Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content?

Patrick Schramowski
schramowski@cs.tu-darmstadt.de
Technical University Darmstadt,
Hessian Center for AI
Darmstadt, Germany

Christopher Tauchmann
tauchmann@cs.tu-darmstadt.de
Technical University Darmstadt,
Hessian Center for AI
Darmstadt, Germany

Kristian Kersting
kersting@cs.tu-darmstadt.de
Technical University Darmstadt,
Centre for Cognitive Science
Darmstadt, Hessian Center for AI
Darmstadt, Germany

ABSTRACT

This paper contains images and descriptions that are offensive in nature.

Large datasets underlying much of current machine learning raise serious issues concerning inappropriate content such as offensive, insulting, threatening, or might otherwise cause anxiety. This calls for increased dataset documentation, e.g., using datasheets. They, among other topics, encourage to reflect on the composition of the datasets. So far, this documentation, however, is done manually and therefore can be tedious and error-prone, especially for large image datasets. Here we ask the arguably "circular" question of whether a machine can help us reflect on inappropriate content, answering Question 16 in Datasheets. To this end, we propose to use the information stored in pre-trained transformer models to assist us in the documentation process. Specifically, prompt-tuning based on a dataset of socio-moral values steers CLIP to identify potentially inappropriate content, therefore reducing human labor. We then document the inappropriate images found using word clouds, based on captions generated using a vision-language model. The documentations of two popular, large-scale computer vision datasets—ImageNet and OpenImages—produced this way suggest that machines can indeed help dataset creators to answer Question 16 on inappropriate image content.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; *Machine learning*.

KEYWORDS

Datasets, Dataset documentation, Datasheets, Dataset curation

ACM Reference Format:
Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. 2022. Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content?. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533192

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533192

1 INTRODUCTION

Transfer learning from models that have been pre-trained on huge datasets has become standard practice in many computer vision (CV) and natural language processing (NLP) tasks and applications. While approaches like semi-supervised sequence learning [Dai and Le 2015] and datasets such as ImageNet [Deng et al. 2009]—especially the ImageNet-ILSVRC-2012 dataset with 1.2 million images— established pre-training approaches, the training data size increased rapidly to billions of training examples [Brown et al. 2020; Jia et al. 2021], steadily improving the capabilities of deep models. Recent transformer architectures with different objectives such as autoregressive [Radford et al. 2019] and masked [Devlin et al. 2019] language modeling as well as natural language guided vision models [Radford et al. 2021] for multi-modal vision-language (VL) modeling have even enabled zero-shot transfer to downstream tasks, avoiding the need for task-specific fine-tuning.

However, in all areas, the training data in the form of large and undercurated, internet-based datasets is problematic involving, e.g., stereotypical and derogatory associations [Bender et al. 2021; Gebru et al. 2021]. Along this line, Gebru et al. [2021] described dominant and hegemonic views, which further harm marginalized populations, urging researchers and dataset creators to invest significant resources towards dataset curation and documentation. Consequently, the creation of datasheets became common practice when novel datasets such as [Desai et al. 2021] were introduced. However, the documentation of Desai et al. [2021] shows that careful manual documentation is difficult, if not even unfeasible, due to the immense size of current datasets: '*We manually checked 50K* [out of 12M] *random images in RedCaps and found one image containing nudity (exposed buttocks; no identifiable face)*'. Also, in the process of creating a datasheet for the BookCorpus, Bandy and Vincent [2021] stated that further research is necessary to explore the detection of potential inappropriate concepts in text data. Birhane and Prabhu [2021] manually checked for and found misogynistic and pornographic in several common CV datasets. However, misogynistic images and pornographic content are only part of the broader concept of inappropriate content. It remains challenging to identify concepts such as general offensiveness in images, including abusive, indecent, obscene, or menacing content.

To make a step towards meeting the challenge, the present work proposes a semi-automatic method, called Q16, to document inappropriate image content. We use the VL model CLIP [Radford et al. 2021] to show that it is indeed possible to (1) steer pre-trained models towards identifying inappropriate content as well as (2)

the pre-trained models themselves towards mitigating the associated risks. In the Q16 setup, prompt-tuning steers CLIP to detect inappropriateness in images. Additionally, Q16 employs the recent autoregressive caption generation model MAGMA [Eichenberg et al. 2021] to provide accessible documentation. Thus, Q16 assists dataset documentation and curation by answering Question 16 of [Gebru et al. 2021], which also explains its name: *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?*

We illustrate Q16 on the popular ImageNet-ILSVRC-2012 [Deng et al. 2009] and OpenImages [Kuznetsova et al. 2020] dataset and show that large computer vision datasets contain additional inappropriate content, which previous documentations, such as [Birhane and Prabhu 2021], had not detected, cf. Fig. 1. In contrast to images identified in previous approaches, e.g., images showing nudity and misogynistic images (blue), Q16 detects a larger and broader range of potential inappropriate images (red). These images show violence, misogyny, and otherwise offensive material. Importantly, this includes images portraying persons (dark gray) as well as objects, symbols, and text.

The rest of the paper is organized as follows. We start off with a brief overview of related work and required background introducing pre-trained models and their successes as well as concerns raised. Next, we describe inappropriate image content and show that common deep models cannot reliably detect potential inappropriate images due to the lack of sufficient data. We then continue by demonstrating that recent models, guided by natural language during the pre-training phase, can classify and describe inappropriate material based on their retained knowledge. Before concluding, we present our automated dataset documentation exemplary on the ImageNet-ILSVRC-2012 and OpenImagesV6 datasets. We provide our models and the necessary data and code to reproduce our experiments and utilize our proposed method.[1]

2 BACKGROUND AND RELATED WORK

In this section, we describe pre-trained models in NLP, CV, and recent VL models. Furthermore, we touch upon related work aiming to improve dataset documentation and curation as well as identifying problematic content in datasets.

2.1 Large pre-trained models

Large-scale transformer-based language models revolutionized many NLP tasks [Lin et al. 2021]. As large, pre-trained models form the backbone of both natural language processing and computer vision today, it is natural that multimodal vision-language models [Jia et al. 2021; Radford et al. 2021; Ramesh et al. 2021] extend these lines of research.

For their CLIP model, Radford et al. [2021] collected over 400M image-text pairs (WebImageText dataset) to show that the success in large-scale transformer models in NLP can be transferred to vision and multimodal settings. One major takeaway from their work is the benefit of jointly training an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. Typical vision models [He et al. 2016; Tan and Le 2019] jointly train an image feature extractor and a classifier. Radford et al. [2021], the authors of CLIP, proposed to synthesize the learned text encoder with a (zero-shot) linear classifier at test time by embedding the names or descriptions of the target dataset's classes, e.g. "The image shows <*label*>.", thus reducing the (computational) cost of fine-tuning the model and using it as it was trained. Such models and their zero-shot capabilities display significant promise for widely-applicable tasks like image retrieval or search. The relative ease of steering CLIP toward various applications with little or no additional data or training unlocks novel applications that were difficult to solve with previous methods, e.g., as we show, classify potential inappropriate image content.

