

COMP2261 - Machine Learning Report

Michal Pluta

Abstract— Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua quaerat.

Index terms—Machine Learning, K-nearest-Neighbours, Logistic Regression, Hyperparameter Tuning, CNN

I. INTRODUCTION

Mandarin Chinese is one of the most spoken languages with 1.3 billion native speakers globally. In countries with significant Chinese-speaking populations, the ability to accurately and automatically identify characters has applications across a variety of domains. To name a few:

- Recognising mail addresses in the postal service, reducing the dependency on manual sorting.
- Digitisation of cultural heritage, where manual digitisation is too difficult or infeasible time-wise.
- Improving accessibility for the visually impaired or those who have reading difficulties.

This report explores the application of machine learning methods on a dataset comprising Chinese (Simplified) characters, akin to the widely recognised MNIST dataset for handwritten digits. The significance of this study lies in the inherent complexity and variety of Chinese script, and the motivation lies in a personal interest in learning the language.

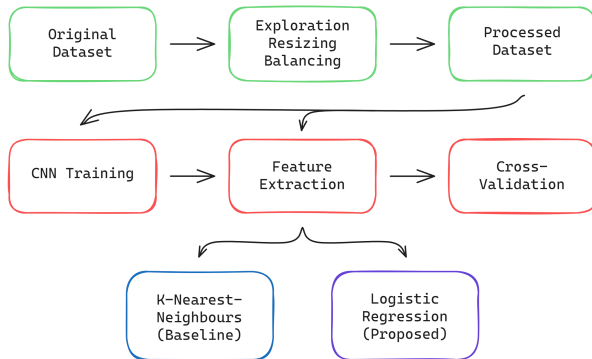


Figure 1: Workflow of the proposed ML System

The primary objective is to demonstrate the performance, tuning process, and limitations of two separate machine learning models.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua quaerat voluptatem. Ut enim aequae doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguere possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam

facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet, ut et.

II. DATASET OVERVIEW

While Chinese has more than 50,000 characters with roughly 6,500 in daily use, we will only deal with a subset part of the HSK 1 curriculum. HSK is a Chinese Proficiency Exam with levels ranging from 1 (Beginner) to 9 (Near-native).

Chinese characters, also known as Hanzi, are composed of strokes, the basic units of writing, arranged in a specific order and direction. The way these strokes combine give rise to a vast array of characters, each with its own unique meaning and structure.

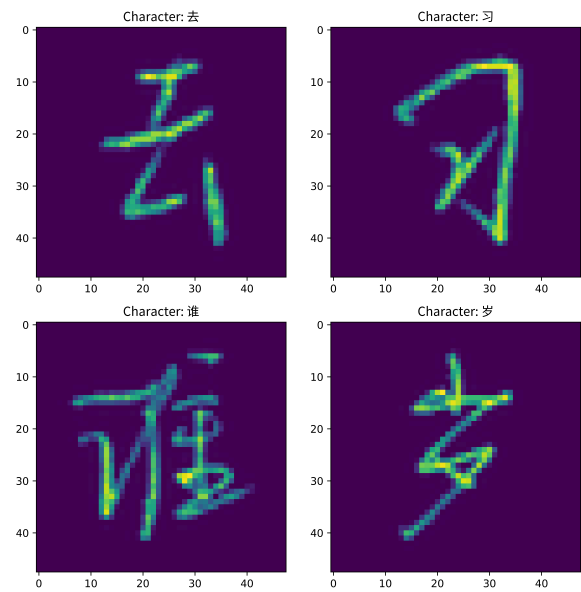


Figure 2: Examples of images in the dataset

The uniqueness of handwritten Chinese characters becomes even more pronounced when considering China's population. This alone leads to a huge variation in style, consistency, size, alignment, and thickness of strokes (Figure 2).

