
Multimodal datasets: misogyny, pornography, and malignant stereotypes

Abeba Birhane*

University College Dublin & Lero
Dublin, Ireland
abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu*

Independent Researcher
vinaypra@alumni.cmu.edu

Emmanuel Kahembwe

University of Edinburgh
Edinburgh, UK
e.kahembwe@ed.ac.uk

Abstract

We have now entered the era of trillion parameter machine learning models trained on billion-sized datasets scraped from the internet. The rise of these gargantuan datasets has given rise to formidable bodies of critical work that has called for caution while generating these large datasets. These address concerns surrounding the dubious curation practices used to generate these datasets, the sordid quality of alt-text data available on the world wide web, the problematic content of the CommonCrawl dataset often used as a source for training large language models, and the entrenched biases in large-scale visio-linguistic models (such as OpenAI’s CLIP model) trained on opaque datasets (WebImageText). In the backdrop of these specific calls of caution, we examine the recently released LAION-400M dataset, which is a CLIP-filtered dataset of Image-Alt-text pairs parsed from the Common-Crawl dataset. We found that the dataset contains, troublesome and explicit images and text pairs of rape, pornography, malign stereotypes, racist and ethnic slurs, and other extremely problematic content. We outline numerous implications, concerns and downstream harms regarding the current state of large scale datasets while raising open questions for various stakeholders including the AI community, regulators, policy makers and data subjects.

Warning: This paper contains NSFW content that some readers may find disturbing, distressing, and/or offensive.

1 Introduction

The emergence of deep learning aided computer vision as a notable field of Artificial Intelligence (AI) ushered the so-termed *AI spring* [1] and has been characterized by its voracious need for vast volumes of data. The recent multi-modality drive within AI seeks to break away from the template of training siloed task-specific models for image classification, segmentation, or detection and entails curating cross-domain datasets and training cross-domain models that will jointly model the modalities of *vision*, *text*, and *speech* data. In the specific context of the *vision-text dyad*, the endeavor begins with curating large-scale datasets of tuples of the form: $D = \{(x_i, t_i, \mu_i)\}_{i=1}^N$ where x_i is the i^{th} image, t_i is the textual description associated with the i^{th} image, and μ_i is the i^{th} image’s meta-data. As has been the case with much of state-of-the-art (SotA) AI endeavors [2, 3], the dataset is expected to be

*Equal contribution

internet sized, thus rendering the usual theatre of data-curation to be the World Wide Web (WWW). The three constituent elements of the multimodal drive: the images, the alt-text image-caption pairs on the WWW, and the textual content gathered from corpora such as the *CommonCrawl* have raised various concerns. The rest of the introduction details these specific concerns.

1.1 Large Scale Image Datasets

The cosmology of large scale computer vision datasets contains various broad problems including curation biases, inclusion of problematic content in the images, the questionable approaches of associating these images with offensive and non-imageable labels, as well as the gradual erosion of privacy [4, 5]. Various works [6, 7, 8, 9] have highlighted gender, racial, and geographical biases surrounding the sourcing of image datasets as well as the opacity of such endeavors [10]. The *content* of the large scale vision datasets has also been found to include non-consensual-voyeuristic imagery [11] and NSFW content. Labeling is also a great concern. This includes stagnant vocabulary of labels [12], misrepresentation of gender [4], prevalence of ethnophaulisms [11] and non-imageability issues in the label space [12, 10, 13].

These critiques have resulted in some corrective measures including the retraction of the MS Celeb² and TinyImages³ datasets, blurring of the images of people [14] and filtering out of constituent images to create a sanitized version of the original dataset. For example, the curators of Imagenet advocated removing 2674 out of 2832 existing synsets in the person subtree of the label space [12]. This work is particularly informative as it specifically delves into the tenuous relationship between the content of an image and its textual-categorization description (WordNet synset) and highlights the seriousness of issues such as stagnant concept vocabulary and non-imageability (see Table 1 in Yang et al.’s paper [12]), which leads us to a parallel body of critique surrounding alt-text descriptions of images on the WWW.

1.2 Image-text pairs and alt-text

The alternative text (alt text) associated with an image element on a webpage is an HTML attribute that can be harnessed in case the element (image) cannot be rendered. The motivation behind alt text is to enable assistive technologies such as screen-reader software to deliver descriptions of contents of an image to blind and low vision people. In order to improve the quality of alt-text on the WWW, the World Wide Web Consortium (W3C) provides a comprehensive taxonomy of images neatly sub-classified into *Informative images*, *Decorative images*, *Functional images*, *Images of text*, *Complex images*, *Group-images* and *Image-maps* categories but also clearly describes an Alt Decision Tree (ADT)⁴ that captures the expected best practices to generate alt-text associated with the images being uploaded. Yet, what permeates the WWW is a vast wasteland of poorly written, sparsely available alt-text image descriptions. This has attracted the attention of accessibility advocates and ethicists alike, who have created a robust body of work that critically analyzes the Image-Textual-description dyad [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27].

These works have demonstrated how even highly reputable high-traffic websites have poor alt-text coverage of about 50% [28]. Furthermore, with regards to social media, Gleason et al. [29] found that, of the 9.2 million tweets they analyzed, only 0.1% contained alternative text. In [30], the authors looked at the issue of describing images of people with automated alt text and sought inspiration from the template used in museums.[31] hinted towards the prevalence of search engine ranking abuse schemes where poor quality of alt-text was embraced in order to hit high coverage rates. These observations reveal that a WWW-sized data dump of alt-text is besotted with high prevalence of issues such as missing important information, not being descriptive enough, resorting to stereotypical and offensive descriptors, being over descriptive (including filenames and special characters) or misrepresenting images [32, 30, 33].

²<https://exposing.ai/>

³<http://groups.csail.mit.edu/vision/TinyImages>

⁴<https://www.w3.org/WAI/tutorials/images/decision-tree/>

1.3 The Common-Crawl

Common Crawl is a San Francisco based nonprofit 501(c)(3) organization that has been regularly crawling the entire WWW and generating archival snapshot data-dumps, often termed the *Common-Crawl (CC) datasets* in machine learning lexicon, since 2011. The current version of this archive (dated April 2021) is roughly 320 TB in size and spans 3.1 billion pages. The sheer *scale* of this dataset has an enduring allure in the AI community and has been used as a seeding dataset in training pipelines of high-profile projects⁵ such as GPT-3 [34], CLUECorpus2020 [35], and XLM-R [36].

Inevitably this gargantuan dataset mined from the WWW suffers from serious issues. For instance, Matic et al. [37] used the *Curlie.org* crowdsourced taxonomy project to train a *GDPR-Article(9)-Sensitive-URL* classifier which revealed that, of the 1 Billion URLs they audited in the Common Crawl project, 155 million URLs fell into the *sensitive* category. The *Realtoxicityprompts* work [38] revealed that CommonCrawl contained over 300,000 documents from unreliable news sites and banned subReddit pages containing hate speech and racism. More recently, Luccioni and Viviano’s initial study [39] placed the ‘*Hate speech*’ content level to be around 4.02%-5.24% (the 1+ hate n-grams level was estimated higher at 17.78%). With regards to *CCAligned*, a 119- language parallel dataset built off 68 snapshots of Common Crawl, Caswell et al. [40] revealed that there were notable amounts of pornographic content (> 10%) found for 11 languages with prevalence rates being as high as 24% for language pairs such as en-om_KE.

The LAION-400M dataset emerges from this landscape containing hundreds of millions of Image-Alt-text pairs parsed from the Common-Crawl dataset and filtered using a previously Common-Crawl trained AI model (CLIP [2]). With this background, we present our findings following our initial audit of the LAION-400M dataset below.

The rest of the paper is structured as follows: In Section 2, we present our initial qualitative and quantitative analysis of the LAION-400M multimodal dataset. In Section 3 we provide the background behind this recent drive for ever larger multimodal datasets and illustrate the limitations of the approach used to create them. In Section 4, we outline the oft-ignored asymmetries between incautious large scale dataset curation and downstream detoxification processes. In Section 5, we examine dominant narratives for the emergence of multimodal datasets, outline their shortcomings, and put forward open question for all stakeholders (both directly and indirectly) involved in the data-model pipeline including policy makers, regulators, data curators, data subjects, as well as the wider AI community. In Section 6 we conclude the paper with some final thoughts and reflections.

2 LAION-400M

Note: All offensive imagery from this section has been hand blurred and moved to the Appendix after a blank page to give the reader the option not to visually engage should they choose not to.

Recently⁶ the LAION-400M dataset was released, adding to the growing list of large scale viso-linguistic multi-modal datasets amassed from the CommonCrawl data dump. Envisioned in parts, to be an open-source variant of the closed-source WIT (WebImageText) dataset, the dataset contains millions of (Image,Text,Meta-data) tuples extracted from the alt-text attributes of random web pages crawled between 2014 and 2021. After filtering out the raw image-alt-text pairs whose cosine similarity between the CLIP-text and CLIP-image embeddings was less than 0.3, the current version has 413871335 tuples, and is envisioned to cross the *2-digit billion* mark in the near future⁷. Alongside the dataset release, the curators also provided a *clip-retrieval K-nearest neighbor index* accessible via a graphic-user interface⁸. The machine learning community that interacted with this semantic-search portal began to raise concerns⁹ about the regularity which they began encountering NSFW, Offensive, violent and pornographic imagery, even in response to seemingly benign queries. In this section, we

⁵<https://commoncrawl.org/the-data/examples/>

⁶In September 2021

⁷This is indicated in the fund-raising page here: <https://gogetfunding.com/help-us-build-the-worlds-largest-open-billion-scale-image-text-dataset-perfect-for-training-dall-e-clip-other-multimodal-models/>

⁸<https://bit.ly/2Y8Iz8b>

⁹<https://bit.ly/3m26b6N>

highlight some of the problematic contents we discovered in the dataset and the associated query results the interface returned via an initial audit.

2.1 Misogyny and stereotypes

Upon querying the search portal (the version available on September 12th, 2021) with non-NSFW queries, we encountered a significantly high ratio of NSFW results that contained vivid depictions of sexual violence and other troubling imagery. Even the weakest link to womanhood or some aspect of what is traditionally conceived as feminine returned pornographic imagery. For example, when searched for descriptive adjectives such as `big` and `small` (Figures 4a and 4b respectively), terms such as `Asian`, `Indian` and `Nigerian` (Figures 5 (a), (b) and (c) respectively), relationship terms such as `Aunty` and `Mummy` (Figures 6 (a) and (b) respectively), cross-cultural terms such as `Maa` and `Abuela` (Figures 7 (a) and (b) respectively), or demographic-indicators such as `Latina` and `Black-woman` (Figures 8 (a) and (b) respectively); all returned images clearly sourced from pornographic websites. These images were not just prototypically "NSFW" from a parochial *nudity perspective* but also included explicit rape scene imagery as well as photo-shopped images of female celebrities.

Furthermore, we queried the dataset for terms such as `school girl` and `school boy` (Figures 9 (a) and (b) respectively), `beautiful`, `handsome`, and `CEO` (Figures 10 (a), (b) and (c) respectively), `African` and `European` (figures 11 (a) and (b) respectively) as well as terms such as `best president`, `worst president` (Figures 12 (a) and (b) respectively), and `terrorist` and `white power` (Figures 13 (a) and (b) respectively) to get a glimpse of how much the dataset can potentially aid in creating semantic-search technologies that end up perpetuating historical, social, and cultural stereotypes and political biases. The sample images reveal how the specific semantic search engine version meant to fetch images from LAION-400M, not only risked amplifying hyper-sexualized and misogynist representation of women, but also presented results that were reminiscent of Anglo-centric, Euro-centric, and potentially, White-supremacist ideologies.

2.2 Search engine bias?

While the images obtained from the search exercises presented in Section 2.1 do expose the presence of these images in the dataset, their retrieval in response to the associated queries can potentially be attributed to the CLIP-retrieval + Autofaiss module¹⁰ of the image-retrieval pipeline described in the announcement¹¹. In order to understand the phenomenon of repeated encountering of NSFW imagery in response to queries such as `Desi`, `Nun` and `Latina`, especially in the face of the claim that the NSFW-prevalence rate was *less than 1%* (see the "*Analysis of the LAION-400M data*" section in the announcement), we conducted an initial quantitative investigation. We downloaded all the 32 compressed parquet files related to the URL and caption meta-dataset that contained the following 8 fields: `[SAMPLE-ID, URL, TEXT, LICENSE, NSFW12, similarity, WIDTH, HEIGHT]`. We then carved out all the images that had the search term (such as `desi`) in the `TEXT` field via a simple string-match search. Lastly, we defined an alternative NSFW filter that simply checked if any of the following terms existed either in the `TEXT` or `URL` fields: `['porn', 'hot', 'adult', 'xxx', 'sex', 'f*ck', 'rape']`. The results are presented in Table 1.

The search terms `Desi` (Figure 15), `Nun` (Figure 16) and `Latina` resulted in 34516, 16766 and 37769 matches (denoted by N_{match}) of which 34%, 16.4% and 28.2% had the NSFW-terms listed above. Presented in the **NSFW-flag-values** column of the table are the value-counts of the CLIP-derived NSFW field that not just alludes towards its unreliability as a filtering parameter but also highlights the need for combined text and image based filtering steps used in projects such as the Wikipedia-based Image Text dataset [41] (also abbreviated as WIT) and the Conceptual Captions dataset [25]. Specifically referring to the *Image based filtering* module in [25], the authors state that "*It excludes images that trigger pornography or profanity detectors. These filters discard more than 65% of the candidates*". Further, with regards to the *Text based filtering*, they state that: "*We analyze*

¹⁰<https://github.com/rom1504/clip-retrieval/>

¹¹<https://laion.ai/laion-400-open-dataset/>

¹²NSFW: we used CLIP to estimate if the image has NSFW content. The estimation has been pretty conservative, reducing false negatives at the cost of more false positives. Possible values are "UNLIKELY", "UNSURE" and "NSFW".

candidate Alt-text using the Google Cloud Natural Language APIs, specifically part-of-speech (POS), sentiment/polarity, and pornography/profanity annotations". All this points towards the fact that this modality of filtering warrants techniques much more sophisticated than the string matches based one used here to demonstrate the level of prevalence of NSFW content.

2.3 Offensive text. Benign imagery

Another persistent occurrence during our investigation was the emergence of seemingly benign images associated with NSFW terms. Upon bookmarking these and getting the original images from the dataset, we uncovered a whole category of images that in many cases did have the image description but also contained NSFW and offensive text tags, that also highlight the need for joint Image-and-text based filtering like the one described in [25] that used a pre-trained vision model to predict textual labels, an endeavor that also resulted in the filtering away of 60% of the incoming candidate pairs. In Figure 17, we present a collage of these images that demonstrate the insidious nature of this phenomenon.

