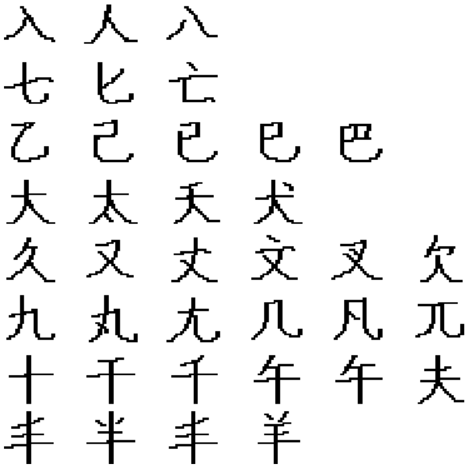
Handwritten Chinese OCR

By: Cary Wu, Hudson Chou

Motivation

OCR system for recognizing handwritten Chinese characters.

Train a model to surpass human accuracy



Google Translate

Pleco

Original database

Provided by National Laboratory of Pattern Recognition (NLPR) and Institute of Automation of Chinese Academy of Sciences (CASIA).

3,755 classes of characters and a total of **300 different writers** for each class

Our Dataset

Number of classes: 70 characters

['祝', '铰', '联', '瑟', '坑', '挑', '研', '件', '哀', '静', '儒', '藐', '瞪', '拿', '赡', '逞', '疫', '蜘', '山', '舟', '训', '圣', '毁', '舶', '煌', '骂', '假', '轮', '谍', '榆', '鼎', '硅', '眠', '朴', '半', '迄', '促', '绎', '觅', '勺', '戳', '酵', '撕', '盂', '弘', '忘', '砧', '蝴', '燎', '题', '尧', '谭', '订', '矗', '襟', '踪', '麓', '规', '挨', '恤', '聪', '卞', '丁', '岔', '租', '充', '各', '猴', '缩', '枚']

Description of datasets

	Dataset Purpose	Number of samples per label	Size of each image
	Training	118	64 x 64
	Validation	55	64 x 64
	Testing	55	64 x 64

