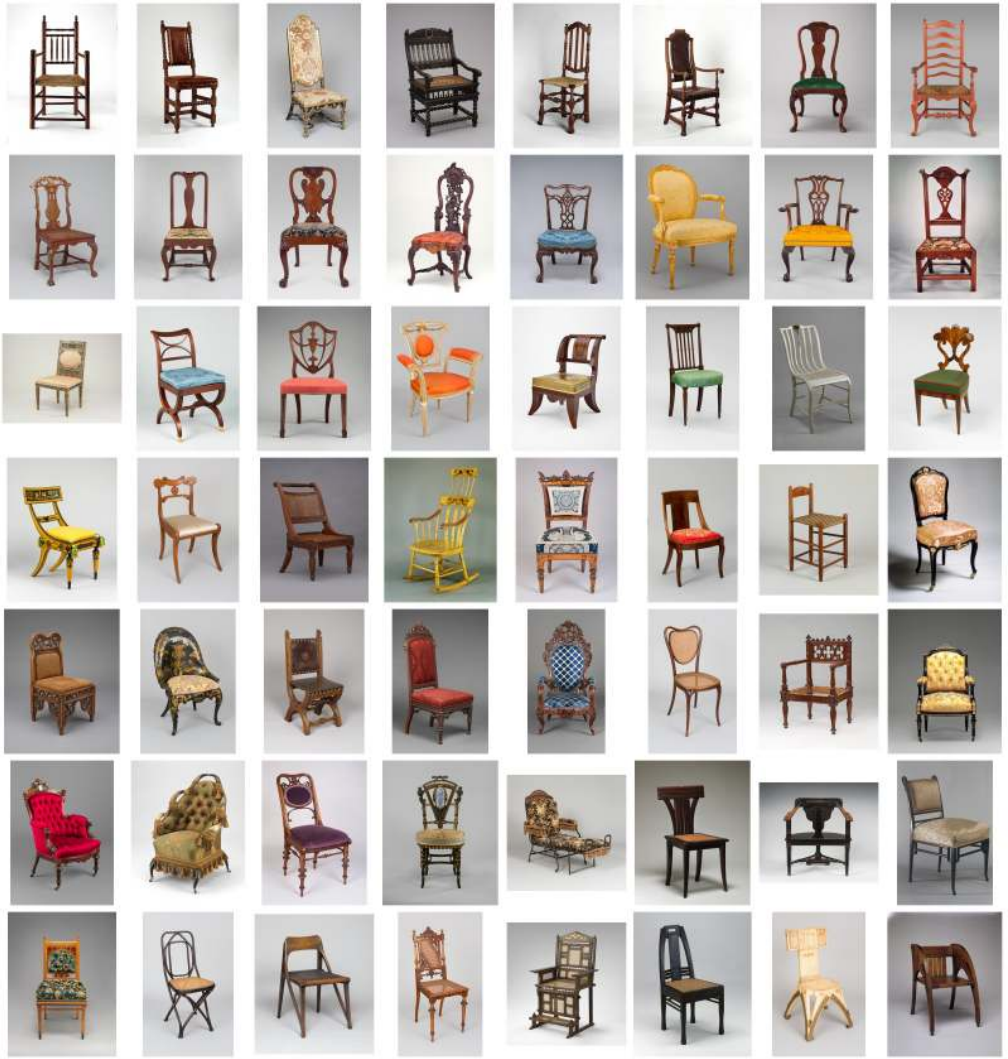
Machine Learning for Design, Week 4

Data and Embeddings

56 chairs from two museums, made between 1650 and 1904



The Metropolitan Museum of Art and the Art Institute of Chicago, open access (CC0)

What goes into your training data, and what does the model learn from it?

Today

1. Choosing the images
2. Embeddings: how CLIP sees them
3. Mapping the set
4. Datasheets
5. Whose work trains the model
6. Discussion
7. Workshop: collect and map your set

Choosing the 56

Art Institute of Chicago: **28** of 36 public-domain chairs

The Met: **28** of about 330 chairs I looked at

Left out, for example:



black-and-white photo



three chairs in one photo



teal backdrop



red backdrop

The Met: McKinley Armchair (1894–96), Armchair (1750–75), Reclining Armchair (after 1866).
Art Institute of Chicago: Rocking Chair, Jacob and Josef Kohn (c. 1881). All public domain.

