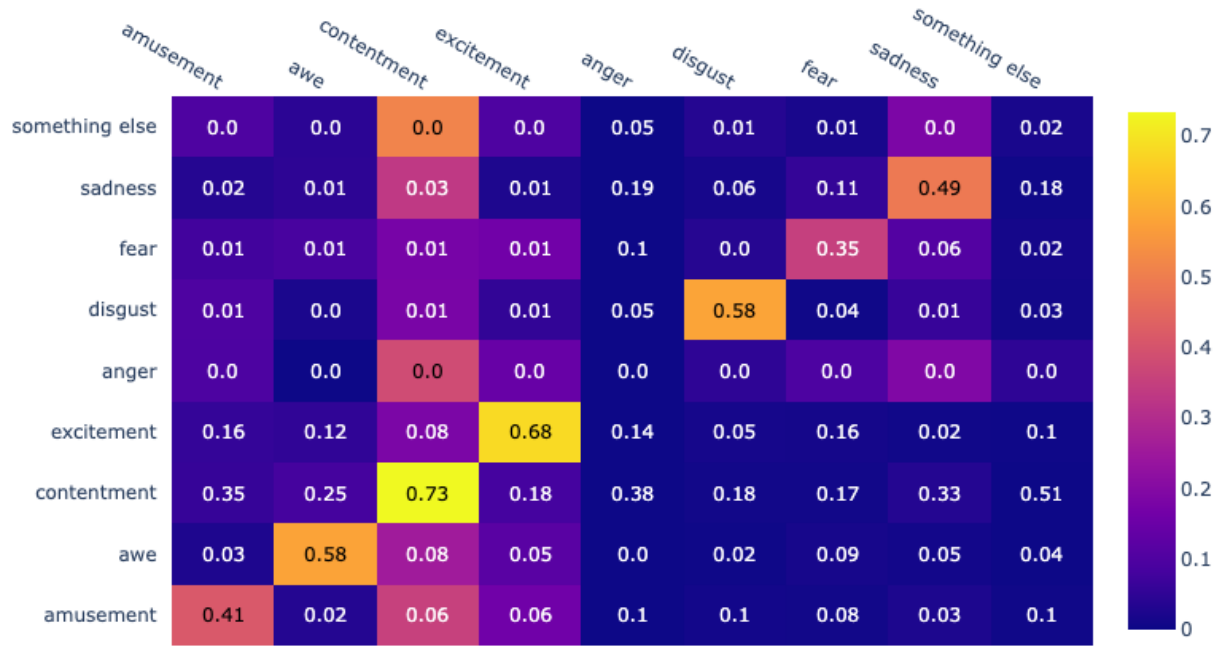

Figure 4: **Confusion matrix for a ResNet-101 pretrained image2emotion 9-way classifier trained and tested with Affection's emotion labels**. Only images and emotion labels for which there is a unique strong majority among the dominant emotions indicated by the annotators are used in this experiment.

D Neural Listeners and Speakers

Figure 8 displays the average test performance of an LSTM + ResNet-101 -based (contrastive) neural listener trained from scratch with ArtEmis explanations. Despite, the fact that this listener uses Affection's annotations for training, it performs on average worse than the non-finetuned CLIP-based model [RKH+21] presented in the main paper. Presumably, this fact, is due to it using simpler components for image and language encoding.

Figure 5 displays examples of some of our neural speakers' characteristic (common) failure modes. The first problem oftentimes faced by *all* of our speaker variants is their inability to recognize the underlying object classes of the depicted objects in the grounding image. Thus, their generations might appear to ground their explanations on objects not actually displayed, e.g., describe properties of a male human when only females are shown. This generic error appears in numerous captioning systems and is not specific only to speakers trained with affective explanations. However, this problem can be more severe in typical affective imagery since such images tend to have more subtle and abstract semantics (e.g., pizza-like-looking wall clock, example (A)). A second but less frequently occurring problem that is also faced by all speaking variants is that they can sometimes create non-sensible emotional assessments, e.g., a human would find it strange to describe a bicycle as being calm (example (B)). Besides these generic problems, the main idiosyncratic problem we observed with the emotion-grounded variant is that it can overfocus (compared to other variants) on language concerning the underlying emotion while missing to ground key visual details. For instance, for image (C), the default variant produces '*I feel sad because the monkey looks like he is* **trapped in a cage**'. Finally, the pragmatic variant, unlike the emotion-grounded one, sometimes might try too hard to use specific visual details in its explanations, creating errors like those seen for image (D) – for which the default variant produces '*The zebras are beautiful and I would love to see them in the wild*'.