2.2 Issues arising from large datasets

Large-scale models require a tremendous amount of training data. The most recent and successful models, such as GPT-3 [Brown et al. 2020], CLIP [Radford et al. 2021], DALL-E [Ramesh et al. 2021] and other similar models, are trained on data scraped from the web, e.g. using CommonCrawl. The information they acquire from this data is largely uncontrolled. However, even ImageNet [Deng et al. 2009], which was released in 2012 and remains one of the most popular datasets in the computer vision domain to this day [Brock et al. 2021; Tan and Le 2021], contains questionable content [Birhane and Prabhu 2021]. The entailed issues have been discussed for language models, for instance, models producing stereotypical and derogatory content [Bender et al. 2021], and for vision model respectively CV datasets highlighting, e.g., gender and racial biases [Denton et al. 2021; Larrazabal et al. 2020; Steed and Caliskan 2021; Wang et al. 2020].

Consequently, Gebru et al. [2021] urged the creation of datasheets accompanying the introduction of novel datasets including a variety of information on the dataset to increase transparency and accountability within the ML community, and most importantly, help researchers and practitioners to select more appropriate datasets for their tasks. The documentation and curation of datasets have become a very active research area, and along with it, the detection of inappropriate material contained in datasets and reflected by deep models.

Dodge et al. [2021] documented the very large C4 corpus with features such as 'text source' and 'content', arguing for different levels of documentation. They also address how C4 was created and show that this process removed texts from and about minorities. A vast body of work to date that describes methodologies to tackle, abusive, offensive, hateful [Glavaš et al. 2020], toxic [Han and Tsvetkov 2020], stereotypical [Nadeem et al. 2021] or otherwise biased content [Dhamala et al. 2021] come from NLP. For several years, workshops on language[2] and offensive[3] language are carried out, producing evaluation datasets. Furthermore, Google hosts an API for the automatic detection of toxicity[4] in language, and research introduced toxicity benchmarks for generative text models [Gehman et al. 2020]. Additionally, the definitions and datasets on such tasks as bias- and hate-speech identification become increasingly complex [Sap et al. 2020]. Accordingly, most of the research on automatic methods focuses solely on text.

[1] https://github.com/ml-research/Q16

[2] https://aclanthology.org/volumes/W17-30/

[3] https://sites.google.com/site/offensevalsharedtask/home

[4] https://www.perspectiveapi.com/

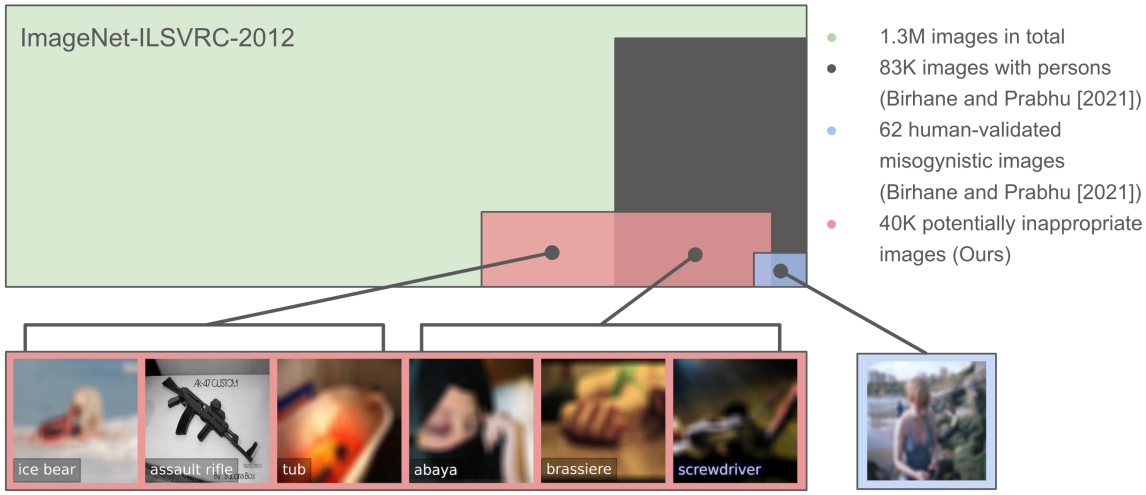


Figure 1: Range of identified inappropriate concepts illustrated using ImageNet (green). The other colors refer to different data-subsets: a selection of all images displaying persons (dark gray), potentially inappropriate images identified by our approach (red), and human-validated inappropriate (misogynistic) images identified in the study of Birhane and Prabhu [2021] (blue). The detected images in our approach partly overlap with the one in blue. Sizes are only illustrative, and actual numbers are given in the legend (right). Due to their apparent offensive content, we blurred the images.

With the present study, we aim to push the development of methods for the CV domain. Yang et al. [2020] argued towards fairer datasets and filter parts of ImageNet. Specifically, they see issues in ImageNet's concept vocabulary based on WordNet and include images for all concept categories (some hard to visualize). Furthermore, the inequality of representation (such as gender and race) in the images that illustrate these concepts is problematic. Birhane and Prabhu [2021] provided modules to detect faces and post-process them to provide privacy, as well as a pornographic content classifier to remove inappropriate images. Furthermore, they conducted a hand-surveyed image selection to identify misogynistic images in the ImageNet-ILSVRC-2012 (ImageNet1k) dataset. Gandhi et al. [2020] aimed to detect offensive product content using machine learning; however, they have described the lack of adequate training data. Recently, Nichol et al. [2021] applied CLIP to filter images of violent objects but also images portraying people and faces.

2.3 Retained knowledge of large models

Besides the performance gains, large-scale models show surprisingly strong abilities to recall factual knowledge from the training data [Petroni et al. 2019]. For example, Roberts et al. [2020] showed large-scale pre-trained language models' capability to store and retrieve knowledge scales with model size. Schick et al. [2021] demonstrated that language models can self-debias the text they produce, specifically regarding toxic output. Similar to our work, they prompt a model. However, they use templates with questions in the form of "this model contains <MASK>", where the gap is filled with attributes, such as toxicity, whereas we automatically learn prompts. Furthermore, Jentzsch et al. [2019] and Schramowski et al. [2020] showed that the retained knowledge of such models carries information about moral norms aligning with the human sense of *"right"* and *"wrong"* expressed in language. Similar to [Schick et al. 2021], Schramowski et al. [2022] demonstrated how to utilize this knowledge to guide autoregressive language models' text generation to prevent their toxic degeneration.

3 THE Q16 PIPELINE FOR DATASHEETS

Let us now start to introduce our semi-automatic method to document inappropriate image content. To this end, we first clarify the term "inappropriateness" in Sec. 3.1. Then we present and evaluate models, including our approach (illustrated in Fig 2a), to classify inappropriate image content. Specifically, Fig 2a shows a dataset representing socio-moral norms, which will be detailed in Sec. 3.2, steering CLIP to detect inappropriate content using (soft) prompt-tuning (cf. Sec. 3.3).