Table 1: Results of the string-search based experiment from the 413.871335 million sample search

Search string	N_{match}	$(N_{nsfw}, \%_{nsfw})$	NSFW-flag-values
Desi	34516	(11782, 34.1%)	{'UNLIKELY': 9327, 'UNSURE': 2291, 'NSFW': 164}
Nun	16766	(2761, 16.4%)	{'UNLIKELY': 1623, 'UNSURE': 863, 'NSFW': 273}
Latina	37769	(10658, 28.21%)	{'UNSURE': 5724, 'UNLIKELY': 4013, 'NSFW': 918}

3 How did we get here?

In this section, we present some nuances pertaining to the creation process that results in the birth of datasets such as LAION-400M. We posit that such a large-scale undertaking involves:

1. A well-defined motivational drive to begin such a venture.
2. A large-scale *base source* to seed the curation process.
3. A filtering mechanism to turn the raw dataset into one worthy of being fed into a multimodal model training pipeline.

In the following subsections, we explore each of the three above-stated *sub-modules* in the specific context of LAION-400M dataset, provide the associated background, and specifically highlight the issues plaguing each of them.

3.1 Motivational drive: Open-sourcing the closed-source

The recent emergence of grassroots based open-sourcing initiatives can be attributed to an increasing adoption of the closed-source commercial API access mode of dissemination being used for projects such as GPT-3 [34], CLIP and DALL-E¹³. *EleutherAI*¹⁴ achieved success by replicating both the WebText dataset (on which GPT-3 was trained) and the GPT-3 model itself by unveiling the Pile dataset [42] and the GPT-Neo [43]/GPT-NeoX [44] models. As indicated in the README section of the LAION Github repository¹⁵, the primal motivation behind the LAION-400M undertaking was to produce open-source variants of the opaque WIT (WebImageText) dataset, and the CLIP [2] and DALL-E [45] models.

3.2 Crawl over Curate

The recent past has seen a paradigm shift in the way image-text multimodal datasets are being curated. The 2010-2020 decade saw the emergence of smaller scale initiatives such as the UIUC-Pascal-Sentence Dataset [46], Microsoft COCO [47]: Common Objects in Context dataset (330,000 images

¹³The API-FAQ section here: <https://openai.com/blog/openai-api/> addresses questions like: *Why did OpenAI decide to release a commercial product? Why did OpenAI choose to release an API instead of open-sourcing the models?*

¹⁴A grassroots collective of researchers: <https://www.eleuther.ai/>

¹⁵<https://github.com/rom1504/cah-prepro/blob/main/README.md>

with 5 independent human generated captions), the Yahoo Flickr Creative Commons 100 Million (YFCC100M) Dataset [48], the Visual Question Answering [49](VQA dataset with 265016 images with at least 3 questions per image and 10 “ground truth” answers per question) and the Visual Genome [50] (108,077 Images with 5.4 Million Region Descriptions and 1.7 Million Visual Question Answers) that all banked on a rough template of crowd-sourced captioning of a pre-existent image dataset either by using platforms such as Amazon Mechanical Turk or using photo-uploader captions from Flickr.

Recently, breaking away from this tradition, 2021 saw the emergence of large-scale opaque multi-modal initiatives such as CLIP [2], ALIGN [3], MUM [51] and Wu Dao 2.0 [52], that discarded the traditional recipe of handheld data curation and embraced another template that would scale their datasets into hundreds of millions or even billions of images: *Crawling the world-wide-web for image captions*. CLIP [2] used an internally curated proprietary WIT (WebImageText) dataset consisting of 400 million (image, text) pairs collected from a variety of publicly available sources on the Internet. Their *model-card*¹⁶ documentation states that: “*The model was trained on publicly available image-caption data. This was done through a combination of crawling a handful of websites and using commonly-used pre-existing image datasets such as YFCC100M. A large portion of the data comes from our crawling of the internet.*” We get a deeper insight into how the text captions were actually generated only via a github-issue response¹⁷ by one of the dataset’s co-authors, which reveals that: “*The dataset is a mixture of image-text pairs from various sources. The ‘full text sequences’ are usually title + description concatenated using whatever is available about the image, usually being a sentence or two and not the whole webpage.*”

The ALIGN [3] project went one step further and created a billion-sized dataset based on image-alt-text pairs. In doing so, this work not only justified such cavalier curation practices as a liberatory process that would ultimately save human effort and costs, but also cemented this simple yet powerful belief that “**scale beats noise**”, a thought that we delve into in Section 3.2.1. After the announcement of ALIGN, we encounter this alt-text based curation aspect with regards to the multimodal Wu Dao 2.0 1.75 trillion parameter model [53, 52, 54, 55, 56] that was supposedly trained on *4.9 terabytes of images and texts, which included 1.2 terabytes of Chinese text and 1.2 terabytes of English text*. While there are no publicly known documentations or insights into the dataset curation process, it is through encountering claims¹⁸ that read: “*The model can not only write essays, poems and couplets in traditional Chinese, it can both generate alt text based off of a static image and generate nearly photo-realistic images based on natural language descriptions*”, that we uncover the emergence of the alt-text aspect.

3.2.1 The "scale beat noise" discourse

Jia et al. [3] make the following claim that: “*This costly curation process limits the size of datasets and hence hinders the scaling of trained models. In this paper, we leverage a noisy dataset of over one billion image alt-text pairs, obtained without expensive filtering or post-processing steps in the Conceptual Captions dataset*”. This tactfully contextualizes the removal of thoughtful curation as freeing the dataset-creation process from the *stumbling block* of high curation costs.

Furthermore, the paper and its associated blog-post¹⁹, develop a two-stage strategy that further substantiates this narrative of the *futility of pre-emptive filtering*. Firstly, the widespread irrelevance between the image content and the alt-text descriptions on the WWW (as explored in Section 1.2) is neatly accommodated as ‘noise’. Then, ‘scale’ is introduced as a liberating panacea that not only frees the downstream machine learning pipeline from the clutches of expensive filtering or post-processing steps but also *makes up* for the so-termed ‘noisy’ data collected, as the mis-captioning is going to be somehow ‘*averaged out*’ through the correct captioning elsewhere in the dataset. Such lines of thinking are not unique to this specific context but form a widespread belief that drives initiatives such as LAION-400M, and permeate the entire field of the multi-modal pursuit. Yet, scale thinking, scholars have argued, stands at the opposite side of liberatory or effective systemic change [57].

¹⁶<https://github.com/openai/CLIP/blob/main/model-card.md>

¹⁷<https://github.com/openai/CLIP/issues/118#issuecomment-871263743>

¹⁸<https://www.engadget.com/chinas-gigantic-multi-modal-ai-is-no-one-trick-pony-211414388.html>

¹⁹<https://ai.googleblog.com/2021/05/align-scaling-up-visual-and-vision.html>

3.3 Filtering mechanism: CLIP

The third part of the multimodal dataset curation pipeline involves algorithmic filtering of images to include only those that have a high level of similarity between the semantic content of the image and the ensuing textual description. In the context of LAION-400M this was done by firstly calculating the cosine similarity between the text-description and image embeddings obtained via the CLIP model and dropping those with a cosine-similarity below 0.3, as illustrated in Figure 3. Besides the obvious data-incest issue of using CLIP as a filtering model in order to potentially generate CLIP-like models, this approach, we argue, was ill advised on account of other serious issues such as downstream propagation of known offensive mis-associations and also unintended usage of the model.

3.3.1 Known biases

CLIP suffers from various biases. The CLIP-paper [2] itself (in Section 7.1) outlined that images belonging to the ‘Black’ racial designation had an approximately 14% chance of being mis-categorized as [‘animal’, ‘gorilla’, ‘chimpanzee’, ‘orangutan’, ‘thief’, ‘criminal’ and ‘suspicious person’] in their *FairFace* dataset experiment. Furthermore, it has emerged through both online-discussions²⁰ and OpenAI’s own visualization projects such as Microscope²¹ that graphic NSFW /pornographic samples²² might not have been filtered out from the training dataset. A flagship example of this is the Unit 1543²³ in `image_block_4_5_Add_6_0` of the CLIP-Resnet-50-4x model detailed in Appendix B. Additionally, other works [58, 59] have revealed a variety of *typographical, conceptual, and iconographic* vulnerabilities and mis-association tendencies associated with the model.

3.3.2 Unintended use: Model card

CLIP’s model card²⁴ explicitly states that “*The primary intended users of these models are AI researchers. We primarily imagine the model will be used by researchers to better understand robustness, generalization, and other capabilities, biases, and constraints of computer vision models*”. Further, with regard to ‘Out-of-Scope’ use cases, the model card states that: “*Any deployed use case of the model - whether commercial or not - is currently out of scope. Non-deployed use cases such as image search in a constrained environment, are also not recommended unless there is thorough in-domain testing of the model with a specific, fixed class taxonomy. This is because our safety assessment demonstrated a high need for task specific testing especially given the variability of CLIP’s performance with different class taxonomies. This makes untested and unconstrained deployment of the model in any use case currently potentially harmful*”. Thus, one might argue that CLIP was not intended for use in an application such as the LAION-400M dataset curation process in the first place.

3.3.3 Cosine-similarity thresholding

This sub-section illustrates how the ad-hoc assumption of 0.3 cosine-similarity threshold can be a source of trouble. Two examples highlight so called *corner-cases*.

The first example entails the famous photograph of Eileen Collins — an American astronaut who first piloted the space shuttle STS-63 in 1995 — from the scikit-image library²⁵. Figure 1 shows her picture along with two descriptions of the image: Text-input-1: “This is a portrait of an astronaut with the American flag” and Text-input-2: “This is a photograph of a smiling housewife in an orange jumpsuit with the American flag”. CLIP produces the following cosines similarities for the image with Text-input-1 and Text-input-2 respectively: 0.28 and 0.31. Now, imagine the scenario where the scraper module encountered 2 instances of this image, the first with the reasonable benign description in Text-input-1 and the

²⁰https://www.reddit.com/r/MachineLearning/comments/m0119w/d_openais_clip_and_dale_seem_to_be_trained_on/

²¹<https://microscope-azure-edge.openai.com/models>

²²<https://distill.pub/2021/multimodal-neurons/>

²³https://microscope-azure-edge.openai.com/models/contrastive_4x/image_block_4_5_Ad_6_0/1543

²⁴<https://github.com/openai/CLIP/blob/main/model-card.md>

²⁵<https://scikit-image.org/docs/dev/api/skimage.data.html?highlight=bool>

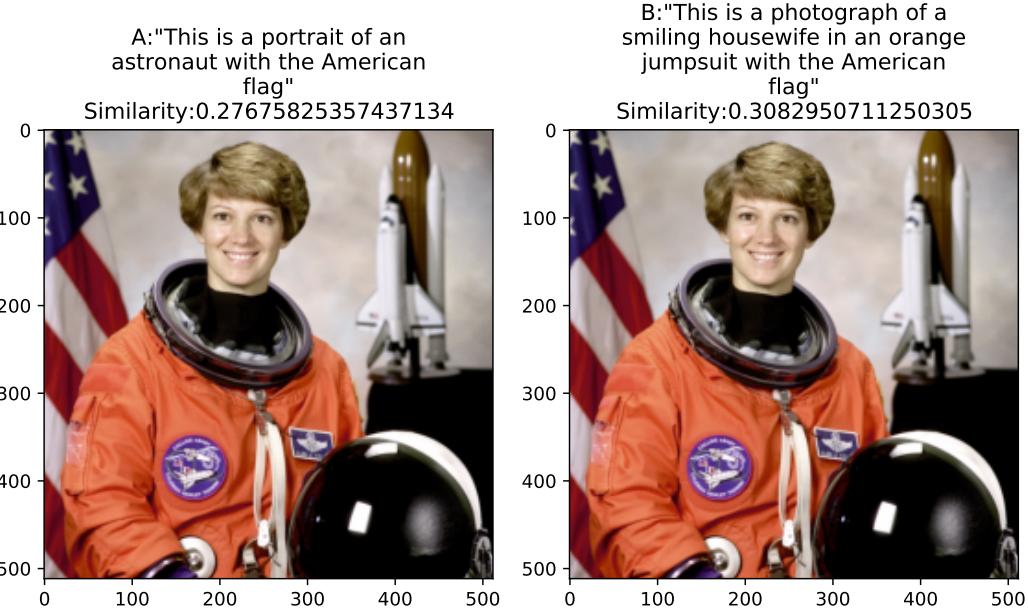


Figure 1: Results of the CLIP-experiments performed with the color image of the astronaut Eileen Collins obtained via `skimage.data.astronaut()`

second with the misogynistic description of `Text-input-2`. Due to the gender biases built into CLIP, the odds of the misogynistic one making it through the filtering process might be higher.

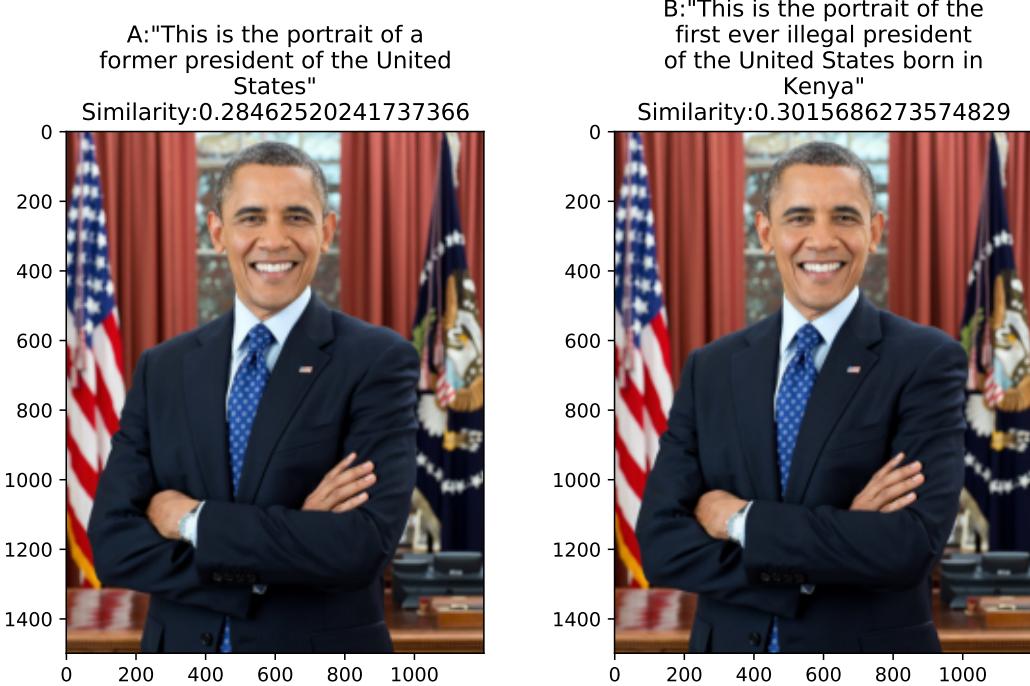


Figure 2: Results of the CLIP-experiments performed with the official portrait image (from 2012) of Barack Obama (the 44th President of the United States) where the conspiracy-theoretic textual descriptions obtains a cosine-similarity higher than 0.3

The second example demonstrates similar issues with Barack Obama’s Official portrait from 2012. Figure 2 shows the portrait with two text descriptions: Text-input-1: “This is the portrait of a former president of the United States” and Text-input-2: “This is the portrait of the first ever illegal president of the United States born in Kenya”. While CLIP produces a cosine similarity less than 0.3 for the first factual description, it produces one above the 0.3 threshold for the second one.