(A)
(Default speaker)
"the pizza looks so delicious and I would love to eat it"

(B)
(Default speaker)
"the bike is so calm and the trees are so beautiful"

(C)
(Emo-grounded speaker)
"this monkey looks sad and I feel sad for him" *

(D)
(Pragmatic speaker)
"the two zebras in the water look like they are having a good time"

Figure 5: **Most common failure modes of our affective neural speakers.** Left-most two examples show *generic* problems that *all* neural variants might suffer from: e.g., misidentifying the underlying visual elements (example A) or making non-sensible emotional judgments (example B). While the third example (C) is sensible, it highlights how an emo-grounded variant can overfocus on the underlying emotion and miss crucial visual details (e.g., the fence). On the contrary, the pragmatic variant (example D) can overcompensate by wrongly mentioning visual details (the default neural speaker simply mentions the zebras in this example). For more details see Section D.

We note that during inference for all neural speaking variants and the results presented in the main paper, we use beam-search with a beam size of 20 and a soft-max temperature for the layer predicting each generated token of 0.3. For the pragmatic variants, the β parameter described in Section 4 controlling the influence of the internal (judging) listener is set to 0.25.

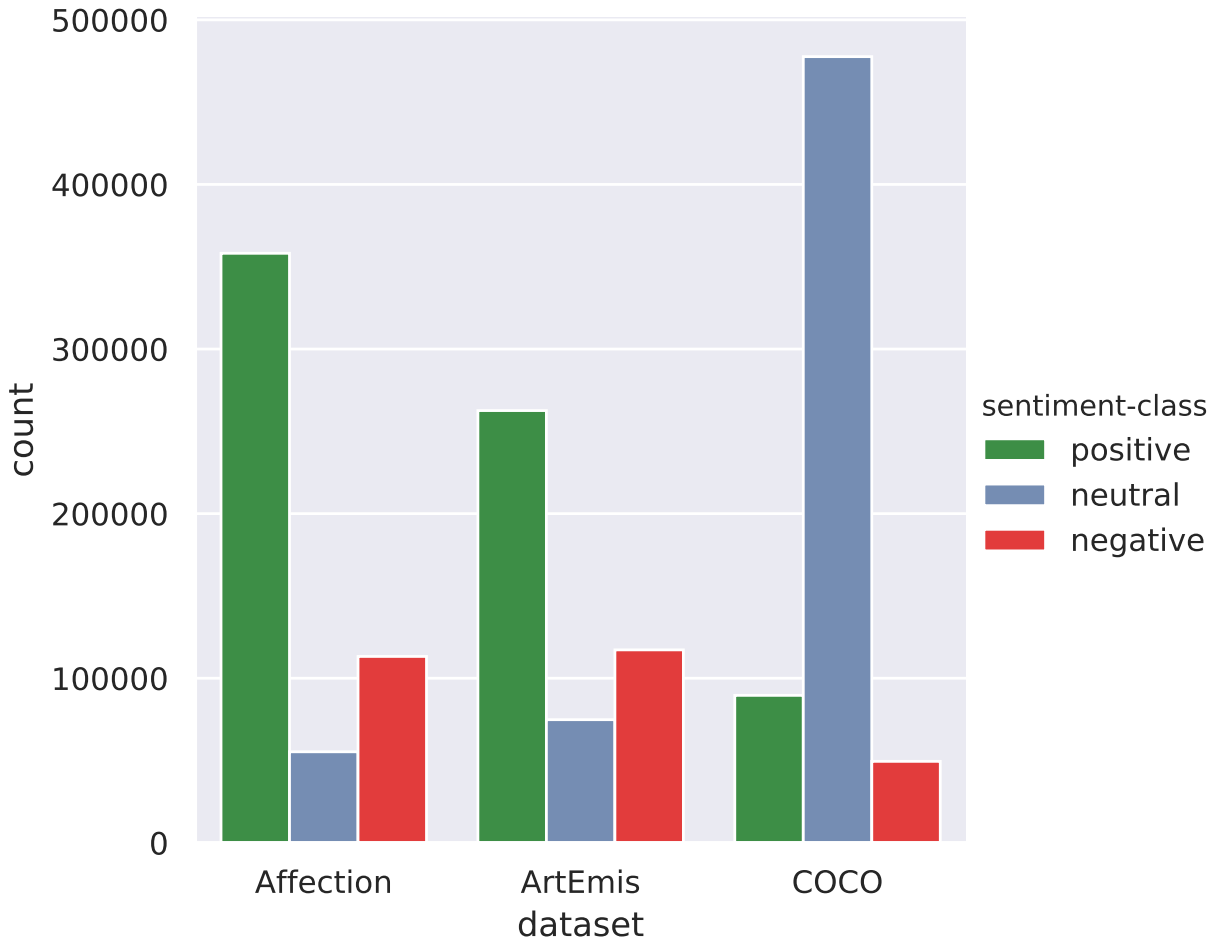

Figure 6: **Sentiment classes per dataset**. Using VADER's [HG14] sentiment classifier to assign the utterances of the shown datasets, in one of three classes. Affection's utterances are on average consider the least neutral.

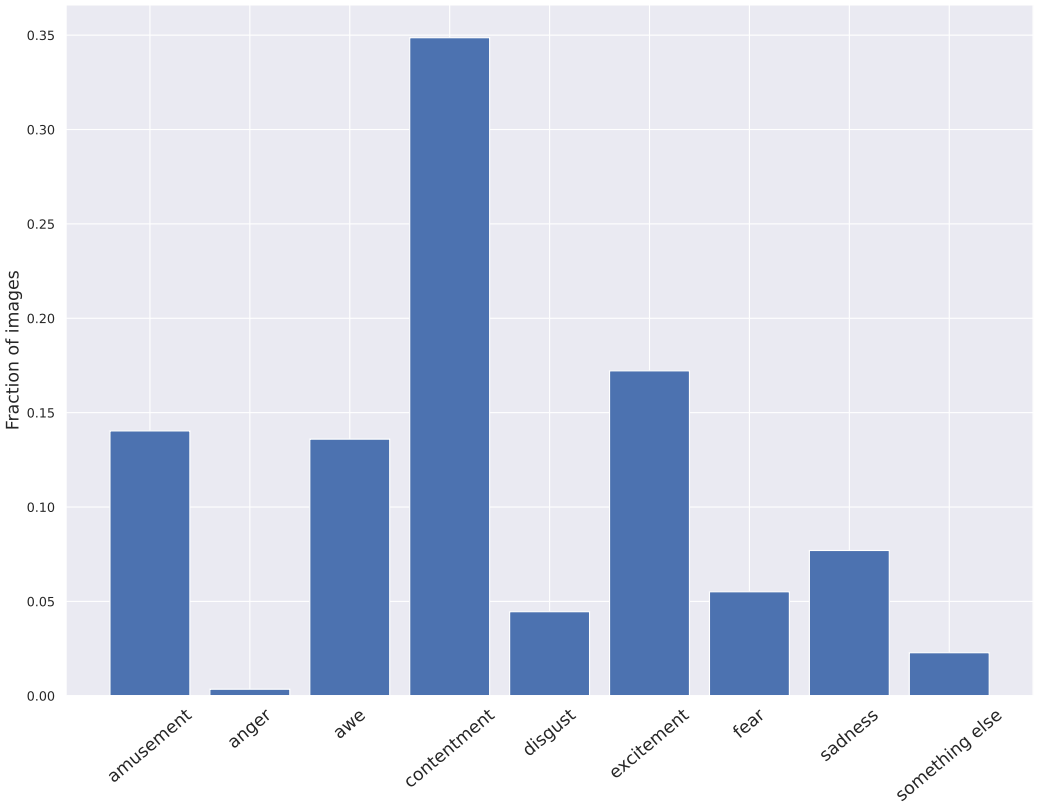

Figure 7: **Fraction of Affection images that have a unique strong majority w.r.t. the dominant emotions indicated by Affections' annotators, per each emotion class**.

