AI for the Media
Assignment 2 – Datasheet
By Marissa Beaty

Motivation

For what purpose was the dataset created?

This dataset was created as part of a graduate course assignment but is intended to complete a mini project training a Pix2Pix generative model to produce realistic architectural images from simplistic line drawings. As a part of this project, we were tasked with creating our own dataset containing images, audio, or video files. This dataset fulfills this task as each image has been added, reviewed, and cleaned by me to ensure it fulfills the requirements of the assignment and of a Pix2Pix model.

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

This dataset was created by my own hand with support from the google extension PinDown. This dataset was created in accordance with the guidelines set forth by the University of Arts London Creative Computing Institute's "AI for the Media" course leaders.

Who funded the creation of the dataset?

The creation of this dataset has not been externally funded.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

This dataset consists of two folders: inputs and outputs. The outputs are a collection of photos containing architectural buildings of varying architectural styles. The inputs are those same photos put through an edge detection program to strip away all details beyond the key lines and shapes defining the buildings' structure. Each input is paired with an output as necessary for training a Pix2Pix model. The dataset has been reviewed several times to ensure each image is unique either in the architectural model it contains, or in the angle at which that model is portrayed.

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

This dataset consists of 1,290 inputs and 1,290 outputs. As the Pix2Pix model takes in one image with an input and output merged together (same height, double the width), the final dataset to train a Pix2Pix model is 1,290 image pairs.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

This dataset is a collection of seven different Pinterest searches to scrape together varying architectural styles and movements (across time periods and cultures) in an effort to create a limited sample of popular architectural styles. This list is not exhaustive, and it should be known that to utilize this dataset or idea on a larger scale, a larger and more representative dataset should be considered. The limited scale of this dataset is in proportion to the time allotted to create the dataset and use the dataset on a training a Pix2Pix model. Not every architectural style will be represented in equal value, which is one of the additional limitations of the data. To ensure the dataset, however, represents a sample of various styles and movements, each Pinterest search was selected with specific wording and then reviewed to determine where gaps may lie. Any gaps were filled by additional Pinterest searches. The wording selected for each set was chosen based on prior Art Historical knowledge obtained through an undergraduate degree on the subject.

Is there a label or target associated with each instance?

This dataset does not require any labels beyond those images working as inputs and those working as outputs.

Are there recommended data splits (e.g., training, development/validation, testing)?

It is encouraged always when training a neural network to split the dataset into a training set and testing set. How you choose to split this up is up to personal preference. For utilizing this dataset, myself I will be following the standard split of 70% training, 20% testing, and 10% validation. Given the size of this dataset, the split above was selected to best train the model and avoid overfitting.

Are there any errors, sources of noise, or redundancies in the dataset?

This dataset is free of errors, noise, and redundancies.

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

This dataset does not rely on external sources. Though the dataset was taken from an external source (Pinterest), all photographs have been downloaded and are saved separate from this source. Any changes to the original source will not affect the dataset. Given the nature of Pinterest searches, researching the same keywords may not create the same set of images as produced for my search. Similarly, the images downloaded on Pinterest were reviewed, and those not fitting the needs of the dataset were removed.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?

This dataset does not contain anything that is confidential or protected.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

This dataset does not contain any offensive, insulting, threatening, or otherwise anxiety inducing materials.

Does the dataset relate to people?

This dataset does not relate to people.

Collection Process

How was the data associated with each instance acquired?

The data in this dataset was acquired using the Google Extension 'PinDown' which allows users to download all images produced from a Pinterest search. In order to create the full dataset, the following searches were completed: "Architecture", "Green Architecture," "Modern Architecture," "Unique Architecture," "Asian Architecture," "African Architecture," "Coastal Architecture," and "Home Exterior." These searches created a dataset close to 7,000 samples of images and video. The dataset needed for my Pix2Pix model, however, required this dataset to be cleaned. Specifically, it required me to first delete any video, as only images were to be used for the model. Then it was reviewed to ensure all images were 1) of an architectural building, 2) were unique to the dataset, and 3) were not obstructed by other stimuli such as people, cars, other buildings, overgrown fauna and flora. In fulfilling this criterion, my dataset narrowed to 1,290 samples. I then put this dataset through a simple program to rename all images using a numerical system. In this way, all output images will be easily labeled and accessible. This labelling system will also be utilized on the input data.