Dates of the chairs



Of 178 Met chair records dated 1875 to 1950, **149** were not public domain.

55 chairs with a year; one is dated only “19th century”

Two Met records


```
"title": "Turned Armchair"
"objectDate": "1650–1700"
"isPublicDomain": true
"primaryImage": "https://images.metmuseum.org/…"
```





```
"title": "Chair"
"artistDisplayName": "Gilbert Rohde"
"objectDate": "1935–36"
"isPublicDomain": false
"primaryImage": ""
```


no image in the open record

The Met Collection API, objects 146 (Turned Armchair) and 481539 (Chair, Gilbert Rohde)

Automatic captions and museum records



“there is a wooden chair with a woven seat and a cane”

Turned Armchair, American, 1650–1700, maple



“there is a wooden chair with a woven seat and a clock on the back”

Armchair, Coromandel Coast, India, c. 1690–1700, ebony, ivory and rattan



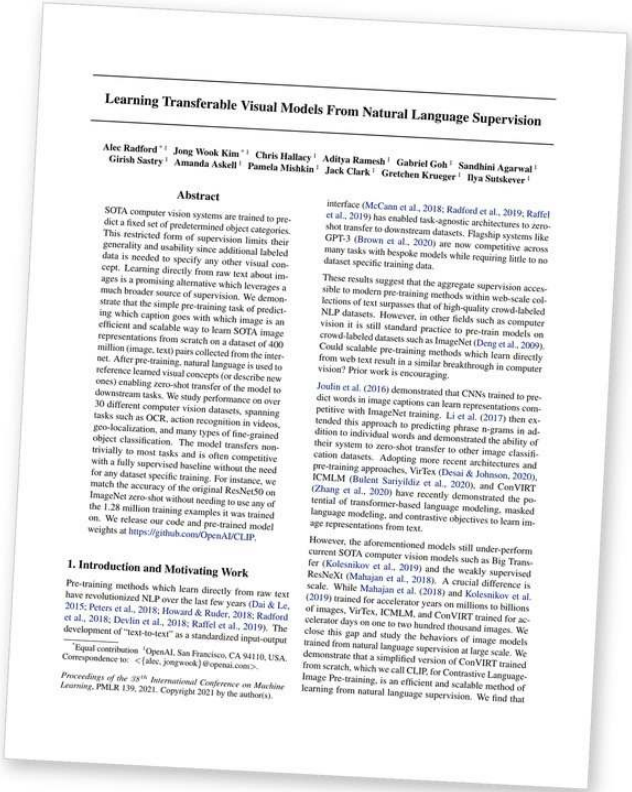
“there is a wooden rocking chair with a sunflower design on it”

Rocking Chair, Hudson and Brooks, ca. 1823, maple, pine, mahogany

Black: BLIP (Salesforce/blip-image-captioning-large), first try. Gray: the museum's record.

2. Embeddings: how CLIP sees them

Today's reading: Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML 2021



Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

SOTA computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study performance on over 30 different computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

1. Introduction and Motivating Work

Pre-training methods which learn directly from raw text have revolutionized NLP over the last few years (Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2019). The development of “text-to-text” as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

