



Japanese Character Classification

日本文字の分類



Author: Chaz Frazer

Table of Contents



Introduction

Business Question &
Application



Data & Info

Sourcing, Datasets,
Data Manipulation



Data Preparation

Preparing images
from binary data for
modeling &
Evaluation



EDA

The story the data
tells & insights



Modeling


Model preparation,
Modeling
techniques & results



Conclusion & Next Steps

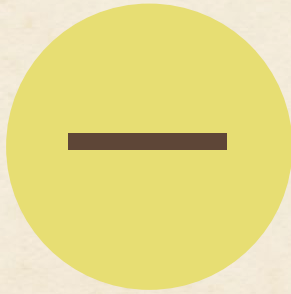
Stretch goals and
further MVP analysis





“相手の理解できる言語で話せば、その人の頭に入る。相手の言語で話せば、その人の心に届く。”

—If you talk to a man in a language he understands, that goes to his head. If you talk to him in his own language, that goes to his heart.



Introduction

Business Questions Asked & Answered



Firstly

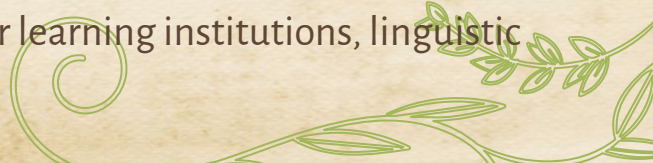
Can a viable product model be created to accurately transcribe, read, and identify Japanese text for the archiving of important literary works? This can be used to preserve the surviving texts of endangered languages from the Ainu and Ryukyu minority groups in Japan.



Secondly

Can this be expanded to create an accurate API that recognizes written Japanese characters for touchscreen devices (ie. dictionaries, translation apps).

Target audience is Japanese and English research orgs, higher learning institutions, linguistic preservation societies, and language students.





Data & Info

Sourcing & Formatting





Source

The data is from the ETL Character Database, which includes over a billion total of Japanese characters hand-written and reorganized by the National Institute of Advanced Industrial Science and Technology (AIST).



Data Properties

Each file contains 5 data sets except ETL8G_33.

Each data set contains 956 characters written by a writer.

Each writer wrote 10 sheets (genkouyoushi) per data set.



Motivations

My background in linguistics provided the platform to dive into computational linguistics for this project. Project to be expanded to create an accurate API that recognizes written Japanese characters for touchscreen devices (ie. dictionaries, translation apps).



Japanese Writing Systems



Kanji

Brought to Japan from China in the 8th century. Pictographs that convey meaning (anthropomorphic and abstract).



Hiragana

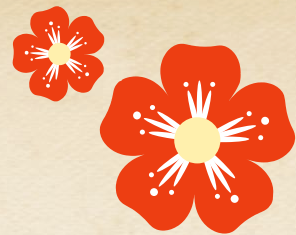
Phonetic 'alphabet' used for participles and to inflect verbs and adjectives.



Katakana

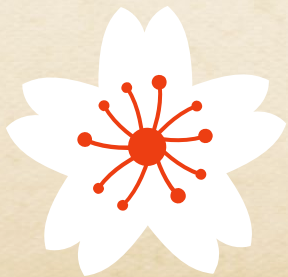
Same phonetic sounds as hiragana. Angular components from kanji. Used for foreign words, sounds, & onomatopoeia.





Data Preparation

From Binary to Black & White



Example Japanese Binary Image Table

Data read from
binary code
and saved to an
.npz file to be
re-read

Images reshaped
from 32x32 to
64x64 pixels

へ			皇	ぜ
イ	人	フ	フ	
ヤ	皇	語		
ぜ	エ	リ	内	
リ	モ	練	げ	



EDA

The story the data tells & insights

Initial Class Imbalance



Hiragana

71 unique hiragana characters (46 singular chars +
29 diphthongs)
113,60 images



Katakana

46 unique katakana characters
64,906 images



Kanji

883 unique kanji characters
139,680 images



Merged Data

215,946 images from combined
three datasets



The slide features a light beige, textured background. In the top-left and bottom-right corners, there are clusters of stylized flowers. Each cluster includes a large red flower with a yellow center and several smaller white flowers with red centers. Green stems and leaves are also visible. In the center, there is a yellow circle containing the Chinese character '伍' (five). Below this, the word 'Modeling' is written in a large, dark brown, serif font. At the bottom, the text 'Model preparation, techniques & results' is written in a smaller, dark brown, sans-serif font.