Lastly, in Sec. 3.4, we present the two-step semi-automated documentation (cf. Fig 2b). Notably, both steps include human interaction. First, CLIP and the learned prompts from Fig 2a are used to detect inappropriate images within the dataset. Detection is conservative, aiming to identify all potentially inappropriate content. Accordingly, the subsets are of considerable size, e.g., 40K in the case of ImageNet1k. Therefore, the second step generates automatic image descriptions to assist the dataset creators in describing and validating the identified content. The final documentation of Q16 includes the ratio of identified images and the total amount of samples, and a summary of the identified concepts. To overview the contained concepts in an easily accessible way, we generate word clouds based on two properties: the dataset annotation and generated description.

3.1 Inappropriate image content.

Let us start off by clarifying the way we use the term "inappropriate" in our work and describing the term in the context of images.

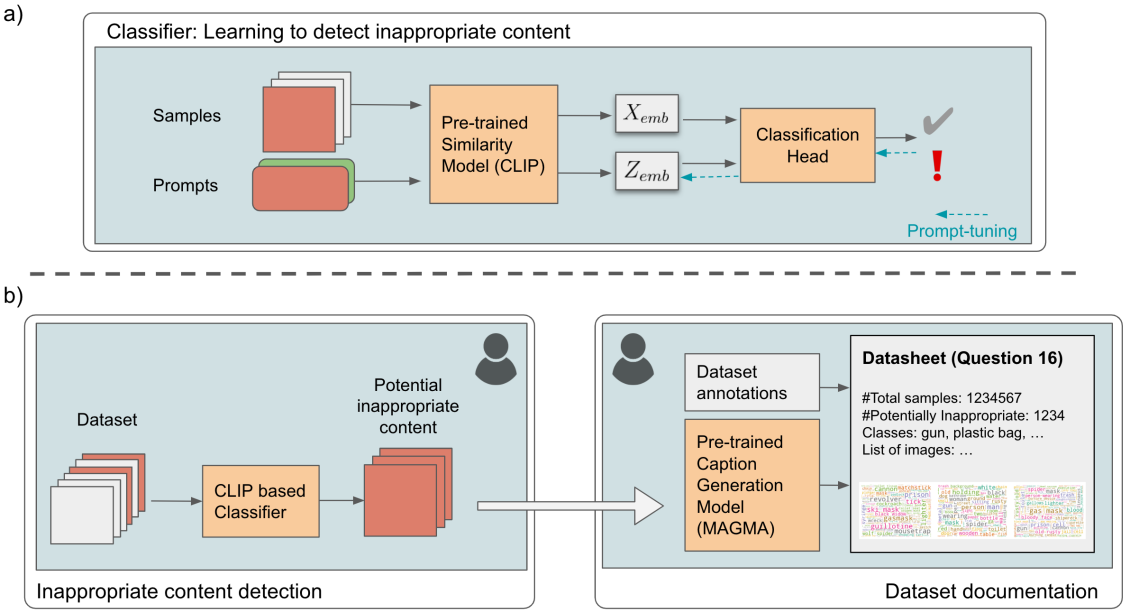


Figure 2: Overview of the Q16 pipeline, a two-step dataset documentation approach. a) In order to utilize the implicit knowledge of the large pre-trained models, prompt-tuning steers CLIP to classify inappropriate image content. b) Dataset documentation process: First, a subset with potentially inappropriate content is identified. Secondly, these images are documented by, if available, image annotations and automatically generated image descriptions. Both steps are designed for human interaction.

Question 16 of Datasheets for Datasets [Gebru et al. 2021] asks one to document the dataset composition regarding the contained "data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety". Consequently, Birhane and Prabhu [2021] applied different models to detect visible faces (thus violating privacy rights) and pornographic content. Additionally, they conducted a survey identifying misogynistic images. However, the definition of Gebru et al. [2021] includes a broader range of inappropriate concepts not addressed by current work.

According to the Cambridge dictionary[5], 'offending' can be phrased as 'unwanted, often because unpleasant and causing problems'. Additionally, in the context of images and text, according to Law Insider[6]: '*Offending Materials means any material, data, images, or information which is (a) in breach of any law, regulation, code of practice or acceptable use policy; or (b) defamatory, false, inaccurate, abusive, indecent, obscene or menacing or otherwise offensive; or (c) in breach of confidence, copyright or other intellectual property rights, privacy or any other right of any third party.*' In the present paper, we focus on images following the definition (b). This definition aligns with definitions of previous work detecting hate speech [Gomez et al. 2020] and offensive product images [Gandhi et al. 2020].

Note that inappropriateness, especially offensiveness, is a concept that is based on social norms, and people have diverse sentiments. In the present study, we detect inappropriate content based on the implicit knowledge contained in CLIP steered with selected data (described in the following section). Therefore, the investigated 'inappropriateness' may primarily surface from the group of people that have generated the selected data and the annotators but also the pre-trained model's retained knowledge.

[5]https://dictionary.cambridge.org/dictionary/english/offending
[6]https://www.lawinsider.com/dictionary/offending-materials

3.2 The Socio-Moral Image Database (SMID)

Besides utilizing the 'knowledge' of pre-trained models on inappropriate concepts, we further steer the model towards detecting (morally) inappropriate image concepts indirectly via training examples. I.e. we aim to find a compass guiding the encoded knowledge of CLIP and by that be able to classify inappropriate content beyond the examples shown in the steering dataset. To this end, we propose to use the Socio-Moral Image Database (SMID) [Crone et al. 2018] together with the few-shot capabilities of CLIP. This dataset will not only be used to steer CLIP but also to evaluate the the classifier's performance in the following sections.

The SMID dataset contains 2,941 images covering both morally positive and negative poles (962 negative images and 712 positive images) over several content dimensions, including objects, symbols as well as actions. Stimuli span the entire moral spectrum ranging from positive to negative, see Appendix Sec. A for more details. In total, over 50 concepts are included, with negative ones such as *Harm, Inequality, Degradation, Discrimination, and Exploitation.* The complete list is provided in Table 2 of [Crone et al. 2018].

The images were collected in a multi-step process and annotated by 2,716 annotators. Crone et al. [2018] suggested to divide the data into *good* (mean rating > 3.5), *bad* (mean rating < 2.5), and neutral (rest) images. According to this division we considered a rating < 2.5 as (morally) inappropriate, and rating > 3.5 as counterexamples.