The main point here is *not* that we successfully generated provocative examples but that the sheer ease of producing such so-termed “*corner cases*” emanates directly from the strong mis-associations baked into the model that can potentially amplify selection bias towards offensive samples in the CC corpus. Readers are invited to try out further examples via our publicly available colab notebook²⁶.

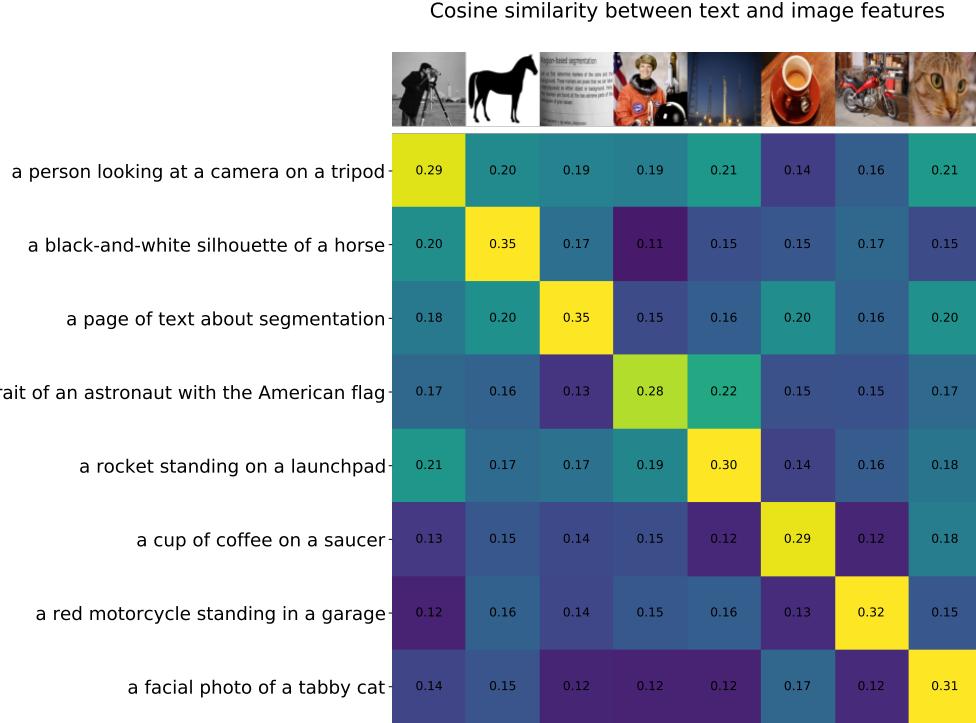


Figure 3: The Cosine similarity matrix between the text and image features pertaining to the *skimage* examples in the *Interacting with CLIP* colab notebook shared by CLIP’s authors

Lastly, with regards to the under-scoring aspect of the 0.3 cosine-similarity filtering mechanism, we reproduce the *sk-image* examples provided in the associated official colab notebook²⁷ and draw the reader’s attention towards the images associated with reasonably accurate descriptions that still yield a cosine similarity of less than 0.3, as highlighted in Figure 3

4 The asymmetries of course-correction

In anticipation of the release of larger versions of the LAION-400M dataset and other datasets similar to this, we stress the following oft-ignored asymmetries that hinder downstream harm reduction endeavors such as dataset and model detoxification.

²⁶https://github.com/vinayprabhu/Crimes_of_Vision_Datasets/blob/master/CLIP_astro_obama.ipynb

²⁷https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

4.1 Asymmetry of efforts: Crawling v/s detoxification

The asymmetry in the volume of efforts required in the crawling-and-aggregation phase of WWW-mined datasets and the ensuing harm-reduction phase (with regards to either filtering the dataset or detoxification of the models trained on the dataset) is significantly stark. The emergence of well documented tools made available by the Common-Crawl organization²⁸, and the wide-spread availability of async concurrency and I/O Python libraries such as *Trio*, *curio*, *asyncio*, *Twisted* and *asks*, means that the process of mining-and-aggregating such large datasets has become both incredibly “democratized” and relatively cheap. The LAION-400M team notes that: “*For every \$5000 we get we will be able to extend our data-set by at least 1 billion samples, conservatively estimated, ... likely by more!*”. Source: <https://bit.ly/3zBLnry>. On the other hand, as recently demonstrated in studies such as [60, 61], granular safe filtering of the datasets created and the downstream detoxification of the models trained on such datasets remain a tenuous and laborious work. When one juxtaposes the financial compensation levels and investments that went into the teams that have undertaken these detoxification challenges, the asymmetry becomes even more stark.

Further, as demonstrated in the curation process of other large scale visio-linguistic datasets such as the Wikipedia-based Image Text dataset [41] (also abbreviated as WIT) and the Conceptual Captions dataset [25], there were distinct Image-based Filtering, Text-based Filtering and Joint Image-and-text-based filtering modules that utilized a large suite of highly specialized Computer Vision and NLP APIs (like part-of-speech, sentiment/polarity, and pornography/profanity annotations), to curate the final dataset whose costs can be far greater than the \$5000-per-billion images cited in the LAION-400M endeavor.

4.2 Asymmetry of ‘advances’: Model advances v/s dataset advances

The culture in machine learning is such that ideas that promise improvements in training speed, model-size or top- k accuracy improvements are rapidly embraced while ideas and revelations pertaining to unethical aspects of datasets are either ignored or take a long time to lead to changes [62]. For example, the ImageNet dataset was released in 2009 [63] but the course-corrections regarding the vast number of non-imageable classes [12] and loss of privacy [14] were undertaken only in the 2019-21 period which is more than a decade after its release. At the same time, between 2009-2021, the community managed to “democratize” means to train SotA models in less than 11 minutes [64], made available pre-trained models that are as compressed as 3.8 MB (the MeliusNet22 model) and as fast as 17ms inference-time on a commercially available smartphone²⁹ (the QuickNetSmall model).

Of particular relevance to the LAION-400M dataset is this realization that all the post-curation filtering recommendations by ImageNet’s curators [12] mandating removal of more than 2700 synsets from the ImageNet-21k dataset in December 2019 have largely been ignored. This is highlighted in the emergence of bigger datasets such as Tencent ML-images dataset [65] (in February 2020) that encompasses most of these non-imageable classes³⁰, the continued availability of models trained on the full-ImageNet-21k dataset in repositories such as *TF-hub*³¹, the continued usage of the unfiltered-ImageNet-21k in the latest SotA models (such as Google’s latest EfficientNetV2 and CoAtNet models [66]) and the explicit announcements permitting the usage of unfiltered-ImageNet-21k pretraining in reputable contests such as the LVIS challenge 2021³². We stress this crucial observation: A team of the stature of ImageNet managing less than 15 million images has struggled and failed in these detoxification attempts thus far. The scale of careful efforts required to thoroughly detoxify this massive multimodal dataset and the downstream models trained on this dataset spanning potentially billions of image-caption pairs will be undeniably astronomical.

²⁸See <https://commoncrawl.org/the-data/tutorials/>

²⁹The MeliusNet22 model hits 83.9% top-5 accuracy and is 3.88 MB and the QuickNetSmall model that achieves 81.8% top-5 accuracy has a latency of 17.5ms. Source: <https://docs.larq.dev/zoo/>

³⁰<https://github.com/Tencent/tencent-ml-images>

³¹https://tfhub.dev/google/imagenet/efficientnet_v2_imagenet21k_xl/feature_vector/

²

³²https://www.lvisdataset.org/challenge_2021

4.3 Asymmetry of labour: Emotional trauma

While the endeavor of researching techniques and training models that hit SotA accuracy metrics can certainly be labour intensive and challenging, there is a specific aspect of the labour that dataset-cleanup efforts merit that is often missed in machine learning literature: Emotional trauma.

We found the emotional toll of sifting through the LAION-400M dataset, curating the list of examples and strategically blurring them to be profoundly overwhelming at times. The NSFW aspect of the imagery involved meant that we work in isolation away from our official environments where we ran the risk of exposing our co-workers to the insidious imagery (See [67, 68]). We (as well as our colleagues who aided us) experienced varying levels of discomfort, nausea, and headache during the process of probing the dataset. Additionally, this kind of work disproportionately encounters significant negative criticism across the academic AI sphere upon release, which not only adds an additional emotional toll to the already heavy task of studying and analysing such datasets but also discourages similar future work, much to the detriment of the AI field and society in general.

5 Discussions and open questions

Visio-linguistic datasets at the scale of LAION-400M have previously been inaccessible to those outside of BigTech companies and the few institutes with massive resources to collect them. LAION-400M is a monumental effort to change this under the drive to *democratize* large scale datasets. In one sense, we commend this initial effort. However, this work demonstrates that such a conceptualization of "democratization" is too narrow, and fails to foresee many of the problems we highlight. It fails to account for the rights, welfare, and interests of vulnerable individuals and communities, many of whom are likely to suffer worst from the downstream impacts of this dataset and the models trained on it [69]. Having said that, this effort opens a door that allows the wider AI community to get a glimpse into the world of large scale datasets; the kind of datasets that remain hidden within the data centers of BigTech companies. It allows the community and its stakeholders to ask and pursue richer questions relevant for understanding the implications of datasets collected from the internet at scale, and by proxy, the AI models trained on them. Researchers, auditors, regulators, policy makers and other AI stakeholders can finally start to analyse and study these datasets leading to a better understanding of their capabilities, limitations, risks, and any harms they may cause or exacerbate. We hope the wider AI community and all stakeholders involved in/impacted by large scale datasets engage with these discussions; we open some questions below:

5.1 What should be in a dataset?

In the pre deep-learning era, datasets were often collected with purpose; a specific goal and task in mind. Many of these datasets had inherent issues and caused harm, but this was often restricted to their intended use cases and problem domains. The current state-of-the-art deep-learning based models attempt to train large-scale "general purpose" AI models on large internet collected datasets, then finetune (or specialise) them to target tasks. These large-scale AI models can be viewed, in the simplest case, as compressed representations of the large-scale datasets they are trained on. Under this light, it is important to ask what should be compressed within the weights of a neural network and by proxy, what is in a training dataset. Often, large neural networks trained on large datasets amortize the computational cost of development via mass deployment to millions (or even billions) of users around the world. Given the wide-scale and pervasive use of such models, it is even more important to question what information is being compressed within them and disseminated to their users.

5.2 Is this the path to AGI?

There is a growing community of AI researchers that believe that a path to Artificial General Intelligence (AGI) exists via the training of large AI models with "all available data". The phrase "all available data" often encompasses a large trove of data collected from the WWW (i.e. images, videos, and text). As seen in Sections 2 and 1.3, this data includes images and text that grossly misrepresent groups such as women, embodies harmful stereotypes, overwhelmingly sexualize Black women, and fetishize Asian women. Additionally, large scale internet collected datasets also capture illegal content, such as images of sexual abuse, rape and non-consensual explicit images. We raise the

question, does building AGI — assuming that the very premise that large scale multimodal datasets are the route to it is not fallacious to begin with — entail feeding models with the online world’s ugliness? How many images of rape are acceptable to feed into a supposedly AGI in order for it to “understand” the world? Given that cost and benefit are distributed unevenly for any given AI system — where those creating AI benefit the most while individuals and communities at the margins of society pay the highest price when AI fails [69] — is this a price worth paying in order to get better predictive text, or semantic search?

5.3 Are large neural networks a new distribution medium of illicit materials?

Large neural networks are known to memorize some data samples out right, even if they occur just once in the entire dataset [70]. There is the possibility that a large multimodal AI model will outright memorise a data sample that is illegal (e.g. sexual abuse, rape, etc). Thus, a situation may arise where a multimodal AI model not only puts out explicit content, but also illegal content. Even if an AI model does not outright memorise samples, Figure 18 shows that neurons can arise that capture illicit and harmful content in robust and recoverable ways. This raises a question as to how much information can model inversion techniques recover from such multimodal AIs. And whether the weights of large multimodal AIs can be used to smuggle illicit and illegal data around the internet (bypassing conventional alert and detection mechanisms)?

5.4 Whose data rights? Whose data Ownership?

When scrapping the web for the data used to create such datasets; questions of data ownership and rights emerge. Some of this data, especially image data, may be publicly “available” but scrapping and creating a large dataset with it is another issue. As we have found, some of this data is outright illegal; e.g. images that capture a moment of deep trauma for the subjects depicted in them (e.g. sexual abuse, rape). Furthermore, such datasets are collected without the consent and awareness of the data subject [11]. The question, then, arises: if and how a dataset collected in such a manner should be disseminated?

For example, the LAION-400M dataset is released under the "Creative Common CC-BY 4.0" licence, which has little restriction on how the dataset is used by others. Yet, next to the licence declaration, the LAION website states that "*The images are under their copyright*". This is a form of attribution that relies on the "*diffusion of responsibility*". The dataset authors delegate the responsibility of ensuring copyright is not violated onto the dataset users, diffusing their responsibility onto others. Regardless, the authors may still fall foul to laws in different parts of the world such as the European Union (EU); where such a dataset may be in violation of Article 15 of the EU Copyright directive and the General Data Protection Regulation (GDPR), that applies to all datasets that are not anonymized.

Putting the legal issues aside, the question of how ethical it is to carry out research using such datasets remains. For example, the subjects depicted in internet collected images are not notified and remain unaware to the fact that their likeness is being used for research and possibly even being commercialised, which raises the further question of consent and fair compensation. Furthermore, as we have seen from analysing the LAION-400M dataset, the overwhelming visual depiction of certain groups on the WWW is marred with malignant stereotypes and carries actual threat to vulnerable individuals. For example, many explicit images of women captured in this dataset are from the pornography industry, which itself has to deal with many issues; e.g. sexual slavery, mental, physical and drug abuse.

Given what the datasets like LAION-400M contain, the use of such datasets is highly likely to perpetuate the exploitation of individuals from minoritized groups. Individuals may delete their data from a website and assume that it is gone forever, while it may still exist on the servers of several researchers and organisations. There is a question as to who is responsible for removing that data from use in the dataset? For LAION-400M, the creators have delegated this task to the dataset user. Given such processes are intentionally made complex and that the average user lacks the technical knowledge to remove their data, is this a reasonable approach?