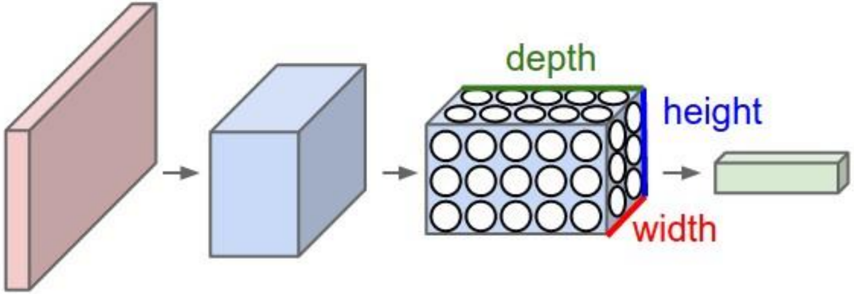
Our Model

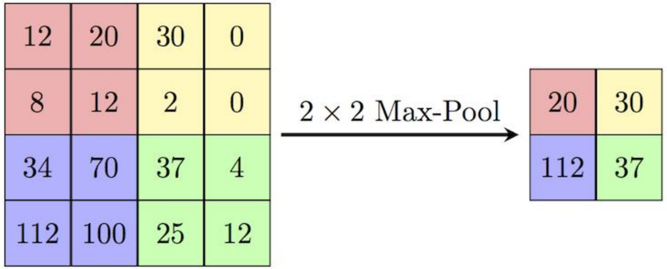
CNN

Max Pooling

Drop Out

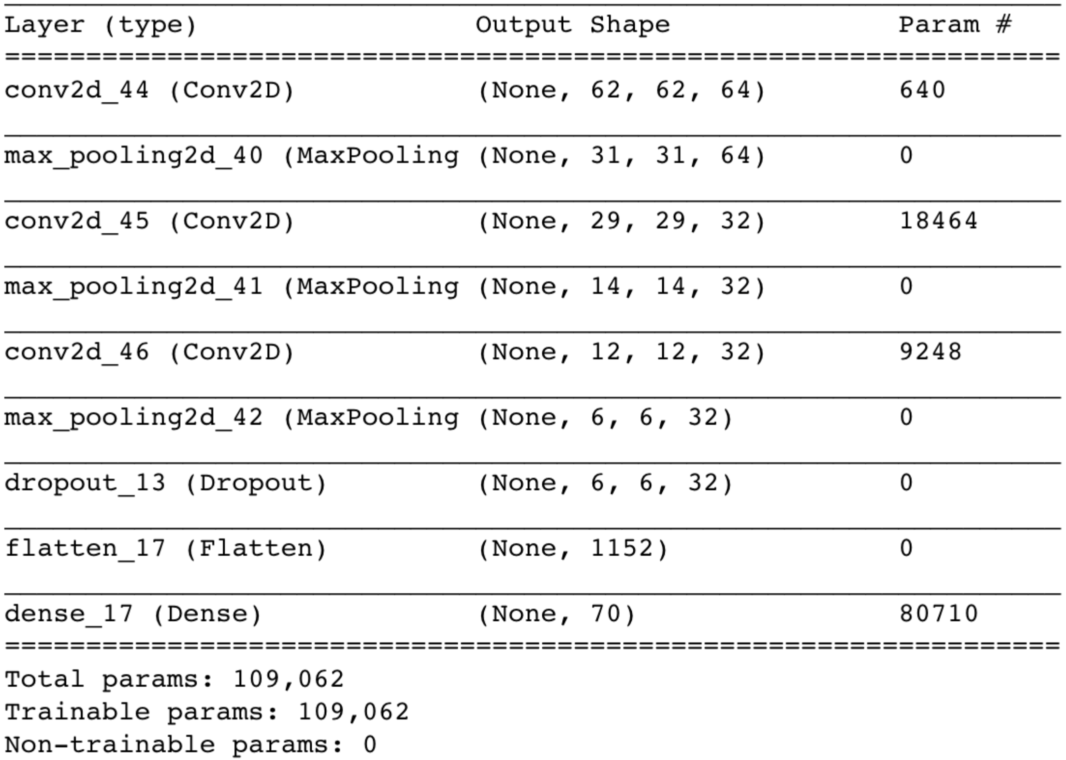
Categorical Cross Entropy





12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
conv2d_44 (Conv2D)           (None, 62, 62, 64)        640       
_________________________________________________________________
max_pooling2d_40 (MaxPooling (None, 31, 31, 64)        0         
_________________________________________________________________
conv2d_45 (Conv2D)           (None, 29, 29, 32)        18464     
_________________________________________________________________
max_pooling2d_41 (MaxPooling (None, 14, 14, 32)        0         
_________________________________________________________________
conv2d_46 (Conv2D)           (None, 12, 12, 32)        9248      
_________________________________________________________________
max_pooling2d_42 (MaxPooling (None, 6, 6, 32)          0         
_________________________________________________________________
dropout_13 (Dropout)         (None, 6, 6, 32)          0         
_________________________________________________________________
flatten_17 (Flatten)         (None, 1152)              0         
_________________________________________________________________
dense_17 (Dense)             (None, 70)                80710     
=================================================================
Total params: 109,062
Trainable params: 109,062
Non-trainable params: 0
```

Manipulating Our Data

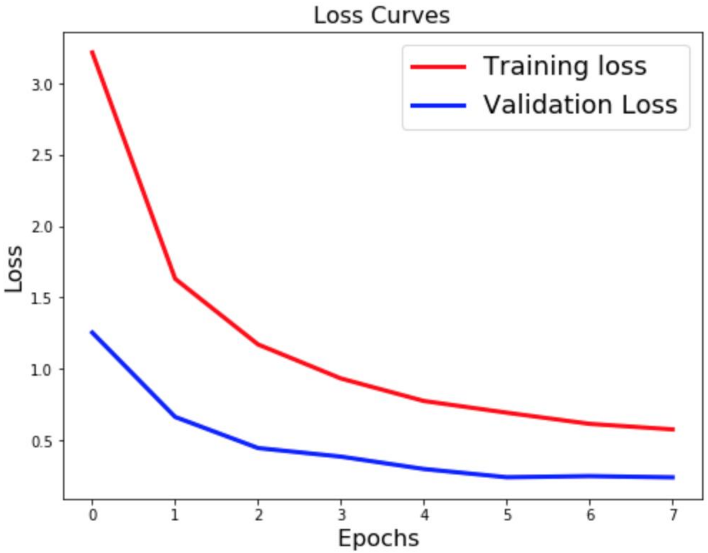
Image Augmentation

Random Shuffling

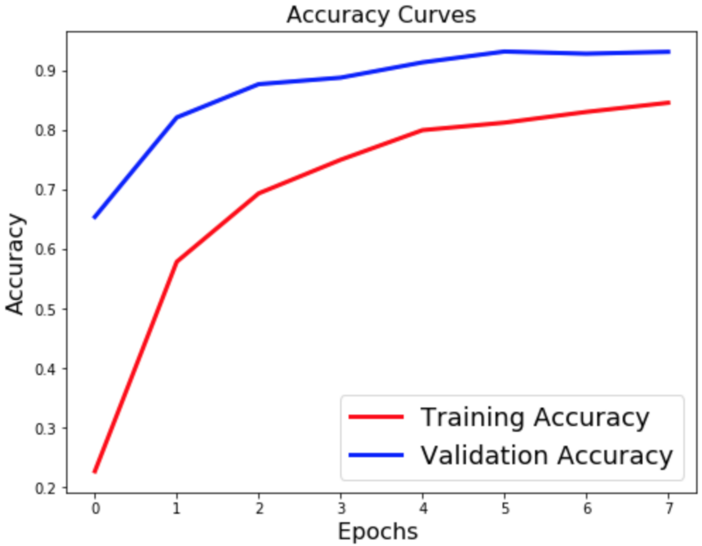
RMSProp vs ADAM