As mentioned above, the downloaded images created only the output dataset. A Pix2Pix model requires an output paired with an input. To create my inputs, I utilized a Canny edge detection model based off code found on StackOverflow that ran through each of my outputs and produced a matching input. This input stripped away the overall design and realism of the image and left it with the structural lines and shapes. Once this was complete my dataset was reviewed once more to ensure all images were paired and no outliers were left. Documentation to the StackOverflow code is included in the ReadMe connected with the Dataset. The separate input and output images were then merged to be one image containing both one input and it's corresponding output.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

This dataset was created using a combination of PinDown, manual processing, and Canny edge detection code. Additional code was used by piecing together research from StackOverflow, Github, and course materials.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The Pinterest searches were selected to capture a wide array of architectural styles ranging along varying time periods and cultures. The original searches did not include "Coastal Architecture" nor "Home Exterior." These were later added on as a review of the dataset revealed gaps in these specific architectural styles. The architectural styles overall were selected based on prior architectural knowledge learned during my bachelor's degree in art history.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process for this dataset was created by myself for my own purposes.

Over what timeframe was the data collected?

This dataset was collected over the course of a week. Though obtaining the data from the Pinterest searches required only a few minutes, processing the data, and running it through the code took several additional days.

Were any ethical review processes conducted (e.g., by an institutional review board)?

As this dataset does not include any sensitive information, nor will it be shared publicly, an ethical review process was not held. Though, it is important to note, that ethical considerations were a part of the creation process when selecting what to include in the dataset and when considering the implications of training a neural network on the dataset. After careful thought, since the dataset was obtained legally, without including sensitive or confidential information, and will not be shared beyond those reviewing my course assignments, any ethical implications were designated inconsequential.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

This dataset does not relate to any person living nor dead.

Preprocessing/Cleaning/Labelling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

To create this dataset, I used PinDown, a Google Extension that allows you to download images from a Pinterest search. Though PinDown saves a significant amount of time in creating an initial dataset, in downloading all the images, it also downloads any advertisements and videos displayed on the search page. When first going through the dataset, these non-relevant images and videos were deleted. Next, the dataset was reviewed to ensure all images were 1) of an architectural building, 2) were unique to the dataset, and 3) were not obstructed by other stimuli such as people, cars, other buildings, overgrown fauna and flora. Any image that contained one of the three above was also removed from the dataset.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw dataset was not saved, specifically because the images (and in some instances video) removed is not positively contributing to the dataset, in that including it would detract from the dataset and training of a neural network.

Is the software used to preprocess/clean/label the instances available?

PinDown is available as a Google Extension. All other processing of the data was done by hand. The code used to create my input data is a simple edge detection model available here and on the ReadMe attached to the dataset: https://stackoverflow.com/questions/58314400/edge-detection-for-multiple-images. I also used an image merging technique from stack overflow available here: https://stackoverflow.com/questions/65342361/photos-side-by-side.

Uses

Has the dataset been used for any tasks already?

This dataset will be used for a Term assignment training a Pix2Pix neural network model to produce realistic architectural images from simple, structural drawings.

What (other) tasks could the dataset be used for? Are there tasks for which the dataset should not be used?

This dataset could also be split and used separately to train a neural network, such as training a network to produce new structural drawings or training a network to produce new realistic architectural images. There are no limitations perceived of this dataset, however, it is encouraged that other tasks consider widening the scope of the dataset in terms of architectural styles and selected images, and or, refining the edge detected images to ensure they fit with the criteria of an alternative project.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

It is important to remember that this dataset is limited in scope, composed of only 1,290 output architectural images. The composition of these images does not contain all representations of architectural styles or ideas, which may limit the types of architectural representative images that can be produced from a neural network. Though I did intend to include as varied a selection as possible on architectural styles (including both different time periods and cultures), the produced images should be critically reviewed as to the potential biases a larger dataset might have avoided. Additionally, all images were run through an edge detection model using the same sigma value (or same standard deviation). Though this worked effectively for most images, some may have benefitted from an alternative sigma value. Given the scope of this project and the size of the dataset, different sigma values were not tested, but would be a good place to start for further or alternative training with the data. Beyond this, there are no other potential harms or risks using this dataset.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

This dataset will be published on Github for instructors of my course to access. This dataset will not be distributed elsewhere by myself or my course instructors.

When will the dataset be distributed?

The dataset will be distributed along with my Term materials at the end of Term on March 13, 2023.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

This dataset will not be distributed under a copyright or other intellectual property license.

Have any third parties-imposed IP-based or other restrictions on the data associated with the instances?

This dataset does not have any third-party IP-based impositions or restrictions placed on it, however, as Pinterest images are only protected vis a vis the Pinterest platform, there is no guarantee that all images associated with the dataset are not protected once removed from the

Pinterest platform. To avoid potential restrictions and impositions, this dataset will only be used for this term project and will not be distributed publicly in order to ensure copyright protections are not violated.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

Export controls and other regulatory restrictions do not apply to this dataset.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

This dataset will be hosted and maintained on my personal computer but made available to members of my Course Instructive team via GitHub for review and grading.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Those interested in contacting me (the creator of this dataset) can do so via m.beaty0520221@arts.ac.uk.

Is there an erratum?

This dataset does not have an erratum.