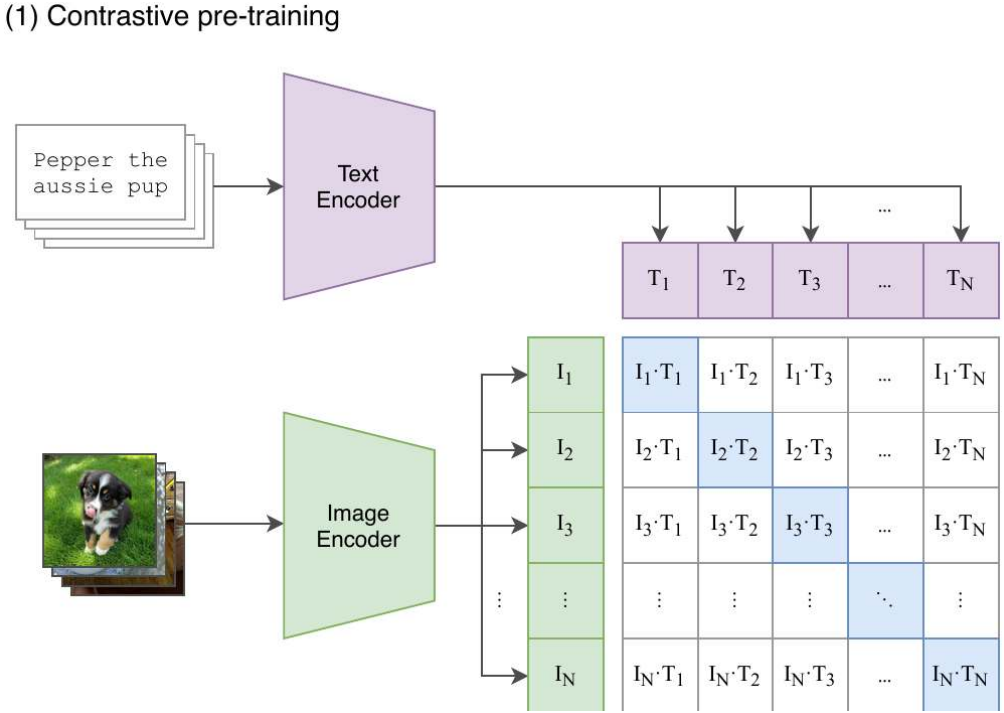
Joulin et al. (2016) demonstrated that CNNs trained to predict words in image captions can learn representations competitive with ImageNet training. Li et al. (2017) then extended this approach to predicting phrase n-grams in addition to individual words and demonstrated the ability of their system to zero-shot transfer to other image classification datasets. Adopting more recent architectures and pre-training approaches, VirTex (Desai & Johnson, 2020), ICMLM (Bulent Sariyildiz et al., 2020), and ConVIRT (Zhang et al., 2020) have recently demonstrated the potential of transformer-based language modeling, masked language modeling, and contrastive objectives to learn image representations from text.

However, the aforementioned models still under-perform current SOTA computer vision models such as Big Transfer (Kolesnikov et al., 2019) and the weakly supervised ResNeXt (Mahajan et al., 2018). A crucial difference is scale. While Mahajan et al. (2018) and Kolesnikov et al. (2019) trained for accelerator years on millions to billions of images, VirTex, ICMLM, and ConVIRT trained for accelerator days on one to two hundred thousand images. We close this gap and study the behaviors of image models trained from natural language supervision at large scale. We demonstrate that a simplified version of ConVIRT trained from scratch, which we call CLIP, for Contrastive Language-Image Pre-training, is an efficient and scalable method of learning from natural language supervision. We find that

[*]Equal contribution [1]OpenAI, San Francisco, CA 94110, USA. Correspondence to: <{alec, jongwook}@openai.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