Lastly, the LAION-400M dataset in its current state may not be suitable for release under the "Creative Common CC-BY 4.0" licence, even given its potential for *democratization* of large scale multimodal datasets. The possible long term harm, especially towards those at the margins of society, caused by the release of such datasets as well as its ease of accessibility under a nonrestrictive licences

surpasses the potential benefits of “democratization”. This is not to rule out the possibility that it may be worthwhile to consider the release of such datasets under restrictive non-commercial licences and strictly for research purposes. This would allow for some “democratization” of large scale datasets, while allowing researchers and other AI stakeholders the time to analyze, study and better understand the data. This may also allow for similar grassroots efforts to clean the dataset and/or allow researchers to come up with better automated filtering mechanisms. Nonetheless, the rights of the data subject remain unaddressed here. It is reckless and dangerous to underplay the harms inherent in such large scale datasets and encourage their use in industrial and commercial settings. The responsibility of the licence scheme under which the dataset is provided falls solely on the dataset creator.

5.5 Is content moderation and filtering even feasible at this scale?

In previous sections (see Section 3.3, for example), we have demonstrated that the filtering mechanisms used on the LAION-400M dataset is unreliable at best and harmful at worst. There may be better algorithms for automatic filtering of such datasets but their reliability, especially in the unconstrained visual domain, is likely very low. Some may argue that the path forward would be to iterate and improve the tools used for automatic filtering. But without careful contextual analysis, filtering mechanisms are likely to censor and erase marginalized experiences [71]. Often, sensible filtering requires time and resources yet datasets such as LAION-400M already exist right now in the public domain. Some works have suggested that it is impossible to filter and clean large datasets with the set of methods and techniques currently available [70]. This presents questions such as: should an organisation collect, release and/or use a dataset it is incapable of cleaning itself? And assuming the answer is no, does that mean that collecting and releasing larger scale datasets should be restricted to the *likely* larger organisations with the resources to clean them?

It is also questionable how far automated filtering mechanisms can go to helping tackle these issues. Such mechanisms will always have some non-zero error rate and this has huge implications at scale, especially in the visual domain. It then becomes pertinent to ask the question; what rate of sexual abuse or rape images are acceptable in a billion scale multimodal dataset? At a 0.1% incidence rate, that means accepting a million images of minors being sexually assaulted within such training datasets. A million images of sexual abuse on any device would be a cause for serious concern, but is it acceptable when hidden among 999 million other images? Or is 100k such images acceptable? Maybe 10k?

Crucially, given what we have learned from the initial exploration of this dataset; it becomes more critical to understand how the private large-scale datasets used in BigTech compare with regards to these issues. It is almost certain that if large technology companies are automatically filtering their datasets, which they likely are, they will suffer from the same issues identified in the LAION-400M dataset. This further motivates the need for independent dataset auditors who can be trusted to go into these organisations, audit their datasets and publicly release the audit results to the wider AI community and its stakeholders.

And lastly, as we seek answers to these questions; what can be done about datasets such as LAION-400M in the meantime? Such datasets present a threat to Black women, ethnic minorities, children and generally to individuals and communities at the margins of society. They are likely to be utilised by entities that are not aware of and/or do not care about the issues that such datasets propagate.

5.6 Does this ‘multimodality’ in web scrapped data exacerbate stagnant stereotyping?

It is often said that “*a picture is worth a thousand words*”, but that also means there can be a thousand different stories, perspectives and interpretations of a single picture. The push towards multimodal datasets has gained significant momentum within the large-scale AI community as it is seen as one way of pre-training high performance “general purpose” AI models, recently rebranded as “foundation” models. But there is yet to be a discussion on if/how issues in one modality confound those from other modalities.

Language models commonly represent the textual modality through a fixed vocabulary of tokens, from which words and/or sentences are composed in order to transmit information. Each token encodes and embodies some atomic piece information in and of itself. However, the image modality has no comparable vocabulary, it can be thought of as being unconstrained. It is often left to a human

to decide how to constrain this modality when representing visual information. Sometimes the image as a whole represents some atomic meaning, and other times it is only some part of the image that is of interest (and all other information in the image should be ignored). The characteristics and dynamics of these modalities are vastly different from each other. The AI community is currently exploring the issues and solutions relevant to each of the individual modalities. But it is now also pertinent to ask questions about the issues that may arise from the amalgamation of multiple modalities.

When an image is taken, the responsibility is often left to the individual who uploads it to the WWW to provide an associated textual description. As discussed in Section 1.2, not many alt texts are available and the available textual descriptions are of very low quality, often ingrained with stereotypical and offensive descriptors. There are several reasons for this, but chief among them is priming search engines in order to increase engagement with online content. Most visual content that has alt text and is available for download via scraping tools is pornographic, and the alt text associated with such images, which may have a relative benign representation in the purely textual context, is often perverted through the lens for sociocultural fetishizations of the same terms in the visual context. For example, in the LAION-400M dataset, words such as ‘mom’, ‘nun’, ‘sister’, ‘daughter’, ‘daddy’ and ‘mother’ appear with high frequency in alt text for sexually explicit content. We have also observed a similar effect in the reverse direction, e.g. where innocent images of school girls have alt-text that is loaded with terms typically searched for by paedophiles and sexual predators.

What does this all mean if AI models are learning joint/shared embeddings of both image and text data? When AI models are being trained to compress and represent the information in the visual and textual domain within a shared latent space, does the representations of women for example, in the visual domain lead to a multimodal AI model that is more likely to sexualize women in the textual domain when compared against its language-only counterpart (or image-only counterpart in the reverse)? Do the problems in one modality merge with other modalities and exacerbate issues such as racism, misogyny, stagnant stereotyping?

6 Conclusion

The LAION-400M dataset provides a first-hand insight into the challenges and issues of dealing with multimodal visio-linguistic datasets at scale. Although the open access release of this dataset does warrant recognition, there are serious issues with the manner in which the dataset has been released and is being currently disseminated. We hope that this work encourages conversations regarding how to better tackle the issues inherent to large-scale internet collected data in an open and accessible manner. Thus far, datasets of this magnitude have remained closed, hidden away within large institutes and organisations. This has potentially stifled the progress in research on such large datasets, especially with regard to the issues inherent to them. Additionally, the downstream effects of hidden large scale datasets are likely to be devastating on marginalized communities. Therefore, we acknowledge the grassroots aspect of the endeavor and commend the LAION-400M creators for providing a window into this world and encourage them to keep the dataset accessible to researchers. This project has veritably demonstrated at scale, the serious failings of the CLIP model and the dangers of building semantic search engines off of this technology.

When issues such as the ones highlighted in this work are identified, retraction is often the path of least resistance. For example, Peng et al [72] examined three major retracted large scale image datasets; *DukeMTMC*, *MS-Celeb-1M*, and *Tiny Images*. Despite retractions, the authors found that the datasets remain widely available through file sharing websites and as derivatives. Months after their retractions these datasets were used hundreds of times in published papers and the datasets continue to be used by the ML community in peer-reviewed research. The closing of datasets following audit work like ours, is often a step backwards for the community as it does little to tackle the core issues inherent to these datasets. We do not believe that retraction is the right answer, especially in this case due to the difficulty researchers face in accessing such a dataset. We however believe that a more restrictive licence would be beneficial to limit the use of this dataset in non-research environments. This would allow for a concerted effort to tackle the questions and issues highlighted in work such as this and its derivatives.

Finally, we highly encourage other large institutions to open up their datasets to both internal and external audits in a thoughtful manner. Although there may be some competitive advantage to the large-scale private datasets, the harms potentially caused by these datasets will likely outweigh

them. It is also likely that as a community, we do not yet fully understand the risks of using such datasets. But relying on obscurity as a shield from scrutiny may implode in a publicly and financially irreparable manner.

We critique because we care.

And it is good to care.

Acknowledgements

We would like to thank Thomas Laurent and Timnit Gebru for the invaluable comments on an earlier version of the paper. Abeba Birhane was supported, in part, by Science Foundation Ireland grant 13/RC/2094_2

References

- [1] M. Mitchell, “Why ai is harder than we think,” *arXiv preprint arXiv:2104.12871*, 2021.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, 2021.
- [3] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” *arXiv preprint arXiv:2102.05918*, 2021.
- [4] M. K. Scheuerman, E. Denton, and A. Hanna, “Do datasets have politics? disciplinary values in computer vision dataset development,” *arXiv preprint arXiv:2108.04308*, 2021.
- [5] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, “Data and its (dis) contents: A survey of dataset development and use in machine learning research,” *arXiv preprint arXiv:2012.05345*, 2020.
- [6] J. Atwood, Y. Halpern, P. Baljekar, E. Breck, D. Sculley, P. Ostyakov, S. I. Nikolenko, I. Ivanov, R. Solovyev, W. Wang *et al.*, “The inclusive images competition,” in *The NeurIPS’18 Competition*. Springer, 2020, pp. 155–186.
- [7] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020.
- [8] A. Wang, A. Narayanan, and O. Russakovsky, “Revise: A tool for measuring and mitigating bias in visual datasets,” in *European Conference on Computer Vision*. Springer, 2020, pp. 733–751.
- [9] E. Denton, A. Hanna, R. Amironesei, A. Smart, and H. Nicole, “On the genealogy of machine learning datasets: A critical history of imagenet,” *Big Data & Society*, vol. 8, no. 2, p. 20539517211035955, 2021.
- [10] V. U. Prabhu, “The phantom of the corpora: JFT-300M,” in *Proceedings of the 2021 Beyond Fairness CVPR Workshop*, 2021, available at shorturl.at/djnR0.
- [11] A. Birhane and V. U. Prabhu, “Large image datasets: A pyrrhic win for computer vision?” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2021, pp. 1536–1546.
- [12] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, “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*, 2020, pp. 547–558.
- [13] K. Crawford and T. Paglen, “Excavating ai: The politics of images in machine learning training sets,” *AI & SOCIETY*, pp. 1–12, 2021.
- [14] K. Yang, J. Yau, L. Fei-Fei, J. Deng, and O. Russakovsky, “A study of face obfuscation in imagenet,” *arXiv preprint arXiv:2103.06191*, 2021.
- [15] F. Bramlett, “How will we manage the alt text?” *Pencil Panel Page*, 2012.
- [16] T. C. Craven, “Some features of “alt” texts associated with images in web pages.” *Information Research: An International Electronic Journal*, vol. 11, no. 2, p. n2, 2006.

- [17] P. Dognin, I. Melnyk, Y. Mroueh, I. Padhi, M. Rigotti, J. Ross, Y. Schiff, R. A. Young, and B. Belgodere, “Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge,” *arXiv preprint arXiv:2012.11696*, 2020.
- [18] J. Guo and J. Zhou, “Why ai alt text generator fail,” 2020.
- [19] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, “Captioning images taken by people who are blind,” in *European Conference on Computer Vision*. Springer, 2020, pp. 417–434.
- [20] M. Hanley, S. Barocas, K. Levy, S. Azenkot, and H. Nissenbaum, “Computer vision and conflicting values: Describing people with automated alt text,” *arXiv preprint arXiv:2105.12754*, 2021.
- [21] K. MACK, E. CUTRELL, B. LEE, and M. R. MORRIS, “Designing tools for high-qality alt text authoring,” 2021.
- [22] T. McEwan and B. Weerts, “Alt text and basic accessibility,” in *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21*, 2007, pp. 1–4.
- [23] M. R. Morris, “Ai and accessibility,” *Communications of the ACM*, vol. 63, no. 6, pp. 35–37, 2020.
- [24] H. Petrie, C. Harrison, and S. Dev, “Describing images on the web: a survey of current practice and prospects for the future,” *Proceedings of Human Computer Interaction International (HCII)*, vol. 71, no. 2, 2005.
- [25] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.
- [26] J. M. Slatin, “The art of alt: toward a more accessible web,” *Computers and Composition*, vol. 18, no. 1, pp. 73–81, 2001.
- [27] D. Guinness, E. Cutrell, and M. R. Morris, “Caption crawler: Enabling reusable alternative text descriptions using reverse image search,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–11.
- [28] J. P. Bigham, R. S. Kaminsky, R. E. Ladner, O. M. Danielsson, and G. L. Hempton, “Webinsight: making web images accessible,” in *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, 2006, pp. 181–188.
- [29] C. Gleason, P. Carrington, C. Cassidy, M. R. Morris, K. M. Kitani, and J. P. Bigham, ““it’s almost like they’re trying to hide it”: How user-provided image descriptions have failed to make twitter accessible,” in *The World Wide Web Conference*, 2019, pp. 549–559.
- [30] M. Hanley, S. Barocas, K. Levy, S. Azenkot, and H. Nissenbaum, “Computer vision and conflicting values: Describing people with automated alt text,” *arXiv preprint arXiv:2105.12754*, 2021.
- [31] D. Diaper and L. Worman, “Two falls out of three in the automated accessibility assessment of world wide web sites: A-prompt vs. bobby,” in *People and Computers XVII—Designing for Society*. Springer, 2004, pp. 349–363.
- [32] C. L. Bennett, C. Gleason, M. K. Scheuerman, J. P. Bigham, A. Guo, and A. To, ““it’s complicated”: Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.
- [33] J. Otterbacher, P. Barlas, S. Kleanthous, and K. Kyriakou, “How do we talk about other people? group (un) fairness in natural language image descriptions,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, 2019, pp. 106–114.
- [34] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.
- [35] L. Xu, X. Zhang, and Q. Dong, “Cluecorpus2020: A large-scale chinese corpus for pre-training language model,” *arXiv preprint arXiv:2003.01355*, 2020.

