Motivation

The data used is from here:
https://www.kaggle.com/datasets/landlord/handwriting-recognition

The dataset is of 400000 handwritten names collected through charity projects.

The purpose of the dataset was to explore different ways to do handwriting to text conversion.

It is unclear who made or funded this dataset but it is in the public domain.

Composition

- The input is the image of the handwritten name, and
- the output is the name in a string.

In the dataset, there are this many in the training set (331,059), testing set (41,382), and validation set (41,382) respectively.

(Note: Although when training my model I used the testing set to train the model.)

There are missing labels for some images. There is no confidential information in this dataset.

Collection process

It is unclear how the data was acquired, if the labels were a part of a larger subset and over what timeframe the data was collected.

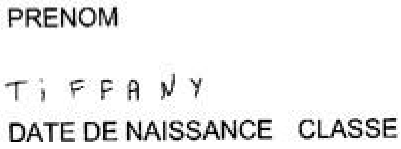
Preprocessing/cleaning/labelling

There were some images without a label so I removed them in the training and validation data.

Also I noticed sometimes, there would be some text outside the handwritten parts which get picked up by the model, effecting the answer.

For example:

For this image, the label would say “Tiffany” but the model would output “Prenom”, “Date De Naissance” and “Classe”

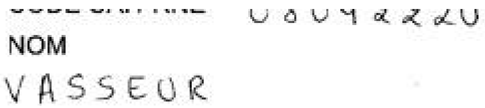


PRENOM

TiFFANY

DATE DE NAISSANCE CLASSE

And for this image, the model would pick up the half cropped text.



NOM

VASSEUR

(Note: In the validation of the model, I would try my best to remove any characters that are not the name so I can test the accuracy of the model when matching the names.)

Uses

The data can be used for anything that takes an image of handwritten names and returns computer text.

There does not seem to be anything that would impact future use or any unfairness against individuals or other risks and harms.

Distribution

The dataset is distributed via Kaggle.

You can use the data however you'd like because it is in the public domain.

Maintenance

The dataset is uploaded and maintained by the Kaggle user "Landlord" from Bengaluru, Karnataka, India. The last time the dataset was updated was 5 years ago in 2020.