Handwriting Recognition: Traditional Chinese

Derry Li

Problem Statement

Inspired by MNIST, we aim to build a CNN Model that accurately recognizes various handwritten (complex) traditional Chinese characters of different styles to help elderlies who can’t type and learners of the Chinese language to learn new characters.

.

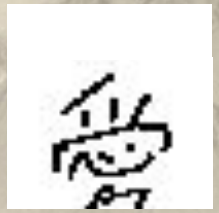
Data

- *Traditional Chinese Handwriting Dataset* from AI-FREE
- 4803 characters
  - **Most commonly used** by Chinese speakers
  - Each with **~50 handwritten versions**

Data

Data Augmentation

Vertical shift



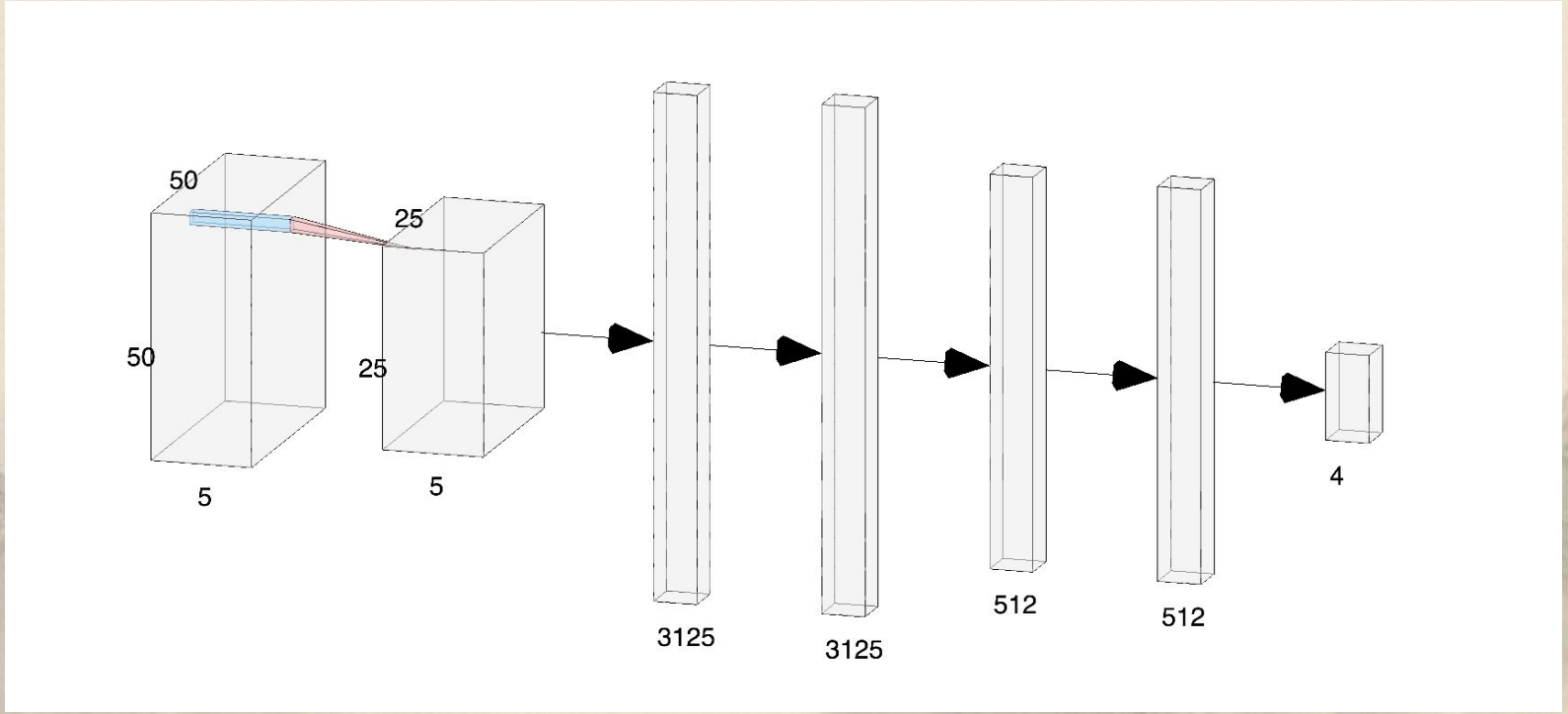
Horizontal shift

Zoom

Base Model

- Trained on **4** characters (**學 無 止 境**)
- Epochs = **50**
- Accuracy: ～ **95%** on train, **85%** on test
- **Conv** → **MaxPool** → **Flatten** → **Dropout** → **Dense** → **Dropout** → **Softmax**