伍

Modeling


Model preparation, techniques & results

Train-Test Split



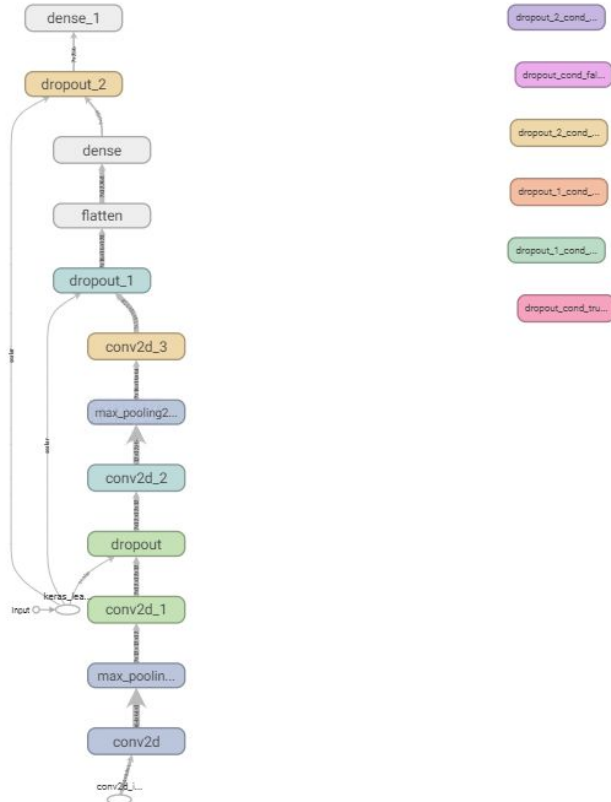
20%
Validation Set

60%
Training Data



20%
Test/Holdout Data

CNN Model



- ImageDataGenerator used to create variation and prevent overfitting
- Models run on AWS EC2 instance using g4dn Nvidia Tesla GPU architecture
- Tensorboard used to live track the model across epochs
- Reduce Learn Rate on Plateau callback utilized to adjust learning rate on the fly if accuracy did not improve after 3 epochs
- Early Stopping callback used to stop model training if accuracy did not improve after 5 epochs

CNN Model Results



Training



Validation



Test

Accuracy	Loss
99.79%	0.68%
99.40%	3.2%
99.73%	0.90%



Confusion Matrix



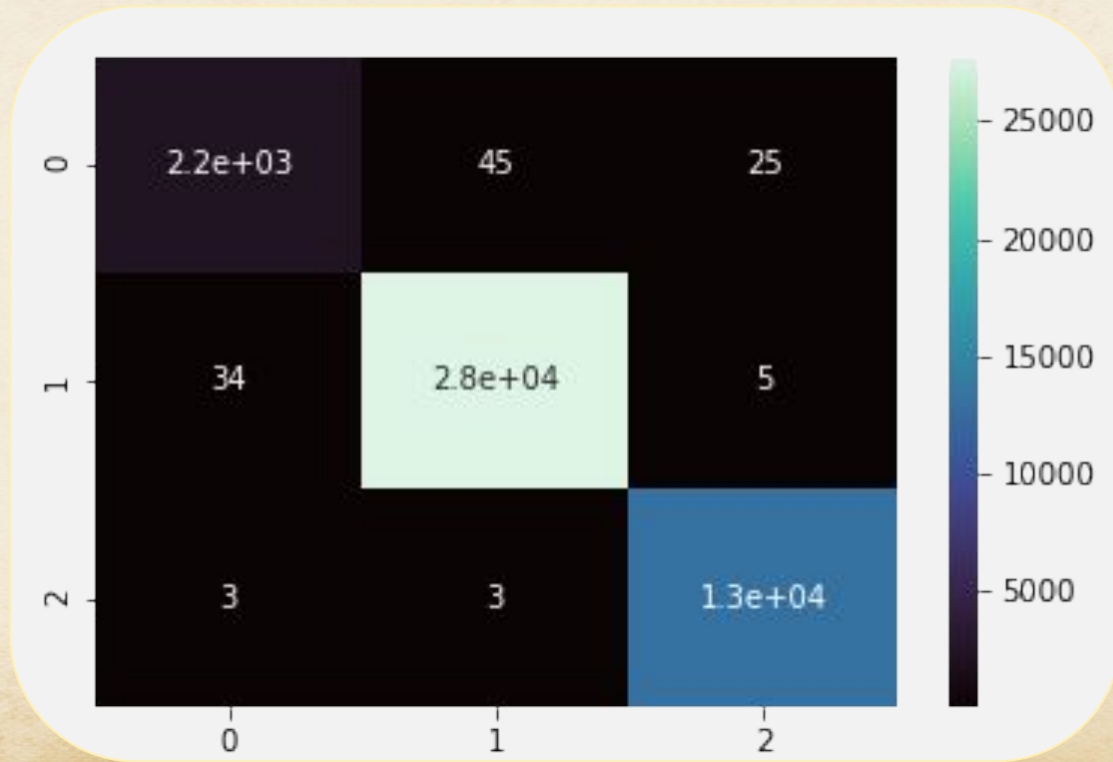
Accuracy / True Positive Rate

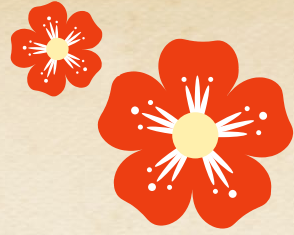
- 0 = Hiragana
- 1 = Kanji
- 2 = Katakana



Actual

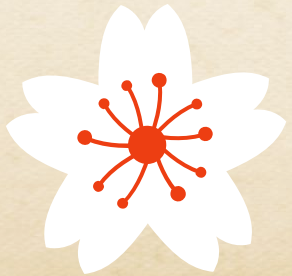
Predicted





Conclusion & Next Steps


Stretch goals and further MVP analysis





Accuracy is Key

Over 200,000 unique
characters trained to a
recognition percentage
of over 99% accuracy



Stretch Goals and Next Steps

Kuzushiji

Work with kuzushiji (Japanese cursive writing) KMINST dataset variations

OpenCV

For live model image recognition using webcam

iOS API

Handwriting recognition app using trained model

Linguistics ECI

The CUNY Endangered Language Initiative strives to preserve our dying languages around the world. Use model as a way to utilize computational linguistics and preserve precious texts and early written Japanese history





ありがとうございました
(Thanks!)

Questions are open!



[Github.com/Mynusjanai](https://github.com/Mynusjanai)



[@mynus_jp](https://twitter.com/mynus_jp)



[.linkedin.com/in/chazfrazier](https://www.linkedin.com/in/chazfrazier)