III. DATASET TRANSFORMATIONS

The original dataset [1] contains 178 classes of images, split using an 80/20 train-test split. The images are greyscale with a 1:1 aspect ratio, and in total there are 131,946 samples. To minimise bias and to ensure the train and test sets were representative of the whole dataset, the sets were merged and then shuffle-split through stratification. This also allowed the dataset to be standardised using one directory only.

IMAGE RESIZING

One clear issue with the original dataset were the varying image dimensions.

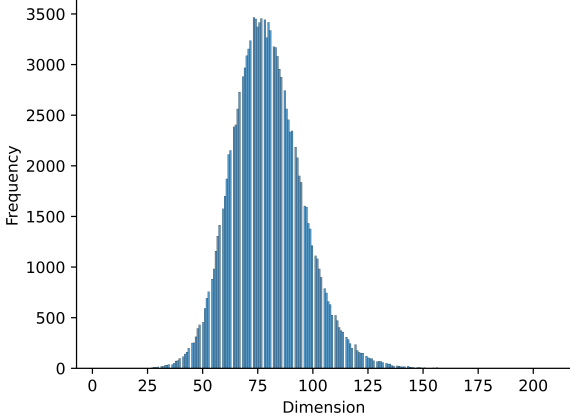


Figure 3: Distribution of image dimensions in the original dataset

This was a problem because the CNN used for feature extraction has specific input tensor dimension requirements, hence, all images had to be standardised. 48×48 was chosen as a suitable image size as it meant most images could be downsampled. Downsampling is often a better technique than up-sampling as it doesn't limit the model's ability to learn key features like object boundaries. Additionally, due to the complexity of Chinese characters, the images could not be downsampled further due to the likely loss of information.

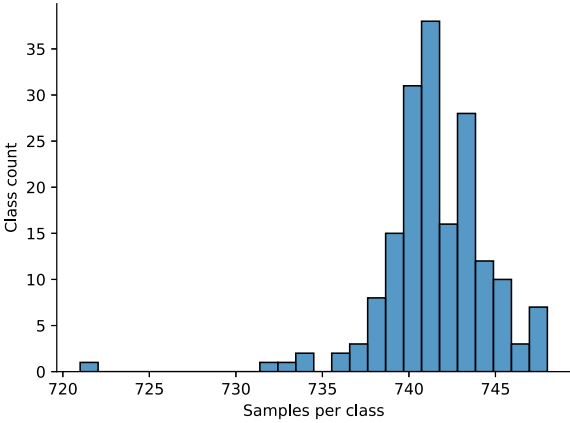


Figure 4: Distribution of the number of samples in each image class.

From further exploration, a class imbalance was evident.

Imbalance metric	Value
Imbalance Ratio	1.03745
Interquartile Range	3.0
Coefficient of Variation	0.00409

Table 1: Dataset imbalance metric values

Evaluating the imbalance ratio, IQR, and coefficient of variation, it was clear the imbalance was minimal and insignificant.

FORMAL DEFINITION

More formally, the images $x^{(i)}$, and one-hot encoded labels $y^{(i)}$ in this system are defined as follows:

$$x^{(i)} \in \mathbb{R}^{48 \times 48} \quad (1)$$

$$y^{(i)} \in \mathbb{R}^{178} \quad (2)$$

By implication, the input tensors are defined as

$$x \in \mathbb{R}^{131946 \times 48 \times 48} \quad (3)$$

$$y \in \mathbb{R}^{131946 \times 178} \quad (4)$$

This allows us to define the CNN's feature extraction operation, $\text{extract}()$ as

$$\text{extract}(x^{(i)}) = f^{(i)} \in \mathbb{R}^{512} \quad (5)$$

FEATURE EXTRACTION

Since the dataset was comprised of images, relevant features had to be extracted from each image. One of the most effective ways of doing this is by using a CNN (Convolutional Neural Network).

While pre-trained CNNs such as Resnet50 were initially considered due to their simplicity and outstanding performance, they yielded feature vectors which were simply too large (2048), resulting in unmanageable training times. One way to handle this would be to perform dimensionality reduction using a technique such as PCA, however, this is just as, if not more computationally intensive.