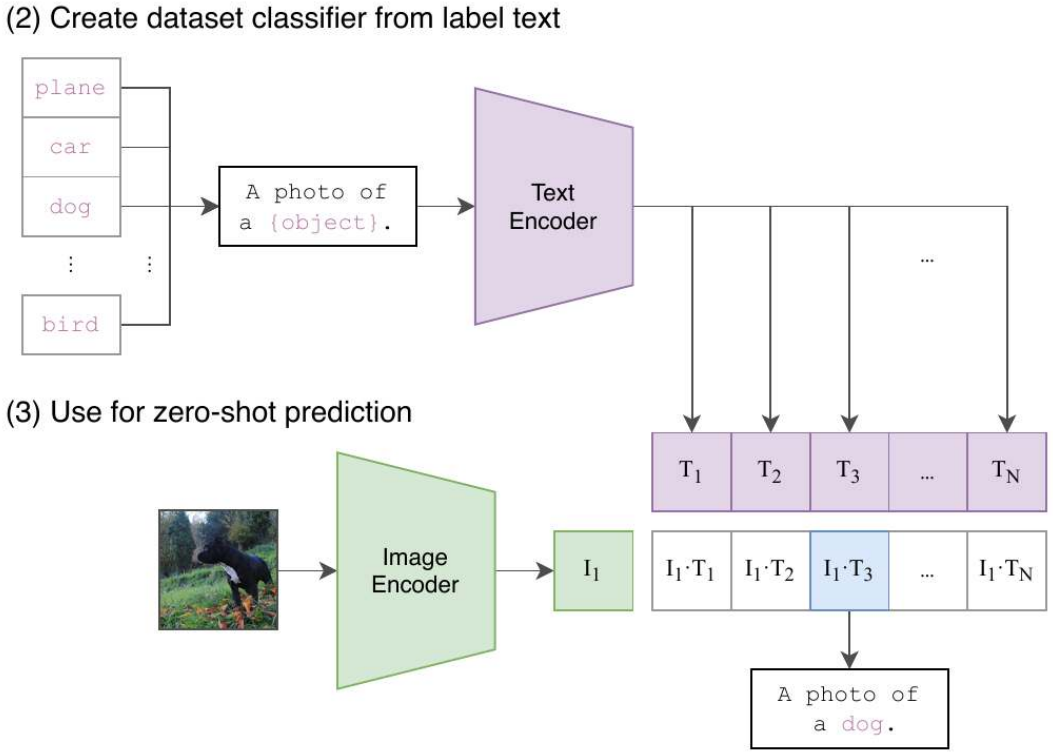
CLIP



Trained on **400 million** (image, text) pairs from the internet

Radford et al., ICML 2021, Fig. 1

Zero-shot labels



Any list of words becomes a classifier, with no extra training.

Radford et al., ICML 2021, Fig. 1 (2) and (3)

Zero-shot on our chairs: 21 of 43

	CLIP said: side chair	armchair	rocking chair
title: side chair	**12**	12	4
title: armchair	2	**8**	4
title: rocking chair	0	0	**1**

12 of 28 side chairs called armchairs



Turned Armchair
CLIP: rocking chair



Side Chair
CLIP: armchair



Armchair
CLIP: rocking chair



Side Chair
CLIP: armchair

CLIP ViT-B/32, “a photo of a side chair / armchair / rocking chair”, against the museum titles; 13 chairs titled only “Chair” or similar left out

Other kinds of images



Sketches: **25.2%** for a model trained on ImageNet photos, **60.2%** for zero-shot CLIP

Radford et al., ICML 2021, Fig. 7 (right): accuracy on each whole dataset, with bananas as the example class