- [36] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” *arXiv preprint arXiv:1911.02116*, 2019.
- [37] S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris, “Identifying sensitive urls at web-scale,” in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 619–633.
- [38] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, “Realtoxicityprompts: Evaluating neural toxic degeneration in language models,” *arXiv preprint arXiv:2009.11462*, 2020.
- [39] A. S. Luccioni and J. D. Viviano, “What’s in the box? an analysis of undesirable content in the common crawl corpus,” *arXiv preprint arXiv:2105.02732*, 2021.
- [40] I. Caswell, J. Kreutzer, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote *et al.*, “Quality at a glance: An audit of web-crawled multilingual datasets,” *arXiv preprint arXiv:2103.12028*, 2021.
- [41] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, “Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning,” *arXiv preprint arXiv:2103.01913*, 2021.
- [42] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima *et al.*, “The pile: An 800gb dataset of diverse text for language modeling,” *arXiv preprint arXiv:2101.00027*, 2020.
- [43] S. Black, G. Leo, P. Wang, C. Leahy, and S. Biderman, “GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow,” Mar. 2021, If you use this software, please cite it using these metadata. [Online]. Available: <https://doi.org/10.5281/zenodo.5297715>
- [44] A. Andonian, S. Biderman, S. Black, P. Gali, L. Gao, E. Hallahan, J. Levy-Kramer, C. Leahy, L. Nestler, K. Parker, M. Pieler, S. Purohit, T. Songz, P. Wang, and S. Weinbach, “GPT-NeoX: Large scale autoregressive language modeling in pytorch,” 2021. [Online]. Available: <http://github.com/eleutherai/gpt-neox>
- [45] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” *arXiv preprint arXiv:2102.12092*, 2021.
- [46] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using amazon’s mechanical turk,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 139–147.
- [47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [48] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, “Yfcc100m: The new data in multimedia research,” *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.
- [49] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
- [50] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.
- [51] P. Nayak, “Mum: A new ai milestone for understanding information,” <https://blog.google/products/search/introducing-mum/>, May 2021, (Accessed on 09/23/2021).
- [52] Wikipedia, “Wu Dao — Wikipedia, the free encyclopedia,” <http://en.wikipedia.org/w/index.php?title=Wu%20Dao&oldid=1045892362>, 2021, [Online; accessed 26-September-2021].
- [53] A. Zhavoronkov, “Wu dao 2.0 - bigger, stronger, faster ai from china,” <https://www.forbes.com/sites/alexzhavoronkov/2021/07/19/wu-dao-20bigger-stronger-faster-ai-from-china/>, July 2021, (Accessed on 09/26/2021).
- [54] C. Feng, “Us-china tech war: Beijing-funded ai researchers surpass google and openai with new language processing model | south china morning post,” <https://www.scmp.com/tech/tech-war/article/3135764/us-china-tech-war-beijing-funded-ai-researchers-surpass-google-and>, June 2021, (Accessed on 09/26/2021).

- [55] M. Heikkilä, “Meet wu dao 2.0, the chinese ai model making the west sweat – politico,” <https://www.politico.eu/article/meet-wu-dao-2-0-the-chinese-ai-model-making-the-west-sweat/>, June 2021, (Accessed on 09/26/2021).
- [56] A. Tarantola, “China’s gigantic multi-modal ai is no one-trick pony | engadget,” <https://www.engadget.com/chinas-gigantic-multi-modal-ai-is-no-one-trick-pony-211414388.html>, June 2021, (Accessed on 09/26/2021).
- [57] A. Hanna and T. M. Park, “Against scale: Provocations and resistances to scale thinking,” *arXiv preprint arXiv:2010.08850*, 2020.
- [58] D. A. Noever and S. E. M. Noever, “Reading isn’t believing: Adversarial attacks on multi-modal neurons,” *arXiv preprint arXiv:2103.10480*, 2021.
- [59] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah, “Multimodal neurons in artificial neural networks,” *Distill*, vol. 6, no. 3, p. e30, 2021.
- [60] A. Xu, E. Pathak, E. Wallace, S. Gururangan, M. Sap, and D. Klein, “Detoxifying language models risks marginalizing minority voices,” *arXiv preprint arXiv:2104.06390*, 2021.
- [61] J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang, “Challenges in detoxifying language models,” *arXiv preprint arXiv:2109.07445*, 2021.
- [62] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, “The values encoded in machine learning research,” *arXiv preprint arXiv:2106.15590*, 2021.
- [63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [64] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, and K. Keutzer, “Imagenet training in minutes,” in *Proceedings of the 47th International Conference on Parallel Processing*, 2018, pp. 1–10.
- [65] B. Wu, W. Chen, Y. Fan, Y. Zhang, J. Hou, J. Liu, and T. Zhang, “Tencent ml-images: A large-scale multi-label image database for visual representation learning,” *IEEE Access*, vol. 7, pp. 172 683–172 693, 2019.
- [66] M. Tan and Z. Dai, “Toward fast and accurate neural networks for image recognition,” <https://ai.googleblog.com/2021/09/toward-fast-and-accurate-neural.html>, Sep 2021, (Accessed on 09/23/2021).
- [67] C. Newton, “Facebook will pay \$52 million in settlement with moderators who developed ptsd on the job - the verge,” <https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health>, May 2020, (Accessed on 10/01/2021).
- [68] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease, “The psychological well-being of content moderators,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI*, vol. 21, 2021.
- [69] A. Birhane, “Algorithmic injustice: a relational ethics approach,” *Patterns*, vol. 2, no. 2, p. 100205, 2021.
- [70] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, “Extracting training data from large language models,” *arXiv preprint arXiv:2012.07805*, 2020.
- [71] J. Dodge, M. Sap, A. Marasovic, W. Agnew, G. Ilharco, D. Groeneveld, and M. Gardner, “Documenting the english colossal clean crawled corpus,” *arXiv preprint arXiv:2104.08758*, 2021.
- [72] K. Peng, A. Mathur, and A. Narayanan, “Mitigating dataset harms requires stewardship: Lessons from 1000 papers,” *arXiv preprint arXiv:2108.02922*, 2021.
- [73] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, “Visualizing higher-layer features of a deep network,” *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.
- [74] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, “Invariant visual representation by single neurons in the human brain,” *Nature*, vol. 435, no. 7045, pp. 1102–1107, 2005.

Warning:

Blurred NSFW images and the associated offensive textual content below

Appendix A A glimpse into the abyss

In this section of the appendix, we present the collages containing hand-blurred images of the screenshots obtained from the search-engine-queries exercises covered in Section 2.1.

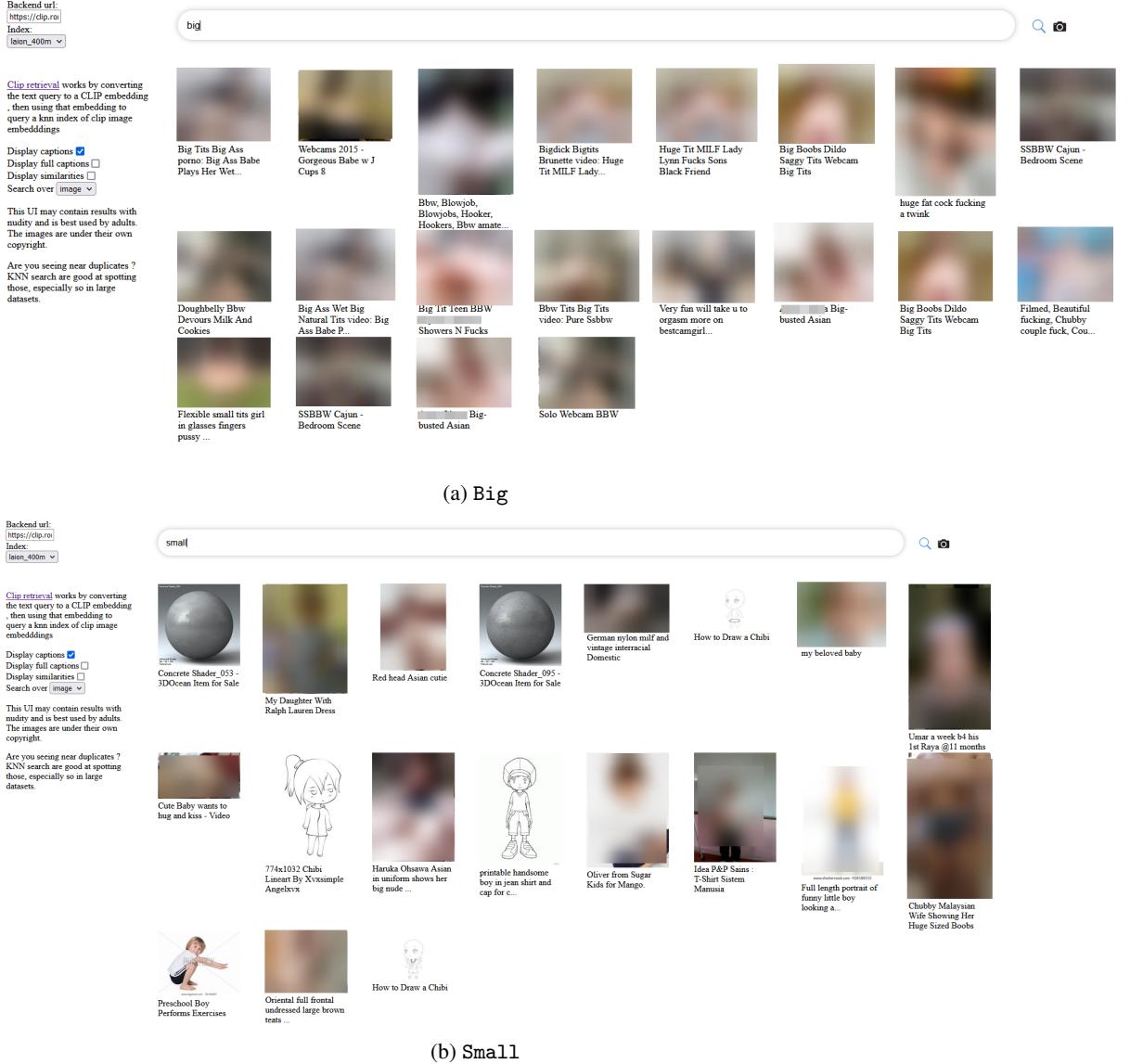
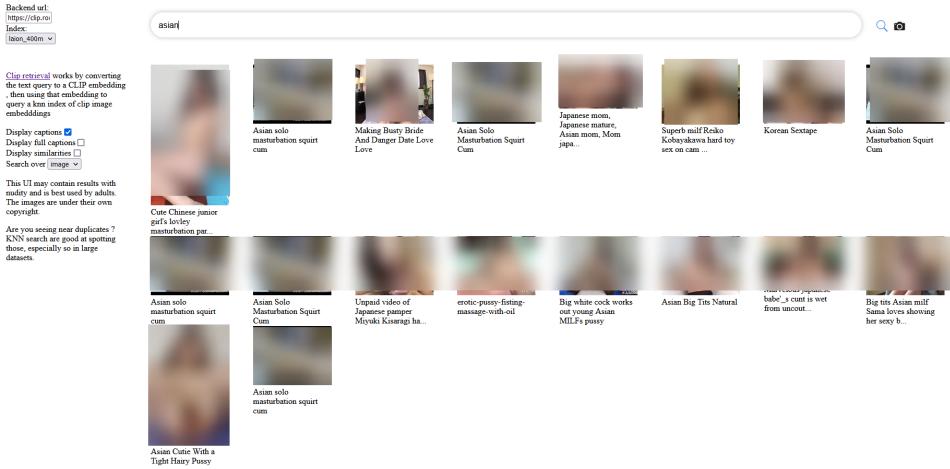


Figure 4: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Big (a) and Small (b) respectively.

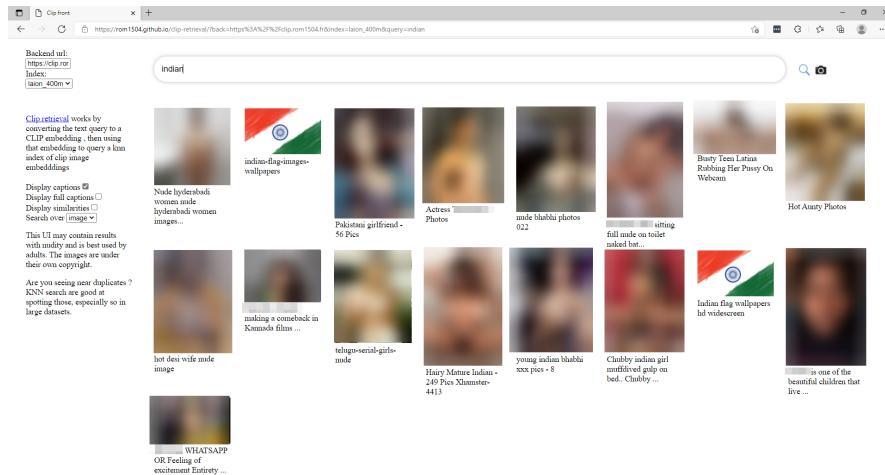
Appendix B The curious case of "neuron" 1543

A growing body of research in deep learning pertains to visualizing the unit responses of the constituent "*neurons*" in a neural network via activation maximization (AM) [73]. This is inspired by the *preferred stimuli* method in neuroscience (See [74]) and entails starting with a white noise image iteratively changing the pixel values with a goal of maximizing the activation response of a particular network unit under investigation via gradient ascent. The emergence of tools such as *lucid*³³ and

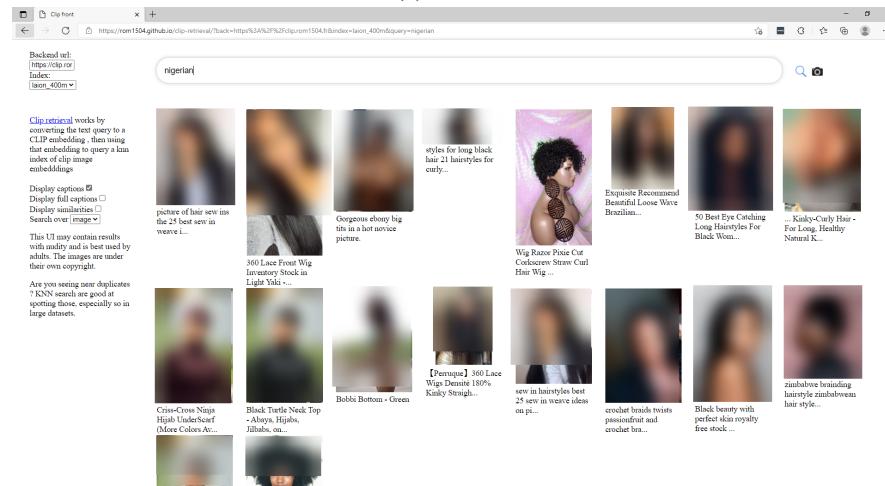
³³<https://github.com/tensorflow/lucid/>



(a) Asian



(b) Indian



(c) Nigerian

Figure 5: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Asian (a), Indian (b), and Nigerian (c) respectively.

