COMP2261 - Machine Learning Report

Michal Pluta

***Abstract*— Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam 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. 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 magnam aliquam quaerat voluptatem. Ut enim aeque 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 distinguique 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 voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam.

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 HSK 1 (Beginner) up to HSK 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 are combined gives rise to a vast array of characters, each with its unique meaning and structure. 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.

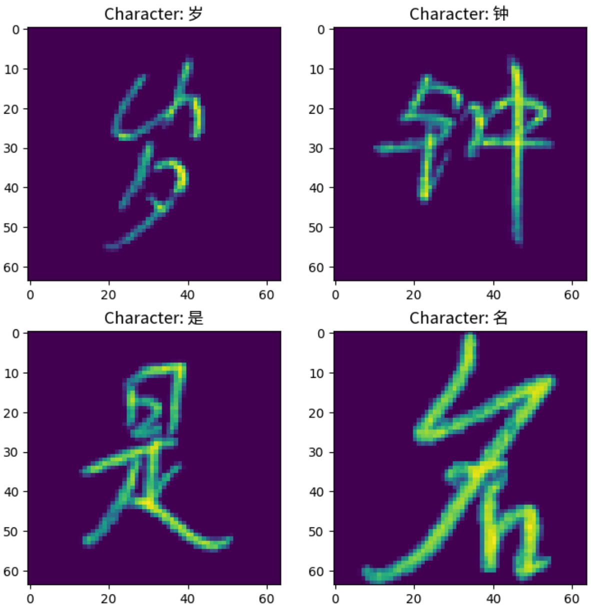

Figure 1: Examples of images in the dataset

III. Dataset transformations

The original dataset [1] contains 178 classes of images, split using an 80/20 train-test split. The images were all greyscale with a 1:1 aspect ratio, and in total, there were 131,946 images. 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 meant the dataset could be standardised by working with one directory only.

Image Resizing

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

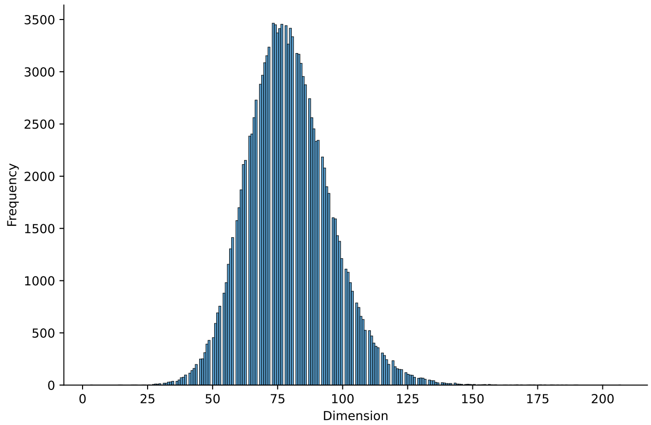

Figure 2: Distribution of image dimensions in the original dataset

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 upsampling 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 image dimensions could not be reduced further due to loss of information.

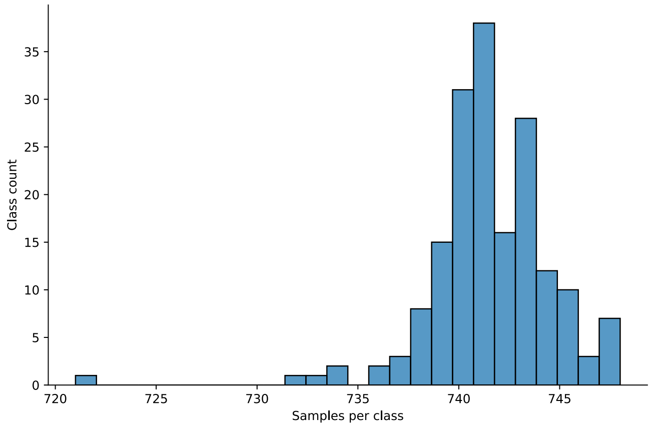

Figure 3: Distribution of the number of samples in each image class.

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.

Feature Extraction

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque 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 distinguique 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.

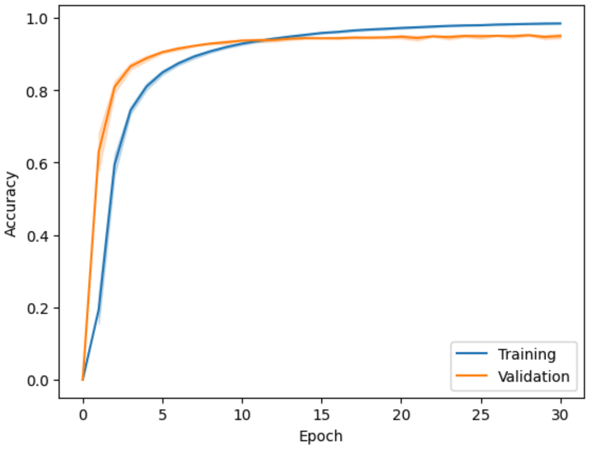

Figure 4: Distribution of image dimensions in the original dataset

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque 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 distinguique 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 voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam

ex eo est consecutus? – Laudem et caritatem, quae sunt vitae.

Formal Definition

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

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

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

By implication, the input tensors are defined as

$$x \in \mathbb{R}^{131946 \times 48 \times 48} \tag{3}$$

$$y \in \mathbb{R}^{131946 \times 178} \tag{4}$$

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

$$\mathrm{extract}\left(x^{(\mathrm{i})}\right) = f^{(\mathrm{i})} \in \mathbb{R}^{512} \tag{5}$$

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, it is more appropriate to 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.
- **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, high-throughput system is necessary.

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 aeque 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 distinguique 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 voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam ex eo est consecutus? – Laudem et caritatem, quae sunt vitae.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque 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 distinguique 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: Dataset imbalance metric values

Hyperparameter	Default	GridSearch
`solver`	`'lbfgs'`	`'lbfgs'`
`penalty`	`'l2'`	`'l2'`
`C`	`1`	`1`
`l1_ratio` ('saga' only)	`None`	`None`

Table 3: Dataset imbalance metric values

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque 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 distinguique 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 voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam ex eo est consecutus? – Laudem et caritatem, quae sunt vitae sine metu degendae praesidia firmissima. – Filium morte multavit. – Si sine causa, nollem me ab eo delectari, quod ista Platonis, Aristoteli, Theophrasti orationis ornamenta neglexerit. Nam illud quidem physici.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque 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 distinguique 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 voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam ex eo est consecutus? – Laudem et caritatem, quae sunt vitae sine metu degendae praesidia firmissima. – Filium morte multavit. – Si sine causa, nollem me ab eo delectari, quod ista Platonis, Aristoteli, Theophrasti orationis ornamenta neglexerit. Nam illud quidem physici, credere aliquid esse minimum, quod profecto numquam putavisset, si a Polyaeno, familiari suo, geometrica discere maluisset quam illum etiam ipsum.

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

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque 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 distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos.

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]