Japanese Character Image Classification

日本文字の分類

Author: Chaz Frazer

Table of Contents

“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.

二

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 (漢字)

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

Hiragana (ひらがな)

Phonetic 'alphabet' used for participles and to inflect verbs and adjectives. Curved components from kanji.

Katakana (カタカナ)

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

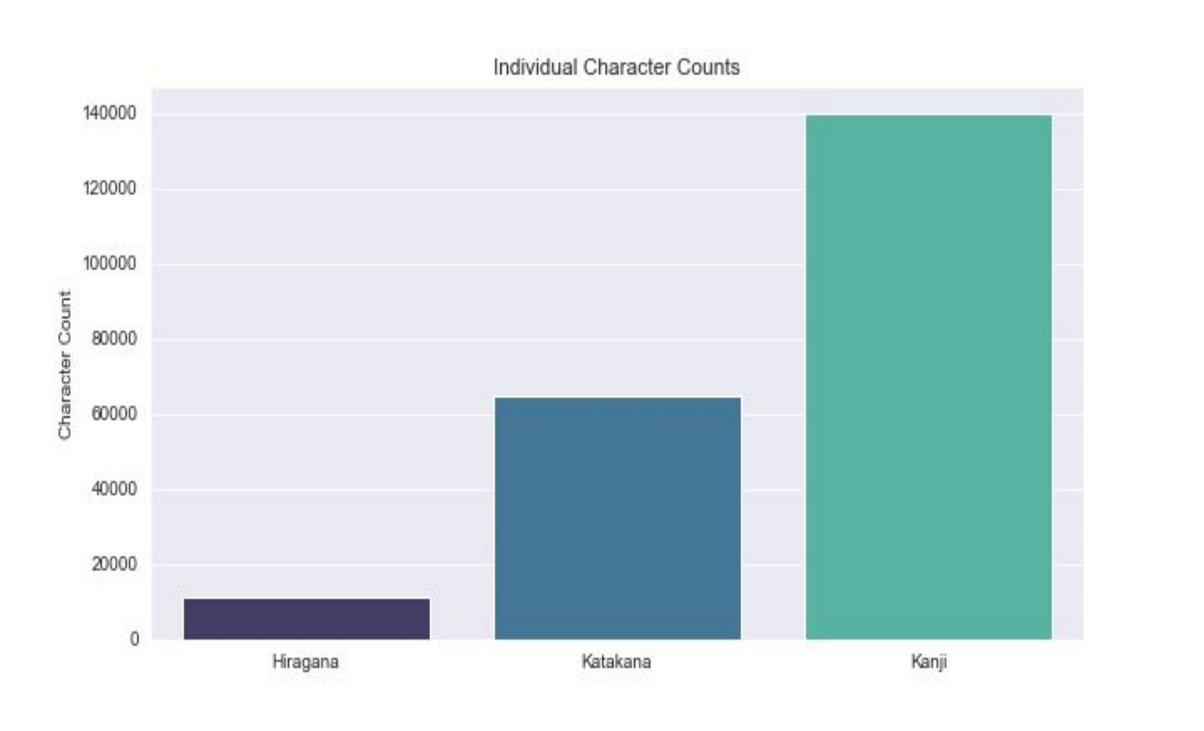
Data read from binary code and saved to an .npz file to be re-read

Images reshaped from 32x32 to 64x64 pixels

四 EDA

The story the data tells & insights

Character Image Counts



- Because of the sheer amount of kanji in Japanese (~3000 in daily use and over 35,000 known to exist), the record number is much larger to account for this robustness in the language
- Note that because only 71 characters exist in hiragana (46 individual + 29 diphthongs), the majority of the entries in ETL8G are kanji
- Reading the Kanji characters from the ETL8G file. Kanji and Hiragana share the same dataset, so to extract only kanji