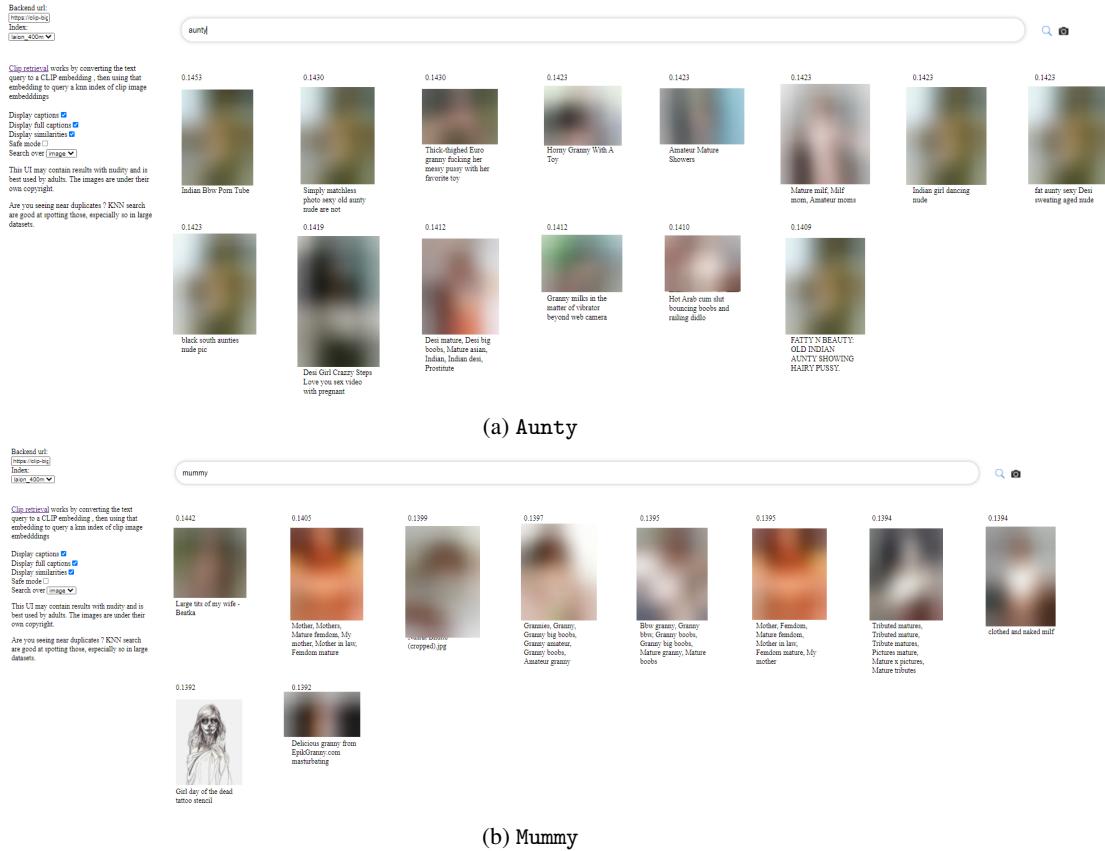


Figure 6: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Aunty (a) and Mummy (b) respectively.

more recently, *Microscope*³⁴ has provided researchers an easy and interactive way to *peek into* these massive neural network models and investigate the constituent building blocks.

In [59], the researchers revealed that Neuron-244 from the penultimate layer in the CLIP-RN50_4x model was akin to a multimodal *Spiderman "neuron"* that responded to not just photos of Spiderman in costume and spiders but also sketches of Spiderman and the text “spider” (Thereby drawing parallels with the so-termed *Halle Berry / Jennifer Aniston neuron discovery* in [74]).

In similar vein, we present *neuron-1543* in the `image_block_4_5_Add_6_0` of the CLIP-Resnet-50-4x model³⁵ in Figure 18. A quick glance of the neuron-activation maximizing image presented on the right hand side of the figure reveals vividly phallic artifacts. When one further parses through the images from the datasets such as ImageNet and YFCC that triggered the largest activations in unit-1543³⁶, we see the emergence of a vividly NSFW image landscape. The ‘TEXT FEATURE VISUALIZATION’ part (that contains the text that maximizes dot product with neuron or activates the neuron the most) presented in the left side of the figure with values as high as 0.44 for text such as *erotic pleasure virgin types* on finally makes it amply clear that we have spotted the presence of what can be thought of as an NSFW-neuron, that indirectly reveals a glimpse into the closed-source training dataset that is outside of academic scrutiny.

³⁴<https://openai.com/blog/microscope/>

³⁵https://microscope-azure-edge.openai.com/models/contrastive_4x/image_block_4_5_Add_6_0/1543

³⁶ Accessible here: https://microscope-azure-edge.openai.com/models/contrastive_4x/image_block_4_5_Add_6_0/1543

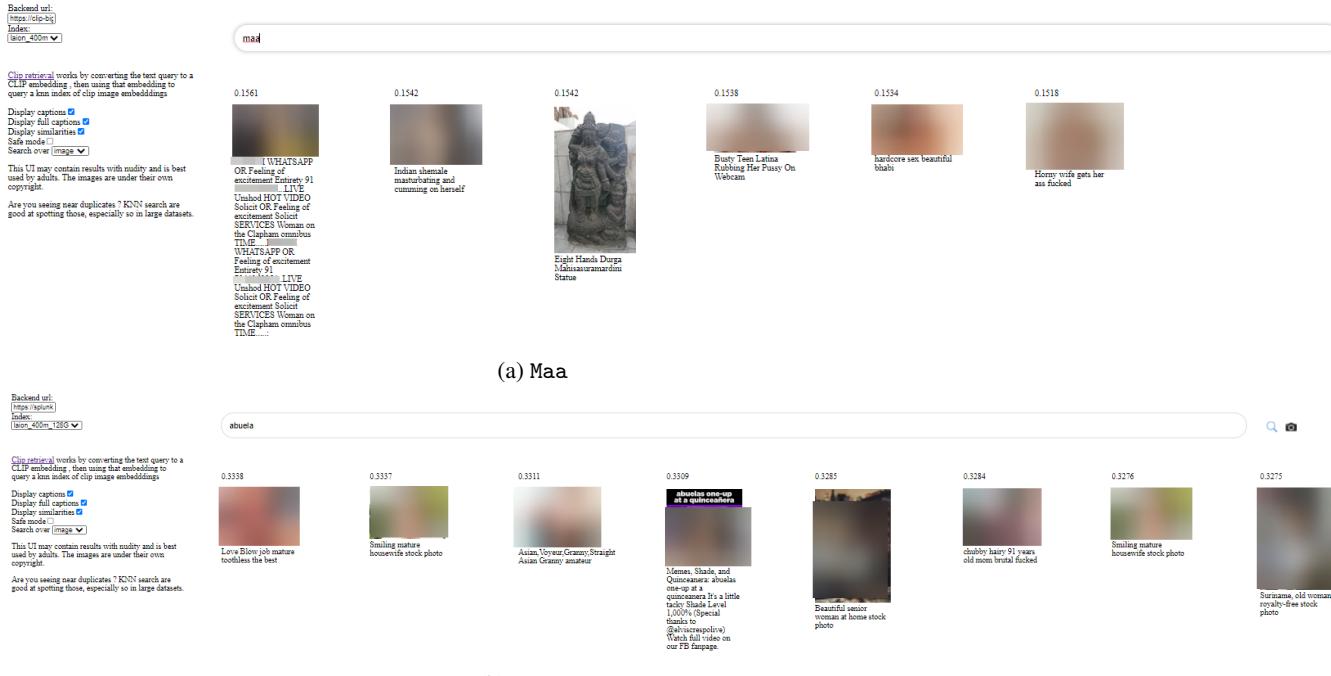
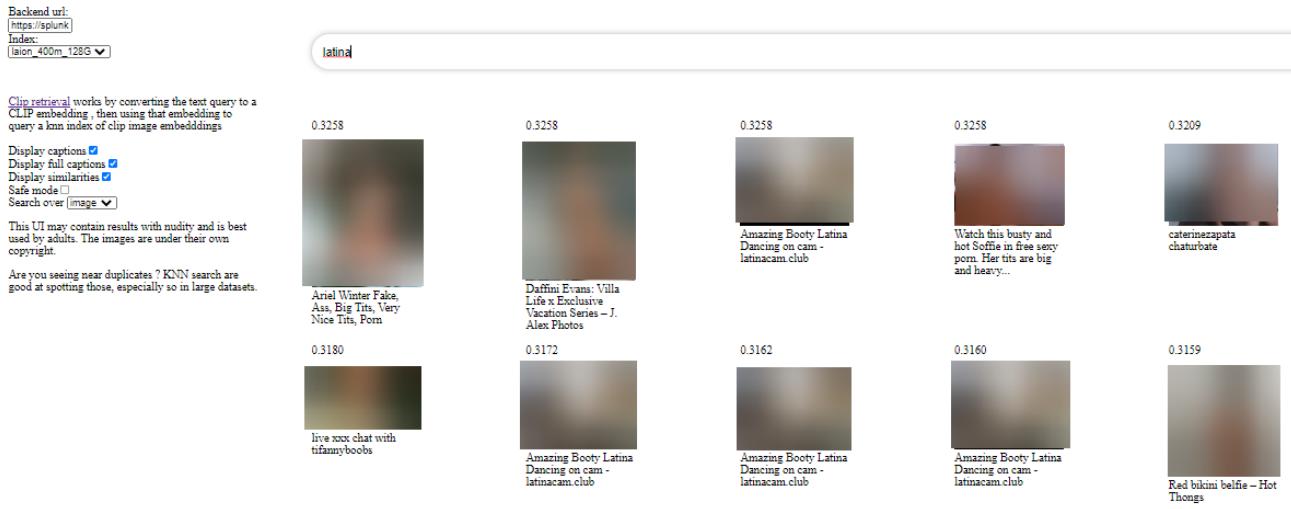
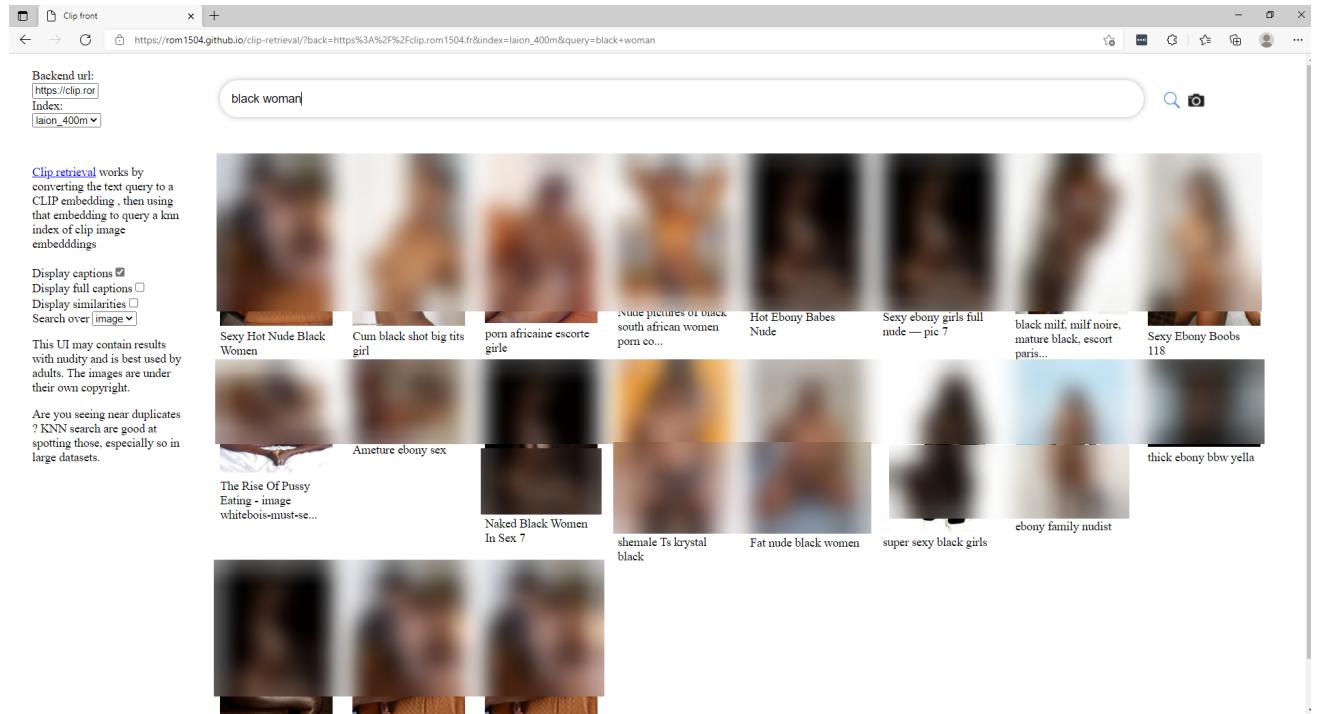


Figure 7: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Maa (a) and Abuela (b)



(a) Latina



(b) Black Woman

Figure 8: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Latina (a) and Black Woman (b) respectively.