Instead, a custom CNN was developed using Keras' deep-learning library. The model features 3 2D-Convolution layers, each followed by a MaxPooling layer. While developing the model, severe overfitting was noticed. This was addressed by adding Dropout layers to the model, which would randomly set a fraction of the input units to 0. This improved the model's ability to generalise.

The CNN was tuned using a Bandit-Based tuner, Hyperband, provided by the Keras library. Once optimised, the model was cross-validated to ensure major overfitting was not present.

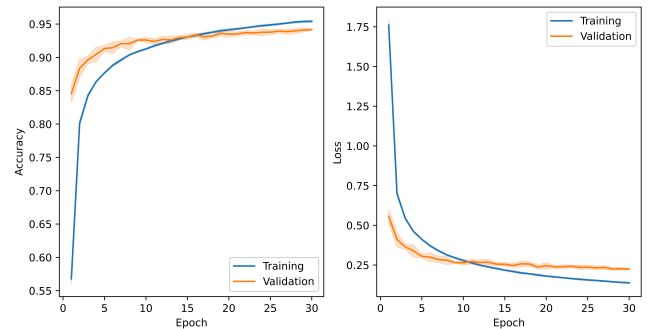


Figure 5: Graph of training and validation accuracy throughout training the CNN

From Figure 5 it is clear that some overfitting is still present, however, this amount is tolerable. The reason for the model performing better on validation data in the first few epochs is due to the Dropout layer. A dropout layer is a regularisation mechanism that is turned on when training and is turned off when testing. This results in the model performing worse on the training data.

IV. EVALUATION METRICS

- **Weighted Accuracy** - Since my dataset has a small amount of imbalance, to be on the safe side I will use the weighted accuracy across all predicted classes instead of average accuracy.
- **F1 Score** - In my classification problem there is no greater negative impact caused by a low sensitivity (Recall). This is not the case with something like medical image classification where low sensitivity could have serious consequences. Therefore, I will use the F1 score which is the harmonic mean of Recall and Precision as they correlate inversely with each other.
- **Training time** - While the total number of Chinese characters remains fairly constant, every so often new characters for complex ideas or newly discovered chemical elements are proposed, and this would warrant retraining/adjusting the model. Lower training times would allow the model to be updated faster.
- **Inference Time** - In applications such as real-time translation, language learning, or even autonomous driving where signs need to be read within fractions of a second, it is crucial to identify text with minimal latency. Additionally, for services with large volumes of data, an efficient, scalable, and high-throughput system is ideal.

V. MODEL EVALUATION

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aequale doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguere possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

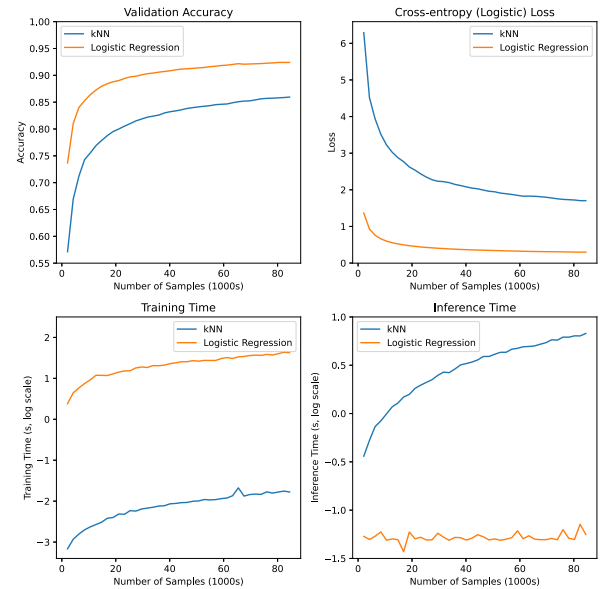


Figure 6: Accuracy of K-Nearest Neighbours for varying k

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aequale doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguere possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