Initial Class Imbalance

Hiragana

71 unique hiragana characters (46 singular chars + 29 diphthongs)
11,360 images

Katakana

46 unique katakana characters
64,906 images

Kanji

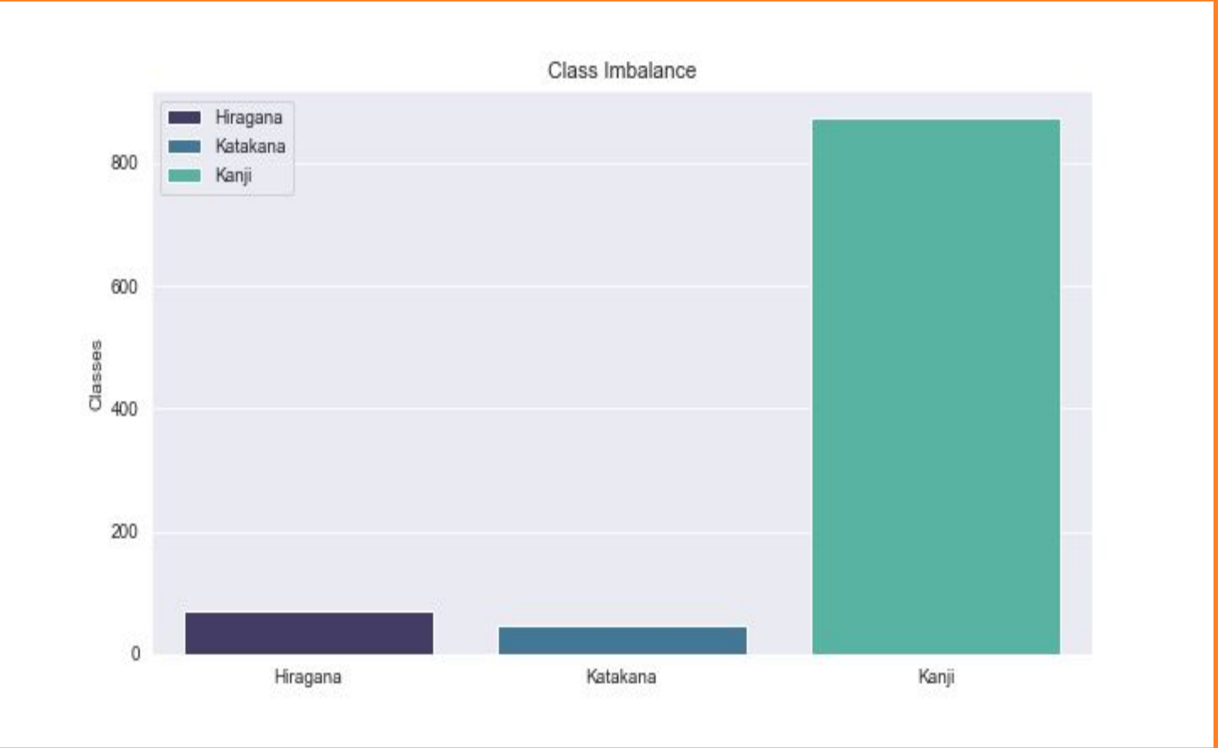
883 unique kanji characters
139,680 images

Merged Data

215,946 images from combined three datasets

Initial Class Imbalance

- Kanji is a sub-sample of 883 characters from the ~3,000 used on a daily basis (speaking & reading)
- This natural class imbalance is a direct result of the structure of the language, and was weighted appropriately during modeling to account for the imbalance
- Hiragana included the 29 diphthongs (character combinations). I chose to omit them from the katakana images for model and data variation