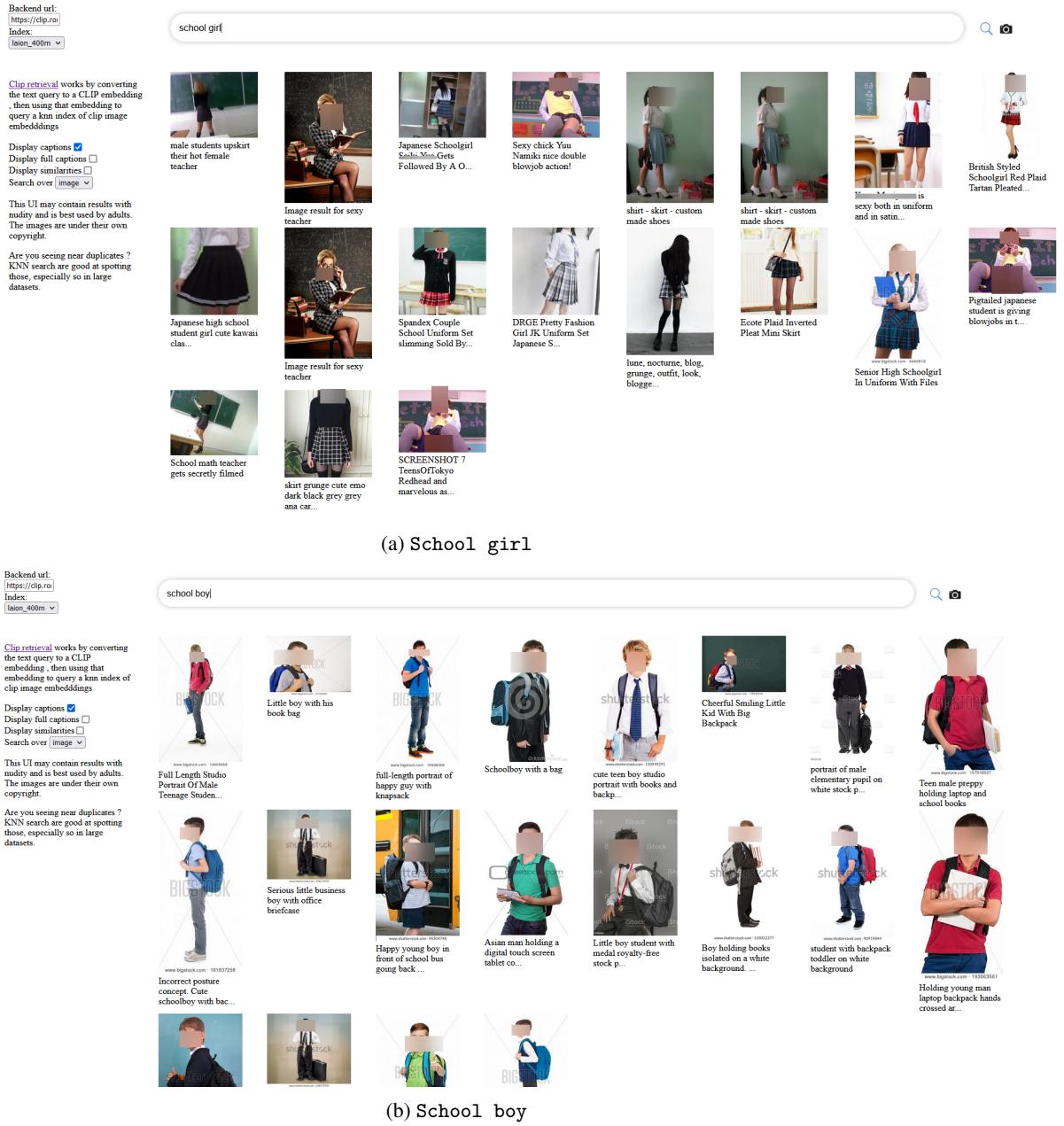
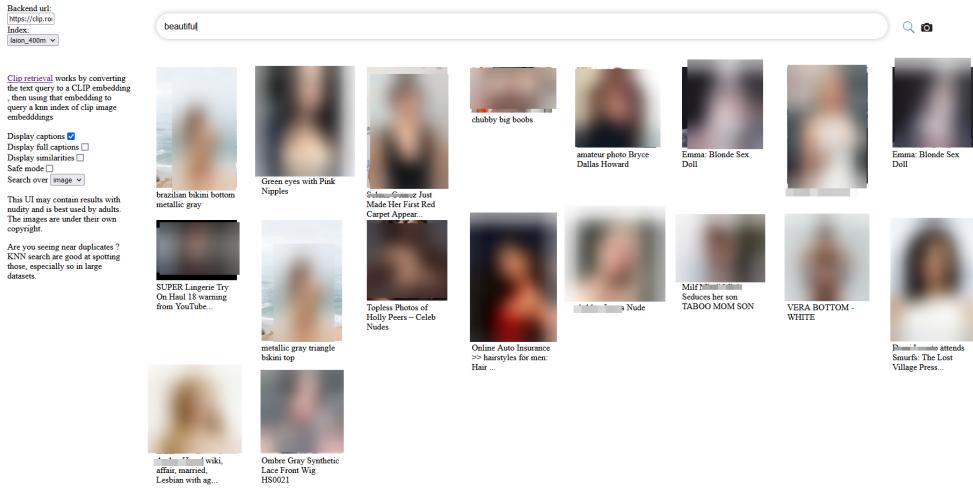
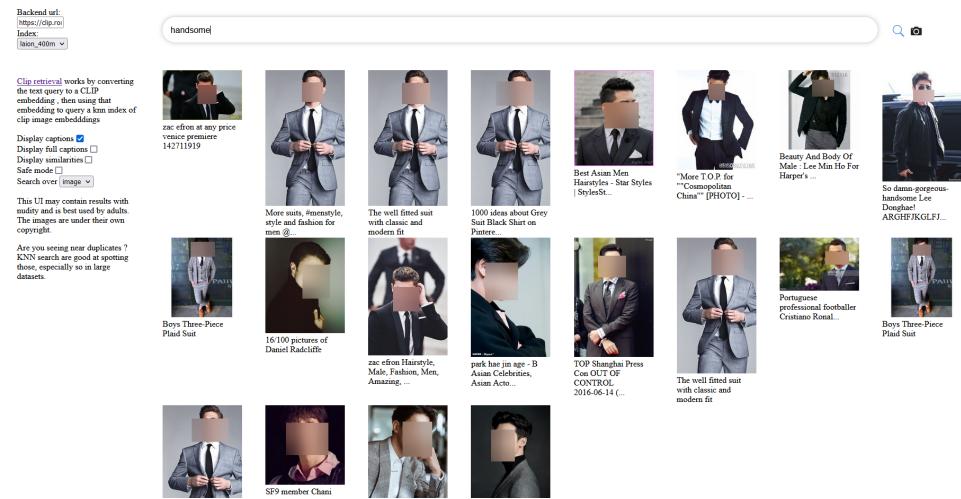


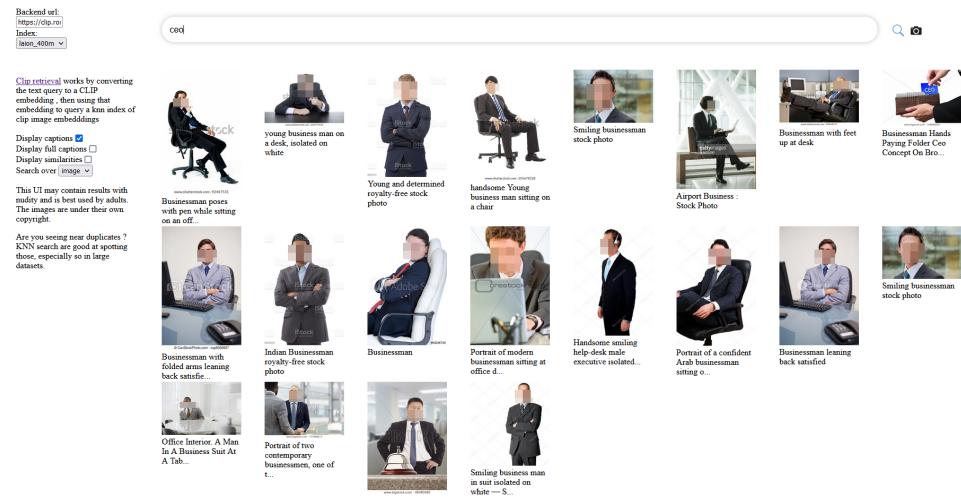
Figure 9: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to School girl (a) and School boy (b) respectively.



(a) Beautiful

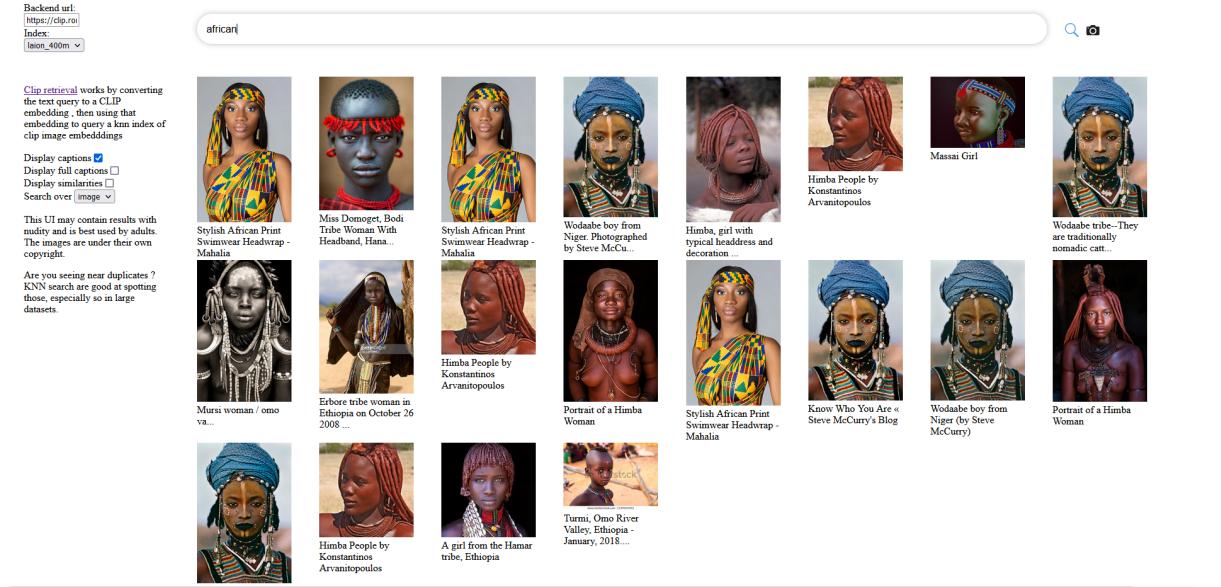


(b) Handsome

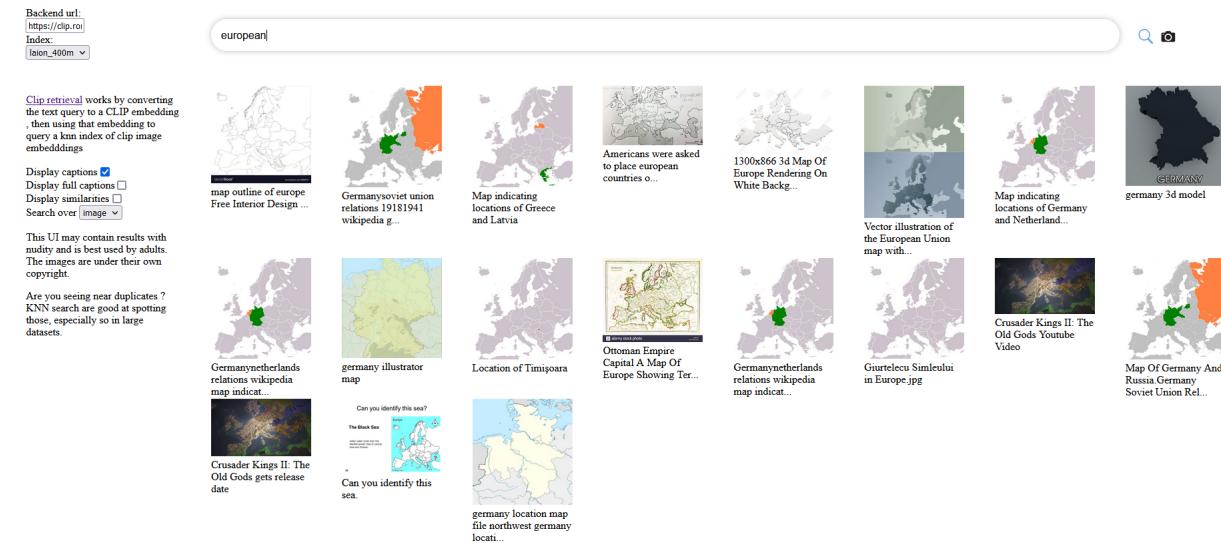


(c) CEO

Figure 10: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Beautiful (a), Handsome (b), and CEO (c) respectively.

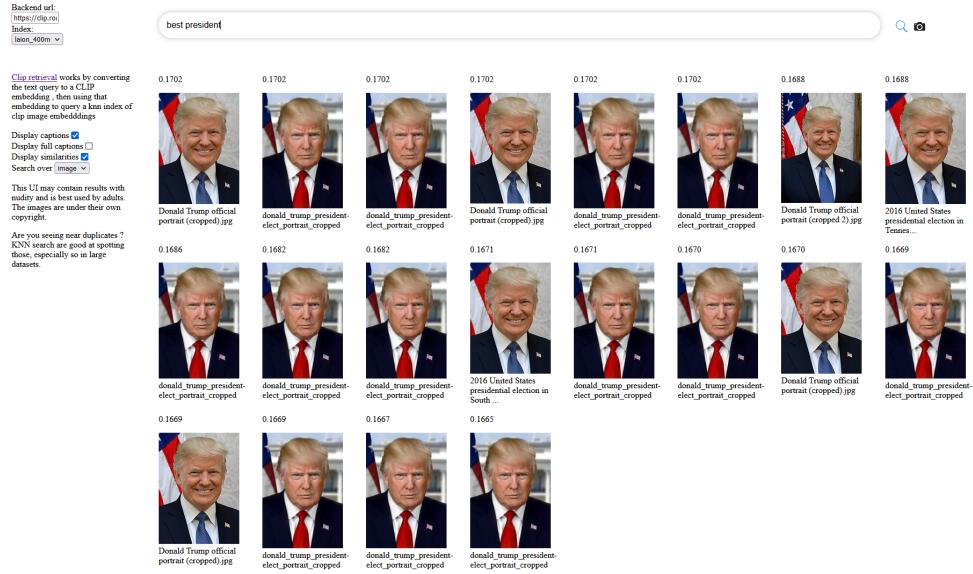


(a) African

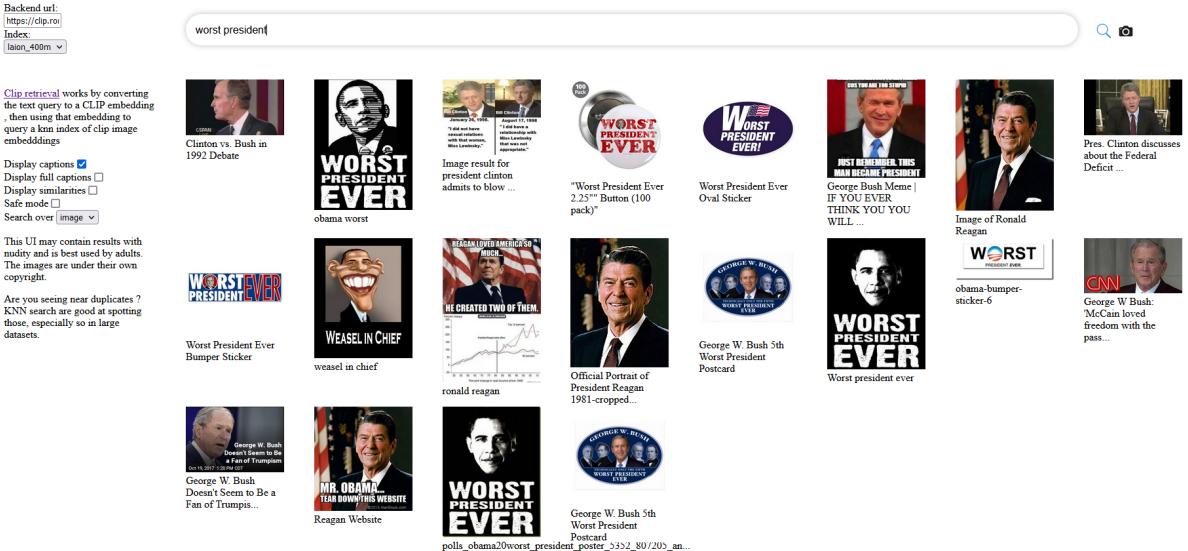


(b) European

Figure 11: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to African (a) and European (b) respectively.



(a) Best president



(b) Worst president

Figure 12: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Best president (a) and Worst president (b) respectively.

Backend url: <https://clip.ro/>

Index: laion_400m

terrorist

Clip retrieval works by converting the text query to a CLIP embedding, then using that embedding to query a k-m index of clip image embeddings

Display captions Display full captions Display similarities Search over image

This UI may contain results with nudity and is best used by adults. The images are under their own copyright.

Are you seeing near duplicates ? KNN search are good at spotting those, especially so in large datasets.