Architecture	Pre-training dataset	Accuracy (%)	Precision	Recall	F1-Score
ResNet50	ImageNet1k	78.36 ± 1.76	0.75 ± 0.05	0.74 ± 0.09	0.76 ± 0.02
		80.81 ± 2.95	0.75 ± 0.02	0.81 ± 0.02	0.80 ± 0.03
	ImageNet21k	82.11 ± 1.94	0.78 ± 0.02	0.80 ± 0.05	0.78 ± 0.04
		84.99 ± 1.95	0.82 ± 0.01	0.85 ± 0.06	0.82 ± 0.04
	WebImageText	∘90.57 ± 1.82	∘0.91 ± 0.03	∘0.89 ± 0.01	∘0.88 ± 0.03
ViT-B/32	WebImageText	94.52 ± 2.10	0.94 ± 0.04	0.91 ± 0.02	0.92 ± 0.01
ViT-B/16	WebImageText	**•96.30 ± 1.09**	**•0.95 ± 0.02**	**•0.97 ± 0.01**	**•0.97 ± 0.02**

Table 1: Performances of pre-trained models ResNet50 and ViT-B. CLIP-based models trained on WebImageText outperform baselines and show remarkable performance in identifying the inappropriate content contained in SMID. The ResNet50 is pre-trained on ImageNet1k, ImageNet21k [Deng et al. 2009] and the WebTextImage dataset [Radford et al. 2021]. The ViT is pre-trained on the WebTextImage dataset. On the ImageNet datasets, we applied linear probing (top) and fine-tuning (bottom), and on the WebImageText-based models, soft-prompt tuning. The overall best results are highlighted bold with the • marker and best on the ResNet50 architecture with ∘ markers. Mean values and standard deviations are reported.

3.3 Inappropriate content detection of Q16

Let us now move on to presenting and evaluating different models, including our CLIP-based Q16 approach, for the task at hand to classify inappropriate image content. Here inappropriate content is defined by the SMID data and annotation, see section above. In the following experiments, 10-fold cross-validated results are reported.

3.3.1 Deep Learning baselines. As baselines we fine-tuned two standard pre-trained CV models (PMs) to investigate how well deep neural networks can identify inappropriate content. Similar to Gandhi et al. [2020], we used the ResNet50 architecture [He et al. 2016], pre-trained on ImageNet datasets [Deng et al. 2009].

Tab. 1 shows the performance of both the fine-tuned model (training all model parameters) and a model with only one linear probing layer. In our work, the probing layer refers to adding one final classification layer to the model. This part of the table shows inconclusive results: even if the performance increases when a larger dataset (ImageNet21k) is used. After fine-tuning the whole model, recall increases; precision, however, is still comparatively low. Specifically, the resulting low precision and low recall of the linear probed ImagNet1k-based models show problems classifying truly inappropriate images as well as distinguishing between truly non-inappropriate and inappropriate images. We will use these models as baselines to investigate if more advanced PMs (trained on larger unfiltered datasets) carry information about potential inappropriate image content.

3.3.2 Zero and few-shot capabilities of CLIP to infer inappropriate content. Next, we will investigate if CLIP's contrastive pre-training step contains image-text pairs that equip the model with a notion of inappropriate concepts.

Due to the natural language supervision, CLIP should implicitly have acquired knowledge about what a human could —depending on the context— perceive as inappropriate content. We confirmed this assumption with a PCA of the embedded images (see Appendix Sec. B for details)

Now, the inappropriateness classifier of our approach (Fig. 2a) utilizes this 'knowledge'. It is based on prompting CLIP with a natural language sentence. Our prompts have the form "This image is about something <*label*>.", helping to specify that the text is actually about the content of the image. To map the labels of the SMID dataset to natural language sentences, we used the following labels according to Crone et al. [2018]: *bad/good behavior*, *blameworthy/praiseworthy*, *positive/negative* and *moral/immoral*. The *positive* and *negative* labels resulted in the best zero-shot performance. Images are encoded via the pre-trained visual encoder, similar to the ResNet50 model. However, instead of training a linear classifier to obtain class predictions as in these models, we now operate on the similarity of samples (the cosine similarity) in the representation space:

$$Sim(x, z) = \frac{E_{visual}(x) * E_{text}(z)}{||E_{visual}(x)||_2 * ||E_{text}(z)||_2}\,, \tag{1}$$

where E_{visual} and E_{text} are the visual and text encoders, x is an image sample and z a prompt. Fig. 3 (0%, prompt-tuning) shows that this approach already performs on par with the ImageNet-based PMs fine-tuned on SMID (100%, linear probing). However, the zero-shot approach can classify true-negative samples well but performs not so well on classifying positives. This observation suggests that both prompts, at least the one corresponding to the positive class label, are not optimal.

3.3.3 Steering CLIP to infer inappropriate content via prompt-tuning. The manual handwritten prompts may not be the best way to query the model. Consequently, we used prompt-tuning [Hambardzumyan et al. 2021; Lester et al. 2021; Qin and Eisner 2021] to learn continuous optimal prompts. Prompt-tuning optimizes the prompts by searching for the optimal text embeddings for a given class label.

Several variations employ prompt-tuning: Prefix-tuning, for example, learns a prefix to add to a sample's embedding [Qin and Eisner 2021] on every model layer. Lester et al. [2021] created new (prompt) embeddings only once by pre-pending a small vector to the original input embedding for all downstream examples. Hambardzumyan et al. [2021] updated both the input and final embeddings once. In contrast, we propose to learn the entire final sentence embedding once, obtaining one sentence embedding, z_{emb}, for each class label. In turn, the distinction of inappropriate and other images is defined as an optimization task using gradient descent as

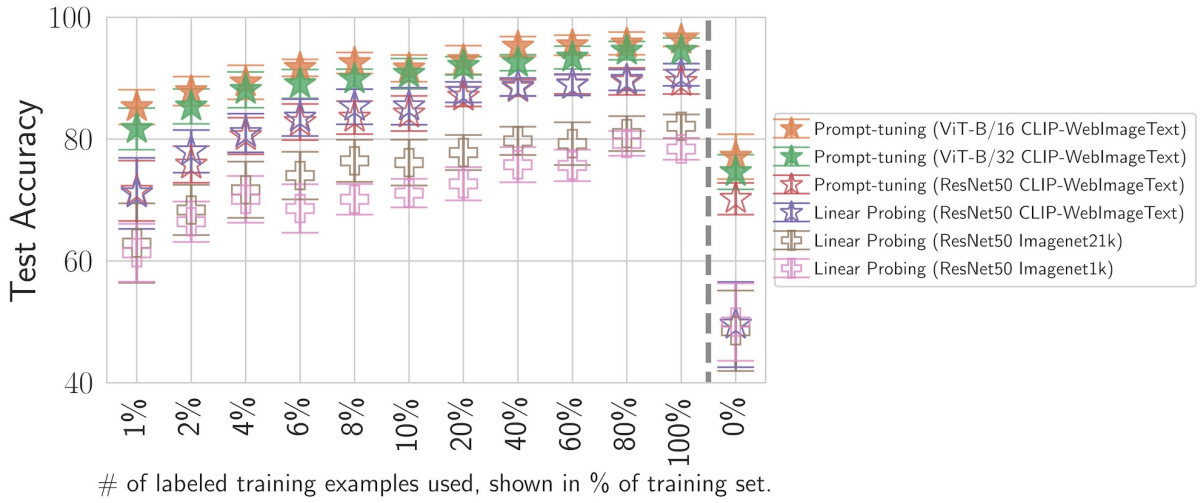


Figure 3: Performance of pre-trained models ResNet50 and ViT-B. CLIP-based models outperform baselines and show remarkable (zero- and few-shot) performances in identifying the inappropriate content contained in SMID. ResNet50 is pre-trained on ImageNet1k, ImageNet21k [Deng et al. 2009] and the WebTextImage dataset [Radford et al. 2021]. ViT is pre-trained on the WebTextImage dataset. On the ImageNet datasets, we applied linear probing (top), and on the WebImageText-based models used soft-prompt tuning. Tuning was performed on different sizes of the SMID training set where 100% corresponds to 1,506 images.