伍

Modeling

Model preparation, techniques & results

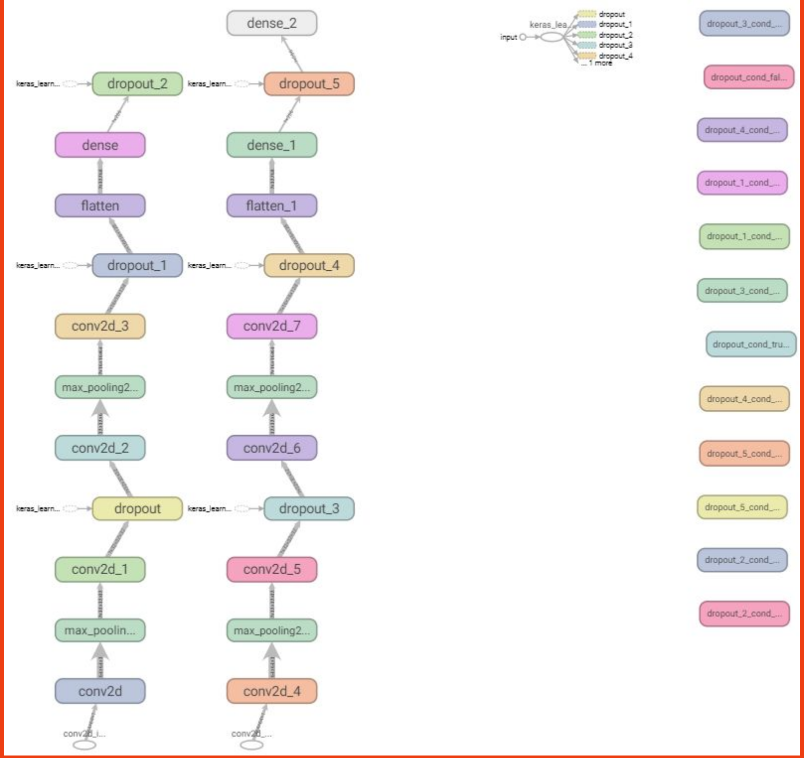
Train-Test Split

20%
Validation Set

60%
Training Data

20%
Test/Holdout Data

Deep Learning Models



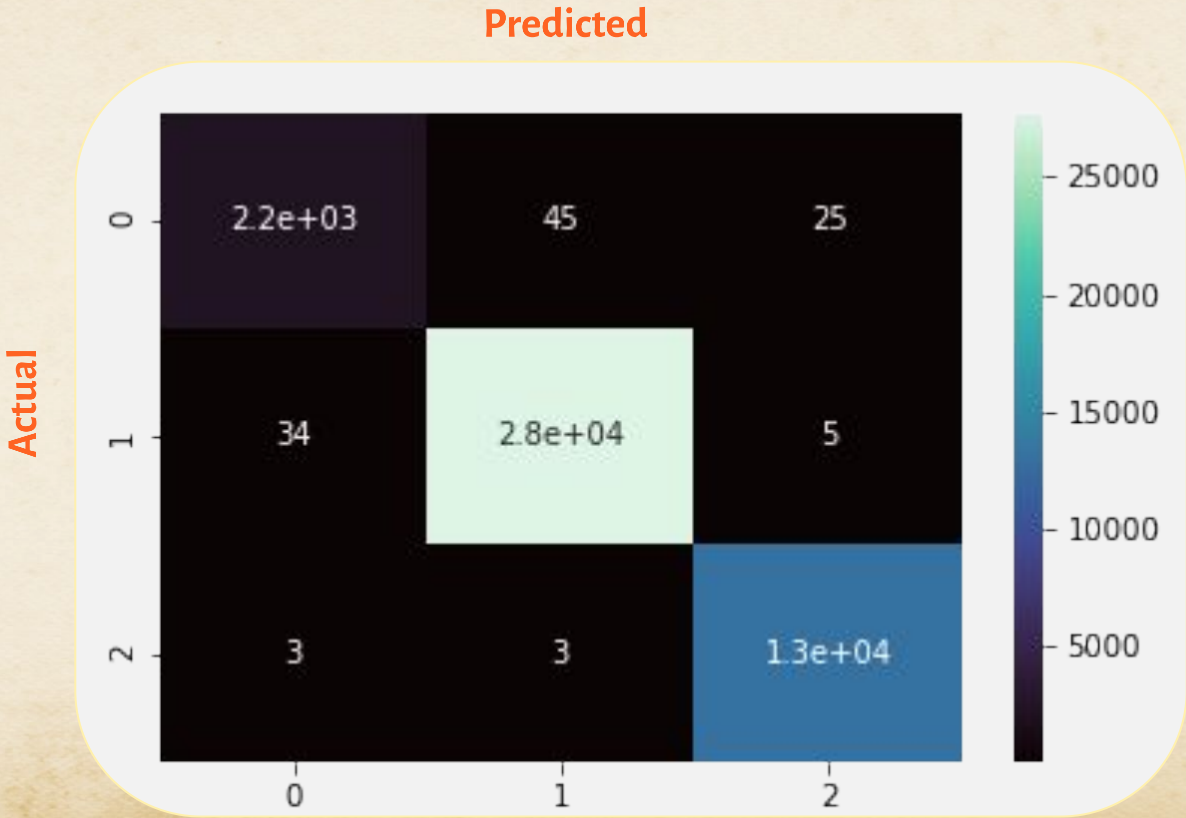
- ImageDataGenerator used to create variation and prevent overfitting
- Models run on AWS EC2 instance using g4dn Nvidia Tesla GPU architecture
- Tensorboard used to live track the model across epochs
- Reduce Learn Rate on Plateau callback utilized to adjust learning rate on the fly if accuracy did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

Tensorboard Model Tracking

CNN Confusion Matrix

Accuracy / True Positive Rate

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



All Model Results

	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Test Accuracy	Test Loss
KNN	92.67%	N/A	92.75%	N/A	95.95%	N/A
Random Forest	94.50%	N/A	94.48%	N/A	94.95%	N/A
CNN	99.79%	0.68%	99.40%	3.2%	99.73%	0.90%
cuDNN (Nvidia)	**99.95%**	**0.017%**	**99.55%**	**0.027%**	**99.48%**	**0.019%**

六

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 ELI

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

References

Electrotechnical Laboratory Data Set (ELTCDB)

http://etlcdb.db.aist.go.jp/

Japanese National Institute of Advanced Industrial Science and Technology (AIST)

https://www.aist.go.jp/

Japan Electronics & Information Technology Industries Association (JEITA)

http://www.jeita.or.jp/english/

ありがとうございました

(Thanks!)

Questions are open!

Github.com/Mynusjanai

www.mynus.gg

Linkedin.com/in/chazfrazer