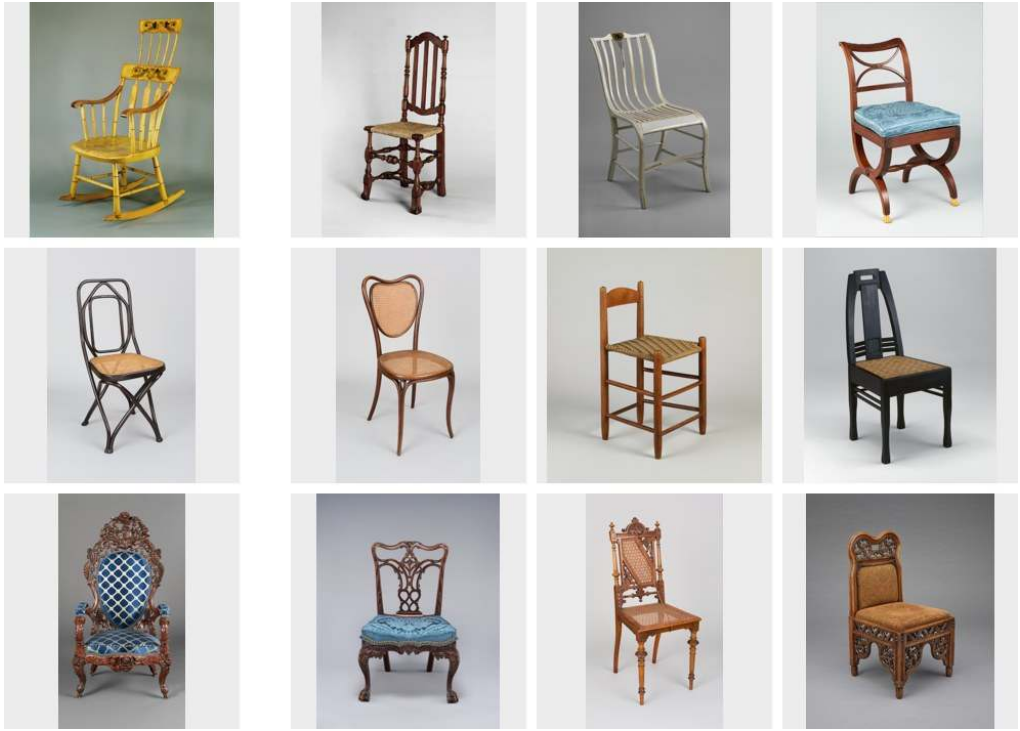
One chair as 512 numbers



−0.039, 0.002, 0.030, −0.010, 0.032, 0.028, …

Turned Armchair (met-146), CLIP ViT-B/32 image embedding, first 6 of 512

Closest chairs by CLIP



Left: a chair from the set

Right: its three closest chairs

CLIP ViT-B/32 image embeddings, cosine similarity

Searching the chairs with words

carved



painted



woven seat



bent wood



Carved: **4 of 4**

The other three: **3 of 4** each

Top 4 of the 56 for “a photo of a carved chair” and the other three phrases, CLIP ViT-B/32

The 56 chairs on a CLIP map

Chairs with
similar numbers
sit close together.



CLIP ViT-B/32, t-SNE to two dimensions

4. Datasheets

Today's reading: Gebru et al.,
“Datasheets for Datasets,” CACM 2021



Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

arXiv:1803.09010v8 [cs.DB] 1 Dec 2021

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases. Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice [1, 13, 24], hiring [19], critical infrastructure [11, 21], and finance [18]. Even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets [4, 5, 12]. For these and other reasons, the World Economic Forum suggests that all entities should document the provenance, creation, and use of machine learning datasets in order to avoid discriminatory outcomes [25].

Although data provenance has been studied extensively in the databases community [3, 8], it is rarely discussed in the machine learning community. Documenting the creation and use of datasets has received even less attention. Despite the importance of data to machine learning, there is currently no standardized process for documenting machine learning datasets.

To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, we propose that every

Authors' addresses: Timnit Gebru, Black in AI; Jamie Morgenstern, University of Washington; Briana Vecchione, Cornell University; Jennifer Wortman Vaughan, Microsoft Research; Hanna Wallach, Microsoft Research; Hal Daumé III, Microsoft Research; University of Maryland; Kate Crawford, Microsoft Research.

A question from the datasheet

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Gebru et al., “Datasheets for Datasets,” arXiv v8 (2021), Section 3.2, Composition

The datasheet in the paper's appendix



Datasheets for Datasets 15

Movie Review Polarity **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary [positive, negative]. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

[1] All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://www.lpdi.epfl.ch/publication/bo2005/bo-thumbs.pdf; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

Figure 1. An example "negative polarity" instance, taken from the file neg/cv452_tok-18650.txt.

those are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie ego-massia . i can forgive many things , but using some hackneyed , whacked-out , screwed-up * non * ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Fig. 1. Example datasheet for Pang and Lee's polarity dataset [22], page 1.

Seven sections:

Motivation

Composition

Collection process

Preprocessing, cleaning and labeling

Uses

Distribution

Maintenance

Gebru et al., arXiv v8 (2021), Appendix, Fig. 1: a datasheet for Pang and Lee's movie review polarity dataset

A datasheet for the chairs

Datasheet: museum chairs

Answers to the questions in Gebru et al., "Datasheets for Datasets" (CACM 2021), for the example set used in weeks 4 and 5 of Machine Learning for Design.

Motivation

- **Purpose:** a set of designs that every student can see, map, fine-tune on and share, so the course can show each step on the same images before students use their own.
- **Created by:** the course instructor, from two museums' open-access collections. No funding.