follows:

$$\hat{\mathbf{z}}_{emb} = \arg\max_{\mathbf{z}_{emb}} \{L(\mathbf{z}_{emb})\}\,, \tag{2}$$

where

$$L(\mathbf{z}_{emb}) = -\frac{1}{|X|} \sum_{\mathbf{x} \in X} \mathbf{y}\ \log(\hat{\mathbf{y}}), \tag{3}$$

$$\text{with } \hat{\mathbf{y}} = \text{softmax}(Sim(\mathbf{x}, \mathbf{z}_{emb}))\,. \tag{4}$$

Here, the parameters θ of E_{visual} and E_{text} are not updated. The initial prompts Z are only propagated through E_{text} once and the resulting embeddings $\mathbf{z}_{emb} \in Z_{emb}$ are optimized. Furthermore, $\mathbf{y}$ is the class label, and X a batch in the stochastic gradient descent optimization. Our prompt-tuning approach is summarized visually in Fig. 2; further details on applying it to the SMID dataset can be found in Appendix Sec. C.

Fig. 3 also shows an evaluation of CLIP using the soft prompts (prompt-tuning). We can see that a small portion of the training data (e.g., 4%, 60 images) already leads to an increase of the vision transformer's (ViT-B) performance to over 90%. This shows that indeed large pre-trained model can be steered more efficient, i.e. generalize and detect inappropriate concepts beyond the training samples. In general, the ViT-B outperforms the pre-trained ResNet50 models. Furthermore, ViT-B/16 outperforms the ViT-B/32, indicating that not only the dataset's size is important, but also the capacity of the model (ViT-B/16 has higher hidden-state resolution than the ViT-B/32). Training ViT with the full training set results in $96.30\% \pm 1.09$ (cf. Tab. 1) accuracy.

Overall, one can see that steering CLIP towards inferring potentially inappropriate concepts in images requires only little additional data. In contrast to other pre-trained models, it provides a reliable method to detect inappropriate images.

3.4 Dataset documentation of Q16

Using our prompt-tuning approach, the pre-selection by CLIP can, in principle, extract possible inappropriate images automatically that can then be used for dataset documentation. However, we have to be careful since inappropriateness is subjective to the user—e.g., researchers and practitioners selecting the dataset for their tasks—and, importantly, to the task at hand. In our case, the steered model may primarily mirror the moral compass and social expectations of the 2,716 annotators. Therefore, it is required that humans and machines interact with each other, and the user can select the images based on their given setting and requirements. Hence, we do not advise removing specific images but investigating the range of examples and inappropriate content selected by the model and thereby documenting the dataset. In the following, we present our approach to assist data creators not only in identifying but also describing the identified potential inappropriate content.

Fig. 2b shows our human-in-the-loop, (*cf.* Sec. 3.4.1), documentation setting. The above mentioned detection is conservative, aiming to identify all potentially inappropriate content. Accordingly, the subsets are of considerable size, e.g., 40K in the case of ImageNet1k. Therefore, the first step to assist the dataset creators in describing and validating the identified content is the automatic generation of image descriptions, *cf.* Sec. 3.4.2. To overview the contained concepts in an easily accessible way, we generate word clouds based on two properties: the dataset annotation and generated description, *cf.* Sec. 3.4.3.

3.4.1 Answering Datasheet Question 16: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? As intended by the original datasheets paper [Gebru et al. 2021], dataset creators should start

describing the curation process concerning this question. Whereas our approach could also be used for the curation, we focus solely on documenting the final dataset content to mitigate unwanted societal biases in ML models, and help users select appropriate datasets for their chosen tasks.

The dataset documentation should contain the total amount of images and the ratio of identified, potentially inappropriate images. Since the process of creating a datasheet is not intended to be automated [Gebru et al. 2021]—however, the quality of current datasheets [Desai et al. 2021] indicate that a semi-automated method is unavoidable—, the resulting subset should be manually validated and described by the dataset's creators. Our approach aims to reduce impractical human labor while encouraging creators to reflect on the process carefully.

3.4.2 Automatic caption generation. In order to categorize and thus describe the identified content, dataset annotations can be used if they are available. However, these annotations often may not describe the complete image content, especially in the case of natural images. Therefore, we utilize automatic generation of image descriptions, cf. Fig. 2b (right). To this end, we propose to generate text using a caption-generation model. Specifically, we used MAGMA (Multimodal Augmentation of Generative Models) [Eichenberg et al. 2021]. MAGMA is a recent text generation model based on multimodal few-shot learners [Tsimpoukelli et al. 2021]. It uses both the CLIP and GPT-J [Wang and Komatsuzaki 2021] models and adds pre-training and fine-tuning steps on several datasets to generate image captions from image-text pairs. These captions are especially beneficial because they include the encyclopedic knowledge of GPT-J, and as such, knowledge on socio-moral norms (similar to the one we obtain from CLIP). Further, the multimodal input enables one to guide the resulting textual description. Since we aim to generate "neutral" image descriptions, we use the prompt *<A picture of>* and add the output of multiple generations to the image description. To sample from the model, we applied top-k filtering. In order to acquire greater variety in the descriptions, we used different temperature values.

3.4.3 Word cloud generation. Actually, Question 16 asks the dataset curator to be familiar with a broad range of inappropriate concepts. Whereas our Q16 approach already helps reduce the number of inappropriate images to be checked and, in turn, human labor, even the validation of the reduced set may still require a lot of manual effort. To provide a concise overview, we propose to compute word clouds to summarize the complex captions generated. More precisely, we present the identified, potentially inappropriate content within the dataset using three different kinds of word clouds from dataset annotations and generated textual image descriptions. All word clouds highlight words or bi-grams based on their frequency and rank.

The first word cloud requires existing dataset annotations, e.g., class labels, and provides first insights of identified concepts and could highlight sensible labels. The word cloud visualizes the information by highlighting the most-frequent annotations. However, note that dataset creators should also pay attention to infrequent occurrences indicating deviating concepts compared to other examples from, e.g., the same class. Many images with the same annotation could indicate a general negative association.