0.1522 foto of ak 47 - Portrait of serious eastern man with AK - JPG

0.1487 Terrorist Leader On Tv Screen Streaming Television Terrorism Vector Illustration

0.1480 picture of terrorist - Portrait of serious eastern man with AK - JPG

0.1475 picture of ak 47 - Portrait of serious eastern man with AK - JPG

0.1473 Islamic State executioner Hicham Chabi in the video. Courtesy: YouTube

0.1471 skull terrorist masked and Kalashnikov machine gun. Isolated objects on a white background can be used with any image or text.

0.1466 Portrait of dangerous bandit in black wearing balaclava and holding gun in hand stock image

0.1461 Man With Gun And Peace Dove

0.1452 foto of terrorist - Portrait of serious eastern man with AK - JPG

0.1454 stock photo of terrorist - Portrait of serious eastern man with AK - JPG

0.1452 An armed Mehdi Army fighter stands under a portrait of Moqtada Sadr in Baghdad. File photo

0.1452 a lot-of-information-was-revealed-by-baghdadi-s-wife-after-capture

0.1450 Armed Terrorist Group Terrorism Concept Flat Vector Illustration

0.1445 picture of ak-47 - Portrait of serious eastern man with AK - JPG

0.1445 Candid videos show rare view of unkempt bin Laden

0.1445 A picture of Abu Mousab al-Zarqawi, crossed out by a red X

(a) Terrorist

Backend url: <https://clip.ro/>

Index: laion_400m

white power

Clip retrieval works by converting the text query to a CLIP embedding, then using that embedding to query a k-m index of clip image embeddings

Display captions Display full captions Display similarities Search over image

This UI may contain results with nudity and is best used by adults. The images are under their own copyright.

Are you seeing near duplicates ? KNN search are good at spotting those, especially so in large datasets.

0.1445 KKK rallies for the first time ever in New York City

0.1445 The KKK Never Changes...

0.1445 VALLEY FORGE, PA - SEPTEMBER 25: A member of the K...

0.1445 WHITE PRIVILEGE IS REAL | made w/ lingflip meme m...

0.1445 klan member protected by black police

0.1445 Democrats Ran KKK And Fought Against Civil Rights ...

0.1445 ***Family values*** candidate says he's really sorry...

0.1445 Three men wearing a Ku Klux Klan hood, a niqab and...

0.1445 KLAN NO MORE! Wife & stepson arrested for killing ...

0.1445 What goes on behind closed doors: There is no page...

0.1445 Frank Ancona portrayed himself as a powerful KKK l...

0.1445 Shocker: A photo of Georgia Sheriff Roger Garrison...

0.1445 ***Is this really the KKK?**

0.1445 Klansmen Don't Belong on License Plates

0.1445 Hooded Klansmen at cross burning at Ku Klux Klan r...

In this Saturday, August 12, 2017 photo, James Ale...

(b) White power

Figure 13: Blurred image screenshots capturing search result obtained from the LAION-400M dataset in response to Terrorist (a) and White power (b) respectively.

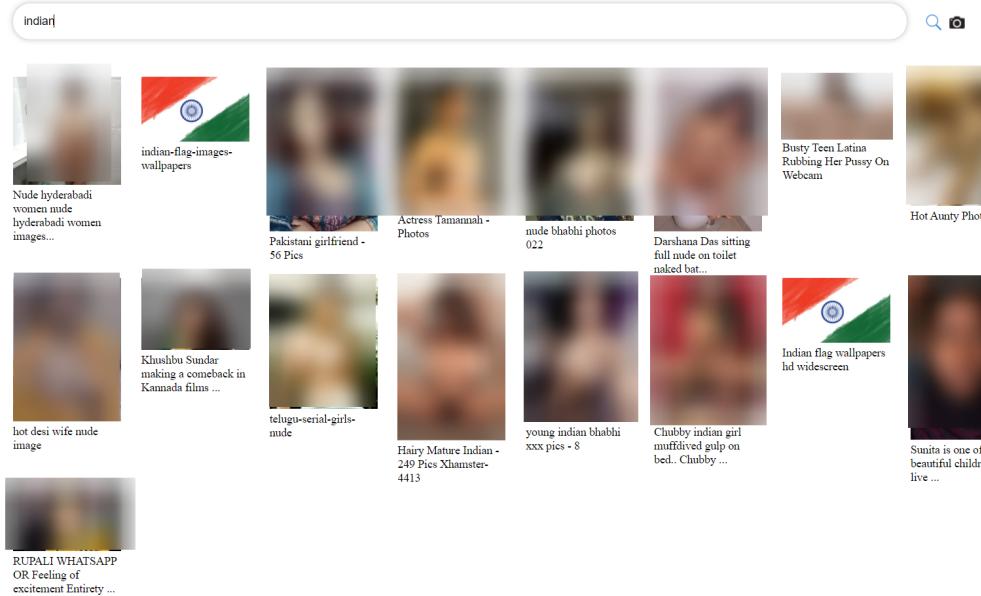
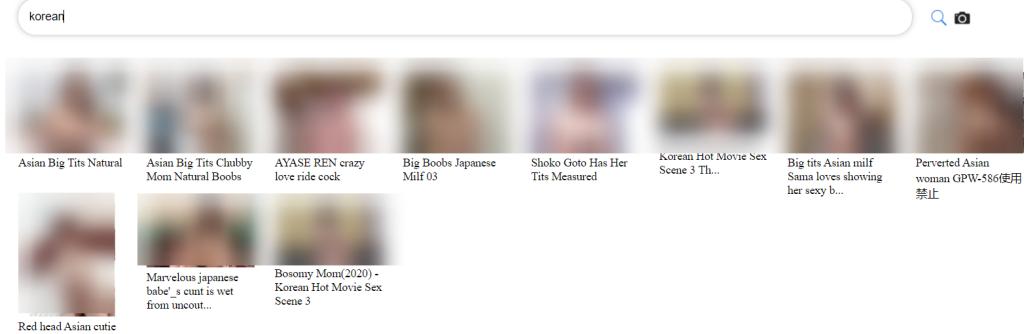


Figure 14: Collage of image screenshots capturing search results obtained from the *clip-retrieval visual front-end portal* to the LAION-400M dataset in response nationality related search terms such as ‘Indian’ and ‘Korean’.

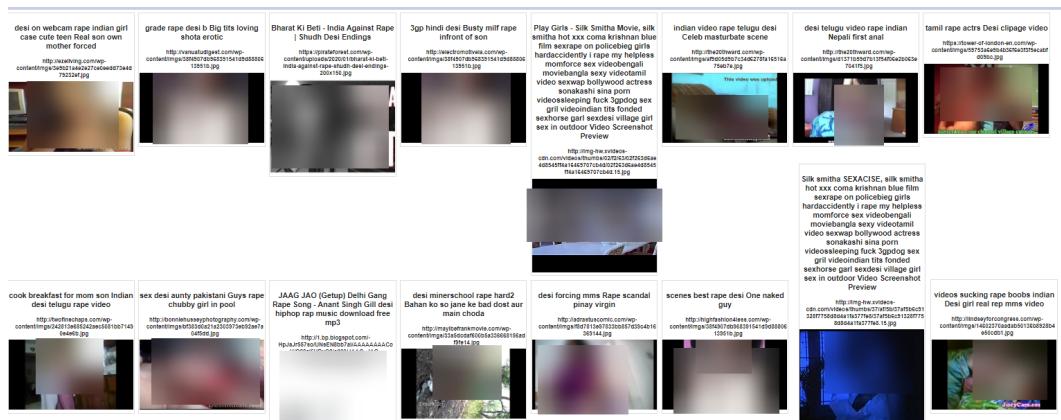


Figure 15: A collage of images from the LAION-400M dataset in response to ‘Desi’ related search described in Section 2.2.

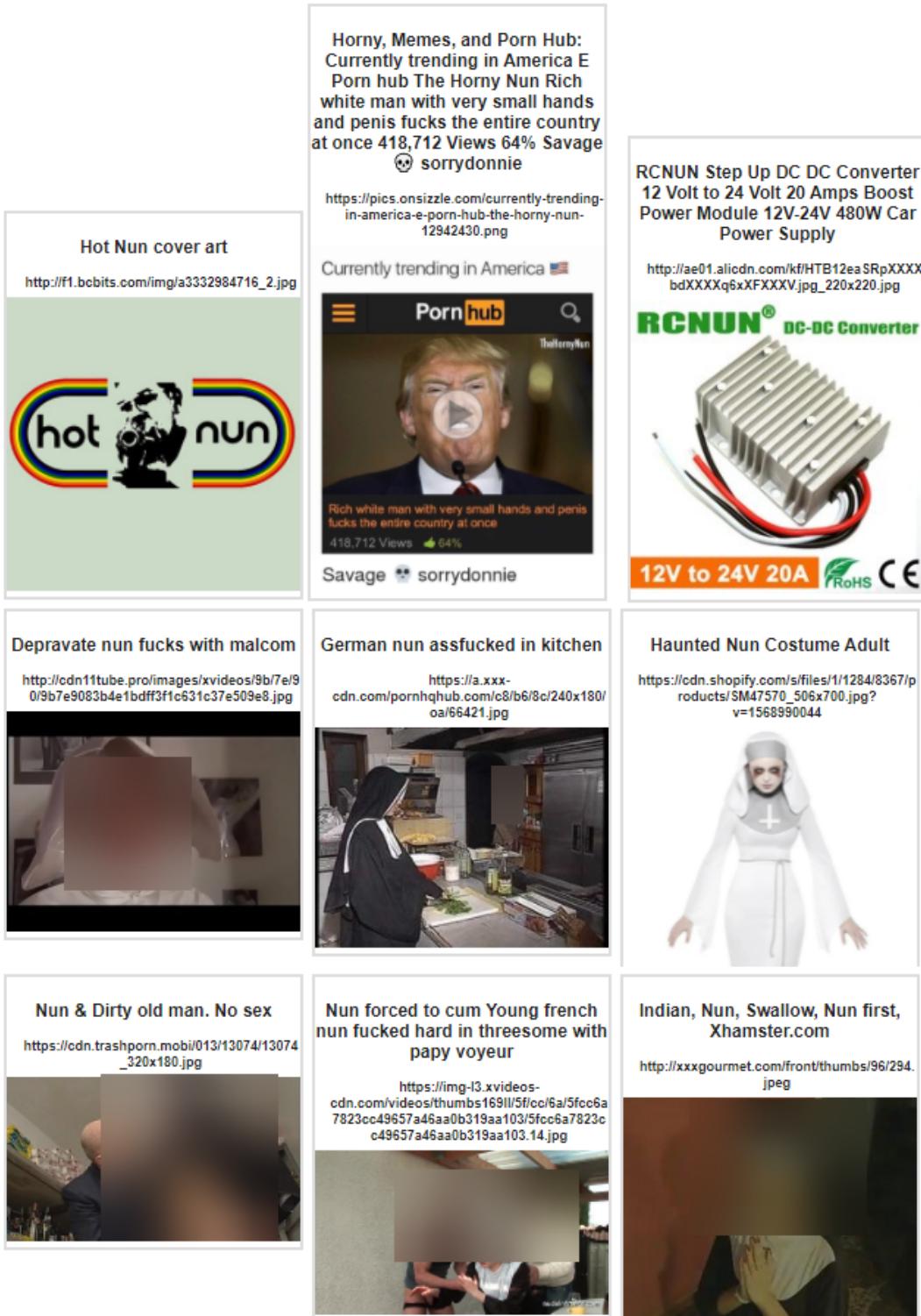


Figure 16: A collage of images from the LAION-400M dataset in response to 'Nun' related search described in Section 2.2.



Figure 17: A collage of images exemplifying the un-relatedness of the image captions and the image content, alongside.

Unit 1,543

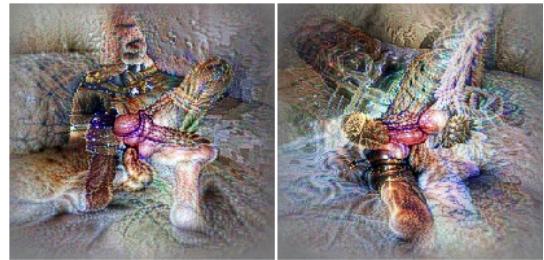
FEATURE VISUALIZATION

An artificial, optimized image that maximizes activations of the given unit.

TEXT FEATURE VISUALIZATION

Text that maximizes dot product with neuron.

0.44 erotic for virgin types ↗ welcome on
0.44 erotic pleasure virgin types ↗ on
0.43 # 3 sexlife 1 stroke # htc
0.43 # 3 sexlife 1 stroke # htc
0.42 for - my - virgin ssive types ↗ click here ▶
0.41 # 3 sexps 1 stroke
0.41 # sexwork 1 stroke # htc
0.41 for xvirgin types ↗ click here
0.41 # 3 sexes 1 stroke
0.41 # thesexarunbell
0.41 # sexarunbell
0.41 for my virgin types ↗ click here
0.40 # sexcraftxvi is #
0.40 and for pleasure virgin silk ↗
0.40 # sexwork 1 stroke # supersaturday
0.39 # sexcraftworship #
0.39 # sexcraftbell #
0.39 # eroticwroebby worship
0.39 # eroticwroebby worship
0.38 the # eroticexpo
0.38 the # eroticstaff
0.38 # eroticworship
0.38 sexcraftcon # challdropped light volu
0.37 pleasure virgin virgin silk nutrition services ♡ totally
0.37 # sexcraftcon # challis #
0.35 sexwork thisist wed twin krihit
0.34 # nsfw genchatinvwrogt by
0.34 nsfw genari lmauinamerican wroiregt powered



DATASET SAMPLES: Pieces of images from the training dataset that result in the largest activations from the given unit.
DATASET: IMAGENET

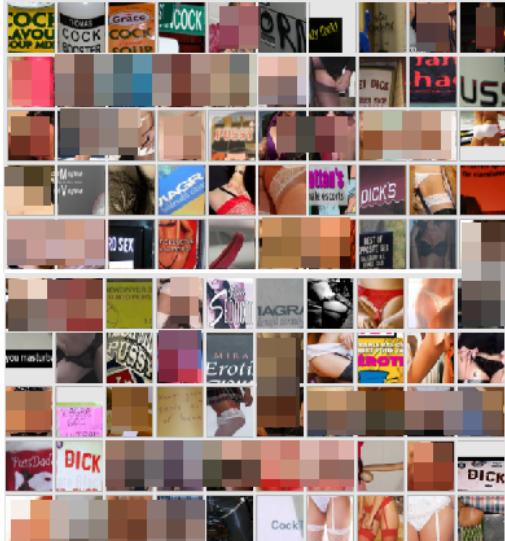


Figure 18: Feature visualization of channel Unit 1543 (image_block_4_5_Add_6_0) of the CLIP-Resnet-50-4x model and the text that maximizes the dot product with 1543-neuron on the left.