Composition

- **Instances:** 56 photographs, one chair in each, on a plain grey or white backdrop.
- **Sources:** 28 from The Metropolitan Museum of Art and 28 from the Art Institute of Chicago.
- **Sample, not the whole collection:** I looked at all 36 public-domain chairs with images at the Art Institute and about 330 chairs at the Met, and picked by hand across dates, makers and regions.
- **Each image has:** the museum, object number, title, maker, date, materials, credit line, licence and a link to the museum's record (credits.csv).
- **Missing:** chairs made after 1904, because most later chairs are not public domain, and most chairs made outside Europe and America. Of the rest, three chairs were made for European buyers (one in India, one in China, and one in the West Indies or Asia, as the record gives both), and one is from Africa.
- **People:** no people appear in the photographs.

Composition: a sample, picked by hand

Missing: chairs after **1904**, and most chairs made outside Europe and America

examples/museum-chairs/datasheet.md in the course notebooks repository

5. Whose work trains the model

Today's reading: Jiang et al., “AI Art and its Impact on Artists,” AIES 2023



AI Art and its Impact on Artists

Harry Jiang
Independent Researcher
Canada
hhj@alumni.cmu.edu

Anonymous Artist
Artist
USA

Deja Workman
Penn State University
USA
dqw5409@psu.edu

Lauren Brown
Artist
USA
labillustration@gmail.com

Mehtab Khan
Yale Law
USA
mehtab.khan@yale.edu

Alex Hanna
The Distributed AI Research Institute
USA
alex@dair-institute.org

Timnit Gebru
The Distributed AI Research Institute
USA
timnit@dair-institute.org

Jessica Cheng
Artist
Canada
chengelingart@gmail.com

Abhishek Gupta
Montréal AI Ethics Institute
Canada
abhishek@montrealethics.ai

Jonathan Flowers
California State University, Northridge
USA
johnathan.flowers@csun.edu

ABSTRACT

The last 3 years have resulted in machine learning (ML)-based image generators with the ability to output consistently higher quality images based on natural language prompts as inputs. As a result, many popular commercial "generative AI Art" products have entered the market, making generative AI an estimated $48B industry [125]. However, many professional artists have spoken up about the harms they have experienced due to the proliferation of large scale image generators trained on image/text pairs from the Internet. In this paper, we review some of these harms which include reputational damage, economic loss, plagiarism and copyright infringement. To guard against these issues while reaping the potential benefits of image generators, we provide recommendations such as regulation that forces organizations to disclose their training data, and tools that help artists prevent using their content as training data without their consent.

ACM Reference Format:

Harry Jiang, Lauren Brown, Jessica Cheng, Anonymous Artist, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Jonathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montreal, QC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3600211.3604681

This work is licensed under a Creative Commons Attribution International 4.0 License.

AIES '23, August 08–10, 2023, Montreal, QC, Canada

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604681

1 INTRODUCTION

In the two years since the publication of [18] which outlines the dangers of large language models (LLMs), multimodal generative artificial intelligence (AI) systems with text, images, videos, voice, and music as inputs and/or outputs have quickly proliferated into the mainstream, making the generative AI industry valued at an estimated $48B [125]. Tools like Midjourney [78], Stable Diffusion [5], and DALL-E [91] that take in text as input and output images, as well as image-to-image based tools like Lensa [97] which output altered versions of the input images, have tens of millions of daily users [47, 127]. However, while these products have captured the public's imagination, arguably to a much larger extent than any prior AI system, they have also resulted in tangible harm, with more to come if the ethical concerns they posit are not addressed now. In this paper, we outline some of these concerns, focusing our discussion on the impact of image based generative AI systems, *i.e.* tools that take text, images, or a combination of both text and images as inputs, and output images. While other works have summarized some of the potential harms of generative AI systems more generally [18, 28, 29], we focus our discussion on the impacts of these systems on the art community, which has arguably been one of the biggest casualties (Section 4) [40, 138].