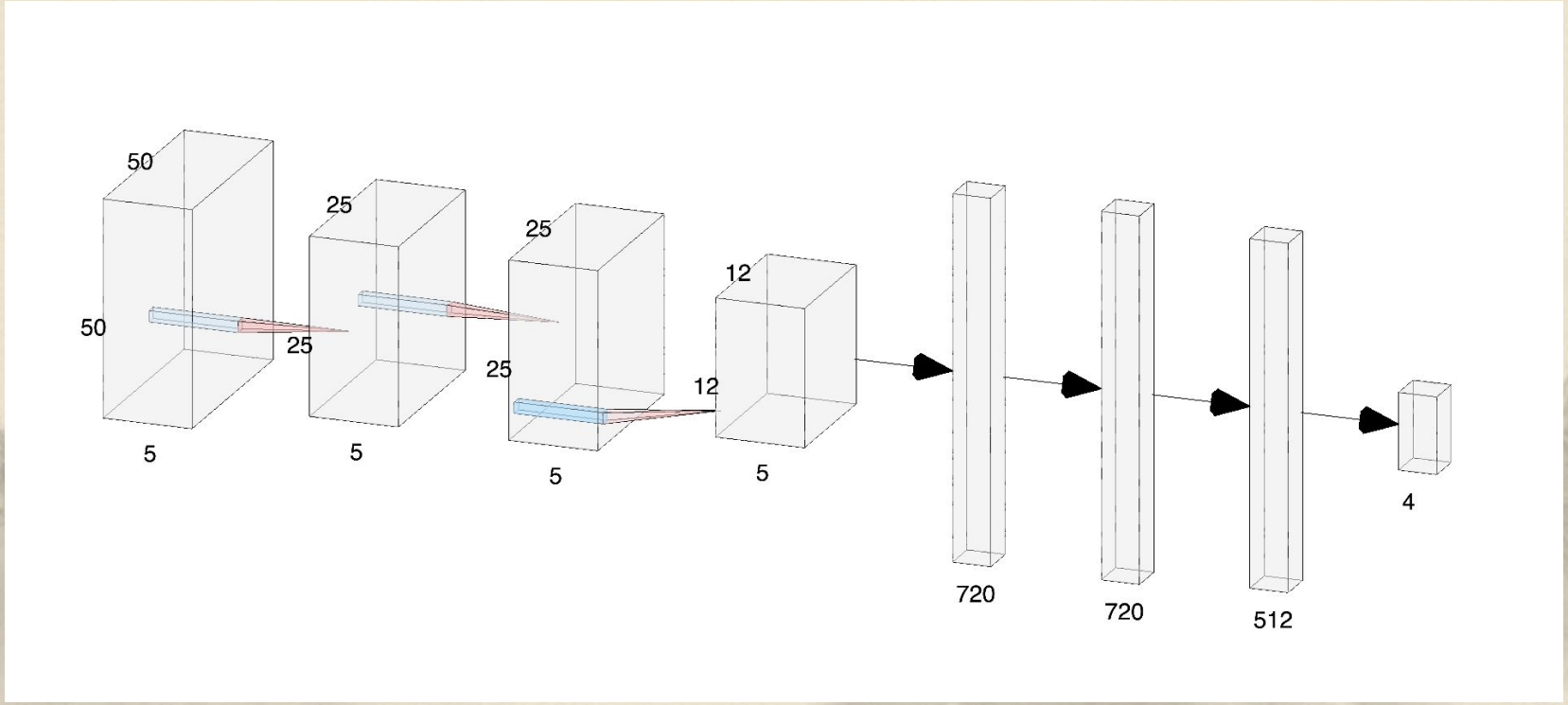
Base Model



Improved Model

- Trained on **4** characters (**學 無 止 境**)
- Epochs = **50**
- **～ 97% accuracy** on test
- **Conv → MaxPool → Conv → MaxPool → Flatten → Dropout → Dense → Softmax**

Improved Model



Improved Model – top 100 characters

- Trained on **100 most commonly-used** characters

  (的, 一, 是, 不, 了, 在, 人, 有, 我, 他, 這, 個 …)
- Epoch = **1000**
- Accuracy: **～ 90%** on train, **76%** on test
- **Conv → MaxPool → Conv → MaxPool → Flatten → Dropout → Dense → Softmax**

Improved Model – All 4803 characters

- Trained on **all 4803** characters
- Epoch = **50**
- ∼ **20% accuracy** on test 😅
- **Conv → MaxPool → Conv → MaxPool → Flatten → Dropout → Dense → Softmax**

Results + Future Work

- **97%** accuracy for 4 characters
- **76%** accuracy for 100 characters
- Gradio **App**!

What's next…

- More **data** → more writers!
- More **styles** (stroke thickness etc.)
- More **epochs** → more **time**!