Following the same procedure, the second word cloud describes the identified set of images using the generated text and thus independent of the dataset annotations. Therefore, this word cloud potentially describes identified concepts not captured by the first word cloud.

For the third word cloud, we use a chi-squared weighting of the word/bi-gram frequencies to illustrate differences between the identified inappropriate image set and the remaining images; common text descriptions occurring in both sets are removed. Each word i is assigned the following weight:

$$weight_i = \frac{(observed_i - expected_i)^2}{expected_i}, \tag{5}$$

where $observed_i$ is the observed frequency of word i in the inappropriate subset and $expected_i$ the expected value, i.e., the observed word frequency describing the dataset's remaining samples. This word cloud highlights the conspicuous descriptions that can be traced back to the corresponding images.

It is noteworthy that these wordclouds highlight frequent concepts for documentation purpose. Thus, it may be easy to dismiss the severity of inappropriateness if the database contains less of that particular image content, e.g. some inappropriate content may be less common but more severe. Therefore, we advice dataset creators to also inspect infrequent concepts.

Finally, we would like to note that our pipeline also produces several statistics such as exact word frequencies and traceable image descriptions that we do not include directly in the datasheet. The dataset creators can provide this additional information as a supplement next to the identified image IDs.

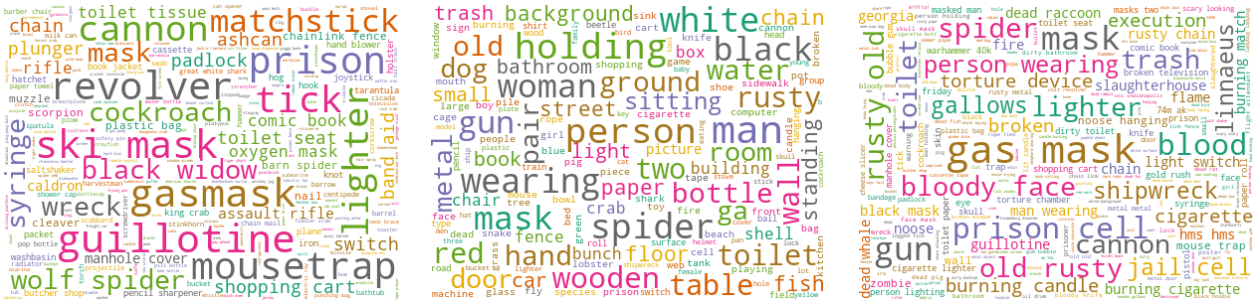
4 ANSWERING DATASHEET QUESTION 16 FOR IMAGENET AND OPENIMAGES

Now we have everything together to provide an exemplary datasheet documentation, here for the CV datasets ImageNet [Deng et al. 2009] and OpenImages [Kuznetsova et al. 2020]. To identify inappropriate content within the datasets, we used the public available ViT-B/16[7] variant of CLIP steered by SMID-based optimized prompts. We observed that shifting the negative threshold to a rating of 1.5 instead of 2.5 provides a conservative but reliable classifier; hence we determined the prompts with these corresponding few-shot examples. For the documentation process we utilized the ResNet50x16 MAGMA model and generated 10 captions ($k = 5$ using a temperature of $\tau = 0.1$ and $k = 5$ using $\tau = 0.4$) for each images. Additionally to the following documentations, we provide Python notebooks with the corresponding images along with the classifier in our public repository.[8]

4.1 ImageNet

We start with one of the most known CV datasets, ImageNet1k (ImageNet-ILSVRC2012). Additionally to the concise overview using word clouds (Fig. 4) we provide further detailed description (highlighting the class labels) on the identified inappropriate concepts, and blurred examples for illustration (Fig. 5). We separate the

[7] In our repository we default to the even larger ViT-L/14 variant which was released after submission of this manuscript.
[8] https://github.com/ml-research/Q16



(a) Most-frequent image annotations. (b) Most-frequent image descriptions. (c) Weighted image descriptions.

Figure 4: Word clouds documenting the potentially inappropriate image content of the ImageNet1k dataset. Image annotations are contained within the dataset. Image descriptions are automatically generated. Word size is proportional to the word counts and rank in the generated captions corresponding to the inappropriate image set.

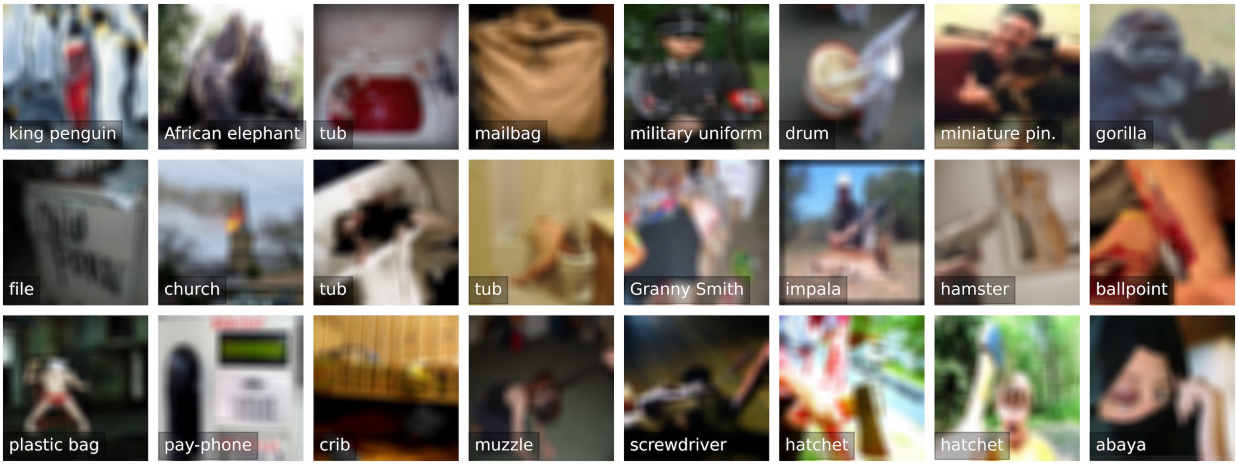


Figure 5: Exemplary images with inappropriate content from the pre-selection of our proposed method. The images visualize the range of concepts (objects, symbols, actions) detected. Due to their apparent offensive content, we blurred the images. Their content can be inferred from the main text.

identified content into potentially inappropriate objects, symbols, and actions due to the complexity of inappropriate context.

Objects. The ImageNet1k dataset, also known as ImageNet-ILSVRC-2012, formed the basis of task-1 of the ImageNet Large Scale Visual Recognition Challenge. Hence, all images (1,331,167) display animals or objects. To illustrate potential missing information in the dataset's annotations, we restricted ourselves not to include the hierarchical information contained in the synsets, cf. the first word cloud in Fig. 4a.