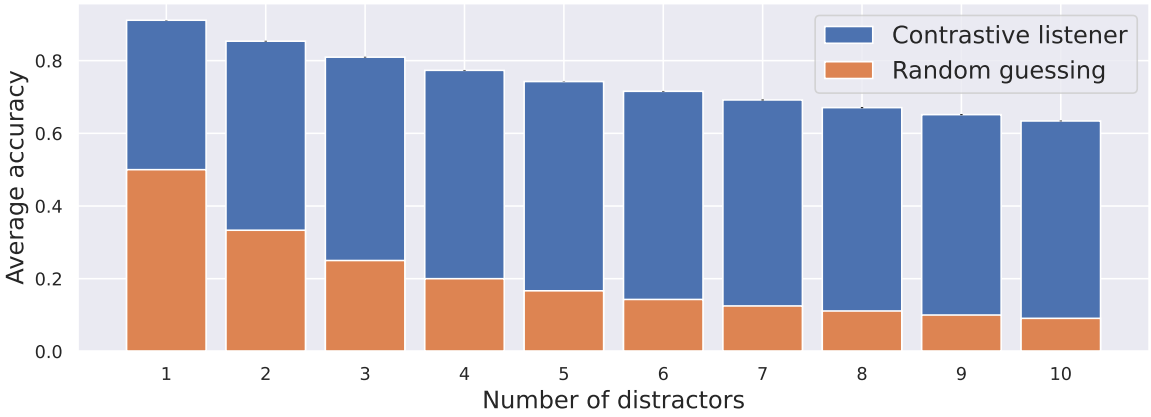

Figure 8: **Listening accuracy of an LSTM + ResNet-101 neural listener, trained with Affection captions**. The performance displayed is a function of the number of distractor images used at inference time and is the average resulting from five random seeds, used when pairing the target with randomly selected distractor images. Random guessing reflects performance when selecting the target uniformly at random. As expected, our neural listener fares significantly better, than random guessing, and also decreases its performance when more distractor images are considered.

Instructions

STEP 1: Look at the image and carefully read the two sentences.

then...
STEP 2: Decide which (if any) of the sentences could have been made **by a human**.

IMPORTANT.

1. The sentences are supposed to be **explanations** about **WHY** a human possibly felt, *or not*, any emotion upon seeing this image.
2. Sometimes humans *or* computers state explicitly the emotion they felt (e.g., happiness), but often, they do not!
3. Sometimes **BOTH** utterances will be from humans or computers.
4. **IGNORE** the **spelling** of the sentences and the lack of *punctuation*.
5. **IGNORE** **any** decision you made about **previously shown** image-sentence pairs in previously solved HITs **you did** for this task. I.e., **treat each HIT independently!**
6. **DO NOT submit more than ~50 HITS.**

Utterance A: this is a spectacular old building that i would love to explore

Utterance B: the architecture of the building is very intricate and detailed

Which utterance(s) were could be made by human(s), to justify a *possible* emotional reaction for the shown image?

- ○ Utterance **A**
- ○ Utterance **B**
- ○ **Both** could be made by humans
- ○ **None** is made by humans (both are made by a computer)

Figure 9: **User interface of emotional Turing test.** Upon reading the instructions (top) and observing the underlying image, each annotator had to select among the four options shown (bottom). In this example, the second utterance (B) is made by a neural speaker, while an annotator of Affection created the first utterance (A).

References

[AOGT22] Panos Achlioptas, Maks Ovsjanikov, Leonidas Guibas, and Sergey Tulyakov. Affection: Learning affective explanations for real-world visual data. *Computing Research Repository (CoRR)*, abs/2210.01946, 2022.

[AOH+21] 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.

[CFL+15] 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.

[DDS+09] 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.

[HG14] 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.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computing Research Repository (CoRR)*, abs/1512.03385, 2015.

[KKKL18] 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.

[KZG+17] 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.

[Lop22] A. Lopez. *Fdupes is a program for identifying or deleting duplicate files residing within specified directories.*, (accessed July 2022). Available at https://github.com/adrianlopezroche/fdupes.

[PWC+15] 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.

[QJHJ16] 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.

[RKH+21] 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.