Hyperparameter	Default	GridSearch
n_neighbours	5	10
weights	'uniform'	'distance'
metric	'minkowski'	'euclidean'

Table 2: Summary of tuned hyperparameters for k-Nearest-Neighbours

Hyperparameter	Default	GridSearch
solver	'lbfgs'	'lbfgs'
penalty	'l2'	'l2'
C	1	5

Table 3: Summary of tuned hyperparameters for Logistic Regression

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aequale doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguere possit, augeri amplificarique non possit. At

etiam Athenis, ut e patre audiebam facete et urbane Stolicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

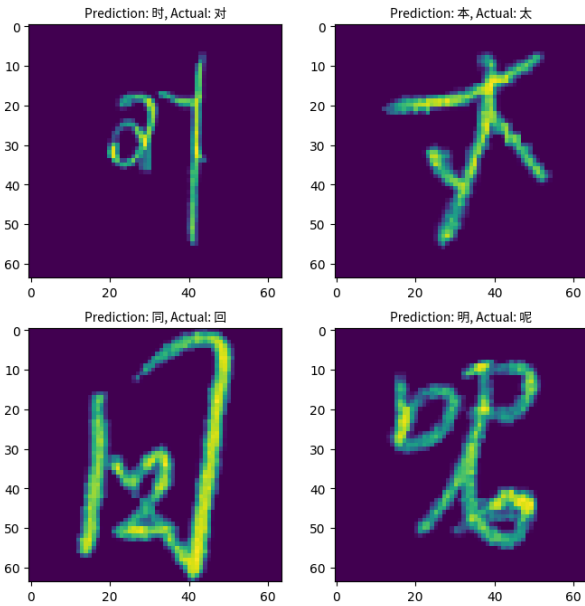


Figure 7: Distribution of image dimensions in the original dataset

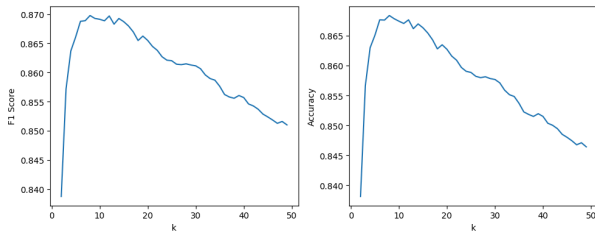


Figure 8: Accuracy vs Training Accuracy of Logistic Regression with varying amounts of training samples

VI. SELF EVALUATION

LECTURES

Throughout the lectures, I was most intrigued by the mathematical concepts that underpin each machine-learning model. It made me glad to know that the mathematical foundations I had practiced tirelessly last year were crucial to understanding both the inner workings of each model as well as the intuition behind how they all were developed. The lectures also showed me that not all models behave like a ‘black box’, and many have visual representations such as kNN and Lloyd’s Algorithm in KMeans.

COURSEWORK

One particular thing this coursework has taught me

MODULE DIFFICULTIES

My main difficulty in the module is being able to effectively translate my theoretical understanding into a practical solution. Ultimately, most of this content is completely new to me, hence I’ve never had a chance to implement or tune any

models independently. Thankfully, the practicals were able to address some of these concerns as I got hands-on experience implementing and adjusting various models.

REFLECTION

I believe the biggest challenge of this assignment was the size of the dataset and the disproportionate computational power I possessed. I approached this project wanting to do image classification related to one of my interests, however, I didn’t fully consider how long it would take to train even a single model with such a vast quantity of data. Moreover, I believe it would have been more effective to use a pre-trained CNN such as Resnet to ensure the feature extraction was as accurate as possible since all the other models rely on this.

UNIQUE CONTRIBUTIONS

While my models do not constitute anything novel, I find it nevertheless an area of research that should be explored.

REFERENCES

- [1] V. Kyzym, “Chinese handwriting recognition : HSK 1”. [Online]. Available: <https://www.kaggle.com/datasets/vitaliikyzym/chinese-handwriting-recognition-hsk-1>