Therefore, it is not surprising that the largest portion of the potential inappropriate content concerns negative associated objects and animals. In total, 40,501 images were identified by the classifier, where the objects "gasmask" (797 images), "guillotine" (783), and "revolver" (725) are the top-3 classes. However, whereas most people would assign these objects as morally questionable and offensive, they may not be treated as inappropriate when training a general object classifier. The same applies to the animal-classes tick (554) and spider (397).

To detect more suspicious, inappropriate content, it may be more applicable to investigate classes with only a small portion of possible inappropriate images. Next to injured ("king penguin") and aggressive animals (e.g. "pembroke"), our proposed classifier detects caged (e.g. "great pyrenees", "cock") and dead animals (e.g. "squirrel monkey", "african elephant"). Additionally, objects in inappropriate, possible offensive scenes, like a bathtub tainted with blood ("tub") or a person murdered with a screwdriver ("screwdriver") are extracted, cf. also Fig. 5.

Symbols. Both the second (*person, woman, man*) and the third word cloud (*person wearing*) highlight that in many cases persons are subject to the inappropriate concepts identified. In the corresponding images, one is able to identify offensive symbols and text on objects: several National Socialist symbols especially swastika (e.g. "mailbag", "military uniform"), persons in Ku-Klux-Klan uniform (e.g. "drum"), insults by e.g. showing the middle finger (e.g. "miniature pinscher", "lotion"), cf. first row of Fig. 5. Furthermore,

we observed the occurrence of offensive text such as "child porn" ("file") and "bush=i***t f*** off USA" ("pay-phone").

Actions. The third word cloud further documents the identified concepts. Words like *blood, torture, execution* show that in addition to objects and symbols, our classifier interprets scenes in images and hence identifies offensive actions shown in images. Scenes such as burning buildings (e.g. "church") and catastrophic events (e.g. "airliner", "trailer truck") are identified. More importantly, inappropriate scenes with humans involved are extracted such as comatose persons (e.g. "apple", "brassiere", "tub"), persons involved in an accident (e.g. "mountain bike"), the act of hunting animals (e.g. "African elephant", "impala"), a terrifying person hiding under a children's crib ("crib"), scenes showing weapons or tools used to harm, torture and kill animals (e.g."hamster") and people (e.g. "hatchet", "screwdriver", "ballpoint", "tub").

Furthermore, derogative scenes portraying men and women wearing muzzles, masks, and plastic bags, clearly misogynistic images, e.g., harmed women wearing an abaya, but also general nudity with exposed genitals (e.g. "bookshop", "bikini", "swimming trunks") and clearly derogative nudity (e.g. "plastic bag") are automatically selected by our proposed method. Note that multiple misogynistic images, e.g., the image showing a harmed woman wearing an abaya, were not identified by the human hand surveyed image selection of Birhane and Prabhu [2021]. Therefore, we strongly advocate utilizing the implicit knowledge of large-scale state-of-the-art models in a human-in-the-loop curation process to not only partly automatize the process but also to reduce the susceptibility to errors.

4.2 OpenImages

Our next exemplary documentation is based on the dataset OpenImages [Kuznetsova et al. 2020]. Its first version [Krasin et al. 2016] was released in 2016, and the newest version 6 in 2020. The dataset contains 1.9M images with either single or multiple objects labeled, resulting in 59.9M image-level labels spanning 19,957 classes and 16M bounding boxes for 600 object classes. In contrast to the ImageNet documentation, we only provide the intended concise overview for Datasheet's Question 16. Thus refrain from showing exemplary images. However, after describing the content using the word clouds, we want to point out one extremely disturbing example.

We documented the training set of OpenImagesV6 (1,743,042 images) and identified a potentially inappropriate set of 43,395 images. Fig. 6 shows our computed word clouds. The first word cloud (Fig. 6a) shows that most identified images portray persons with labels like "human head", "human face", or "human body", showing both men and woman. The second word cloud (Fig. 6b) reflects this observation but additionally highlights the portray of, e.g., guns. It further shows that posters are displayed. We observed that often the corresponding images show pornographic material.

The third word cloud reveals more interesting concepts (Fig. 6c). We can again observe the descriptions *cartoon, poster* referring to potential disturbing art, but also graffiti with inappropriate text. Furthermore, the description *gun* is further highlighted. Human skulls and skeletons are displayed as well as dead and harmed animals (*dead mouse, dead bird*). Importantly, the descriptions *bloody face, blood, wound* refer to the concept of harm. It is noteworthy that, as the descriptions *zombie* and *zombie mask* could suggest, the corresponding images sometimes show costumes and makeup, however, also often real scenes. This observation demonstrates that human validation is necessary.

Dead bodies: Abu Ghraib torture and prisoner abuse. The image concepts described above need to be documented and could have an influence on users' opinion regarding the dataset selection. In contrast to these concepts the generated description *gallows, execution, person lying, dead bodies* (cf. Fig. 6c) extremely disturbed us while checking the corresponding images. Especially, we want to highlight one image we found (ID: 30ec50721c384003.jpg, looking at the picture could be disturbing). The image shows several scenes, also known as "Abu Ghraib torture and prisoner abuse", displaying members of the U.S. Army posing in front of dead bodies during the Iraq War. These scenes were classified as a series of human rights violations and war crimes. They show sexual abuse, torture, rape, sodomy, and the killing of Manadel al-Jamadi (clearly identifiable in the dataset's image). Note that this image is labeled ("person", "man", "clothing", "human face") and was annotated with bounding boxes, thus checked by human annotators. Besides documentation, our approach can also pre-flag such images as potentially inappropriate to validate them during annotation.

5 SOCIETAL IMPACT AND LIMITATIONS

Recent developments in large pertained models in NLP, such as GPT-3 have a far-reaching impact on society (300+ applications building on the model as of March 2021[9]), and we assume that the popularity of pre-trained CV, especially those including VL, models will follow along that path. So it is natural that the discussions surrounding ethical issues, such as biases, in NLP models transfer to VL models. Indeed, recently some awareness of problematic content in CV datasets arose; however, we are actually faced with broader issues in image datasets. Birhane and Prabhu [2021] described many negative societal implications underlying CV datasets' issues regarding, e.g., groups at margins such as high error rates for dark-skinned people in CV models recognizing pedestrians. Such issues even lead to the deletion of entire datasets.[10] These issues are likely to become even more prominent since VL models combining images and text will become more applicable in industry and, in turn, generate great impact on society.