As we argue in Section 3, image based generative AI systems, which we call **image generators** throughout this paper, are not artists. We make this argument by first establishing that art is a uniquely human endeavor, using perspectives from philosophies of art and aesthetics. We further discuss how anthropomorphizing image generators and describing them as merely being "inspired" by their training data, like artists are inspired by other artists, is not only misguided but also harmful. Ascribing agency to image generators diminishes the complexity of human creativity, robs artists of credit (and in many cases compensation), and transfers

Style mimicry by fine-tuning

pensation adds to the already precarious positions that the majority of professional artists are in [70, 92, 138]. In addition to the lack of compensation, using artists' works without their consent can cause them reputational damage and trauma. Users of image generated art can mimic an artist's style by finetuning models like Stable Diffusion on specific artists' images, with companies like Wombo even offering services to generate art in the style tied to specific groups of artists like Studio Ghibli [133]. A number of artists have described this practice as "invasive" and noted the manner in which it causes them reputational damage. After a Reddit user posted

The same method we use next week: a model fine-tuned on one set of images

Jiang et al., AIES 2023, Section 4.2

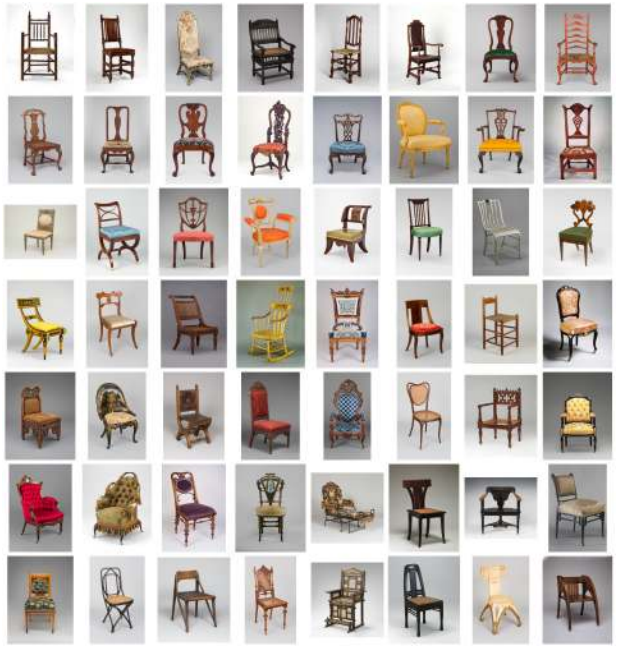
Opt-in consent

In this paper, we have reviewed the chilling impact of image generators on the art community, ranging from economic loss, to reputational damage and stereotyping. We summarized recommendations to protect artists, including new regulation that prohibits training image generators on artists' works without opt-in consent, and specific tools that help artists protect against style mimicry. Our work is rooted in our argument that art is a uniquely human endeavor. And we question who its further commodification will benefit. As

Opt-in: an artist's work is used for training only if they agree

Jiang et al., AIES 2023, Section 8

Our chairs and consent



All 56: public domain (CC0)

Left out: 149 of 178 Met chair records from 1875 to 1950, because they are not public domain

Your set: your own designs, or designs whose owners agreed, and say which in your datasheet

The Metropolitan Museum of Art and the Art Institute of Chicago, open access

Searching two sets of captions





A woven seat: **4** of 4 on the left, **2** of 4 on the right

The same chair in both lists: **1** of 4, for each of the four phrases

Left: BLIP captions. Right: each chair's title and materials from its museum record, standing in for your own captions.
Sentence similarity (all-MiniLM-L6-v2) to "a chair with a woven seat" and "a carved chair", from the week 4 notebook

Discussion

1. Which of your designs would you leave out of your set?
2. What can your captions say that a captioner cannot?
3. What is missing from your set?

Led by this week's discussion leads, with questions from the reading posts

Workshop: collect and map your set

1. Collect 20 to 40 of your own designs
2. Map them with the notebook
3. Caption them two ways: automatic, and in your own words
4. Search both with the same 4 phrases

Due before week 5: your map, the two searches side by side with a few sentences on which captions worked better, and your Motivation and Composition answers

Midpoint next week

7 minutes per team

Next: Week 5

Fine-Tuning and Steering

Read before the studio:

- Amershi et al., “Power to the People,” AI Magazine 2014
- Ruiz et al., “DreamBooth,” CVPR 2023
- Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” ICLR 2022