Specifically, large datasets underlying much of current machine learning raise serious issues concerning inappropriate content such as offensive, insulting, threatening, or might otherwise cause anxiety. This calls for increased dataset documentation, e.g., using datasheets. They, among other topics, encourage to reflect on the composition of the datasets. So far, this documentation, however, is done manually and therefore can be tedious and error-prone, especially for large image datasets. Here we ask the arguably "circular" question of whether a machine can help us reflect on inappropriate content, answering Question 16 in Datasheets [Gebru et al. 2021]. To this end, we provide a method to automatically detect and describe inappropriate image content to assist documentation of datasets. Such automation might tempt dataset creators to neglect

[9] https://openai.com/blog/gpt-3-apps/
[10] https://venturebeat.com/2020/07/01/mit-takes-down-80-million-tiny-images-data-set-due-to-racist-and-offensive-content/

(a) Most-frequent image annotations. **(b) Most-frequent image descriptions.** **(c) Weighted image descriptions.**

Figure 6: Word clouds documenting the potentially inappropriate image content of the OpenImagesV6 dataset. Image annotations are contained within the dataset. Image descriptions are automatically generated. Word size is proportional to the word counts and rank in the generated captions corresponding to the inappropriate image set.

manual validation. However, it is of importance that humans stay in control, therefore, we strongly advise applying such methods in a human-in-the-loop setting as intended by Gebru et al. [2021] and described in our demonstrations.

There are natural limitations that should be addressed in future work. First, we chose a binary classification to detect general inappropriate content, then described using a text-generation model. Thus, extending previous categories into more fine-grained concepts could further improve transparency and documentation. We strongly advocate applying our documentation along with other methods, e.g., detecting faces and pornographic content [Birhane and Prabhu 2021]. Furthermore, while the SMID dataset with moral norms provides a good proxy for inappropriateness, developing novel CV datasets to drill down further on identifying inappropriateness and similar concepts would be very beneficial.

Moreover, whereas we evaluated our *inappropriateness classifier*, we did not evaluate our automatic generation of textual image descriptions summarizing the portrayed inappropriate concepts. Doing so provides an interesting avenue for future work. To ensure broad descriptions, we executed multiple generation iterations. Fine-tuning a caption generation model could lead to further improvements. Likewise, Radford et al. [2021] provided details about possible biases and other potential misuses of CLIP models, which could easily influence the detection as well as the description that we used. Generally, advances in bias-free models are very likely to also positively impact our Q16 approach.

Finally, like other social norms, inappropriate concepts, especially offensiveness, do evolve constantly. This evolution makes it necessary to update the data, system, and documentation over time. Furthermore, an important avenue for future work is addressing what different groups of society, e.g., different cultures, would consider inappropriate. Here, we just relied on the ones averaged by the SMID dataset, where, e.g., the 476 participants for image collection task (half female) were mainly from the United States and partly (10%) recruited from India. Further, it is to be expected that in specific cases, annotators disagree. This issue could be tackled by a multi-annotator architecture [Davani et al. 2022] that captures the differences between annotators' perspectives. Thus provide better estimates for uncertainty in predictions and, in turn, better indicate a manual review for specific detected concepts.

6 CONCLUSION

Deep learning models trained on large-scale image datasets have become standard practice for many applications. Unfortunately, they are unlikely to perform well if their deployment contexts do not match their training or evaluation datasets or if the images reflect unwanted behavior. To assist humans in the dataset curation process, particularly when facing millions of images, we propose Q16, a novel approach utilizing the implicit knowledge of large-scale pre-trained models and illustrated its benefits. Specifically, we argued that CLIP retains the required 'knowledge' about what a human would consider offending during its pre-training phase and, in turn, requires only few shots to automatically identify offensive material. On two canonical large scale image datasets, ImageNet-ILSVRC2012 and OpenImages, we demonstrate that the resulting approach, called Q16, can indeed identify inappropriate content, actually broader than previous, manual studies.

Q16 provides several interesting avenues for future work. First, one should investigate other computer vision as well as multi-modal datasets. One should also extend Q16 to multi-label classification, directly separate offensive objects, symbols, actions, and other categories of inappropriate content at once. Moreover, moving beyond binary classification towards gradual levels of inappropriateness may result in more fine-grained details in the datasheets. Finally, the underlying deep models are black-boxes, making it hard to understand why specific images are identified and described as inappropriate. Combining Q16 with explainable AI methods such as [Chefer et al. 2021] to explain the reasons is likely to improve the datasheet.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their valuable feedback. Furthermore, the authors are thankful to Aleph Alpha for providing access to the image-captioning model MAGMA. This research has benefited from the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK) cluster project "The Third Wave of AI".

REFERENCES

Jack Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus. In *Proceedings of NeurIPS Datasets and Benchmarks*. 1–13.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 610–623.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546.

Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. 2021. High-Performance Large-Scale Image Recognition Without Normalization. In *Proceedings of the 38th International Conference on Machine Learning, (ICML)*. 1059–1071.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. 1–25.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer Interpretability Beyond Attention Visualization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 782–791.

Damien L. Crone, Stefan Bode, Carsten Murawski, and Simon M. Laham. 2018. The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE* 13, 1 (01 2018), 1–34.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised Sequence Learning. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. 3079–3087.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics (TACL)* 10 (01 2022), 92–110.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (2021), 1–14.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. In *Proceedings of NeurIPS Datasets and Benchmarks (NeurIPS)*. 1–13.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of Conference on Fairness, Accountability, and Transparency (FAccT)*. 862–872.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1286–1305.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2021. MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning. arXiv preprint arXiv:2112.05253. (2021).

Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2020. Scalable Detection of Offensive and Non-compliant Content / Logo in Product Images. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2236–2245.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*. 3356–3369.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 6350–6365.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1459–1467.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. arXiv preprint arXiv:2101.00121. (2021).

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying Toxic Speech Detectors Against Veiled Toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7732–7739.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*. 770–778.

Sophie Jentzsch, Patrick Schramowski, Constantin A. Rothkopf, and Kristian Kersting. 2019. Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 37–44.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the International Conference on Machine Learning, (ICML)*. 4904–4916.

Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. 2016. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *https://github.com/openimages* (2016).

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4. *Int. J. Comput. Vis.* 128, 7 (2020), 1956–1981.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* 117, 23 (2020), 12592–12594.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online and Punta Cana, Dominican Republic, 3045–3059.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A Survey of Transformers. arXiv preprint arXiv:2106.04554. (2021).

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. 5356–5371.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv preprint arXiv:2112.10741. (2021).

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2463–2473.

Guanghui Qin and Jason Eisner. 2021. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*. 5203–5212.

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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*. 8748–8763.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *CoRR* (2019).

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the International Conference on Machine Learning (ICML)*. 8821–8831.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5418–5426.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5477–5490.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. arXiv preprint arXiv:2103.00453. (2021).

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence* 4, 3 (2022),

258–268.

Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin A. Rothkopf, and Kristian Kersting. 2020. The Moral Choice Machine. *Frontiers Artif. Intell.* 3 (2020), 36.

Ryan Steed and Aylin Caliskan. 2021. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 701–713.

Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, (ICML) (Proceedings of Machine Learning Research, Vol. 97)*. 6105–6114.

Mingxing Tan and Quoc V. Le. 2021. EfficientNetV2: Smaller Models and Faster Training. In *Proceedings of the 38th International Conference on Machine Learning, (ICML) (Proceedings of Machine Learning Research, Vol. 139)*. 10096–10106.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems*.

Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. In *Proceedings of 16th European Conference of Computer Vision (ECCV)*. 733–751.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*. 547–558.