```
ImageDataGenerator(
        zoom_range=0.2,
        width_shift_range=0.1,
        height_shift_range=0.1,
```

Results

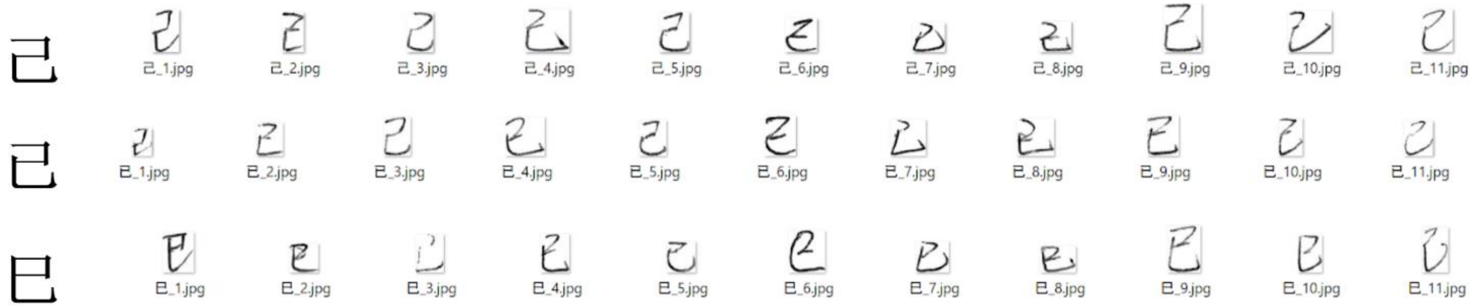


Validation Loss: 0.25



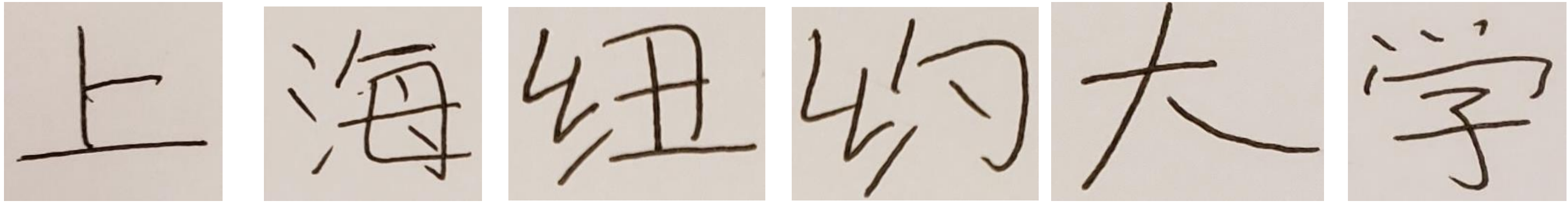
Validation Accuracy: 0.93

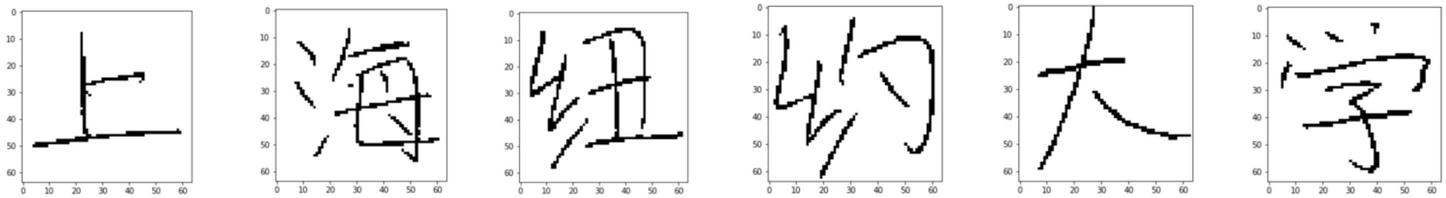
Predicting similar characters

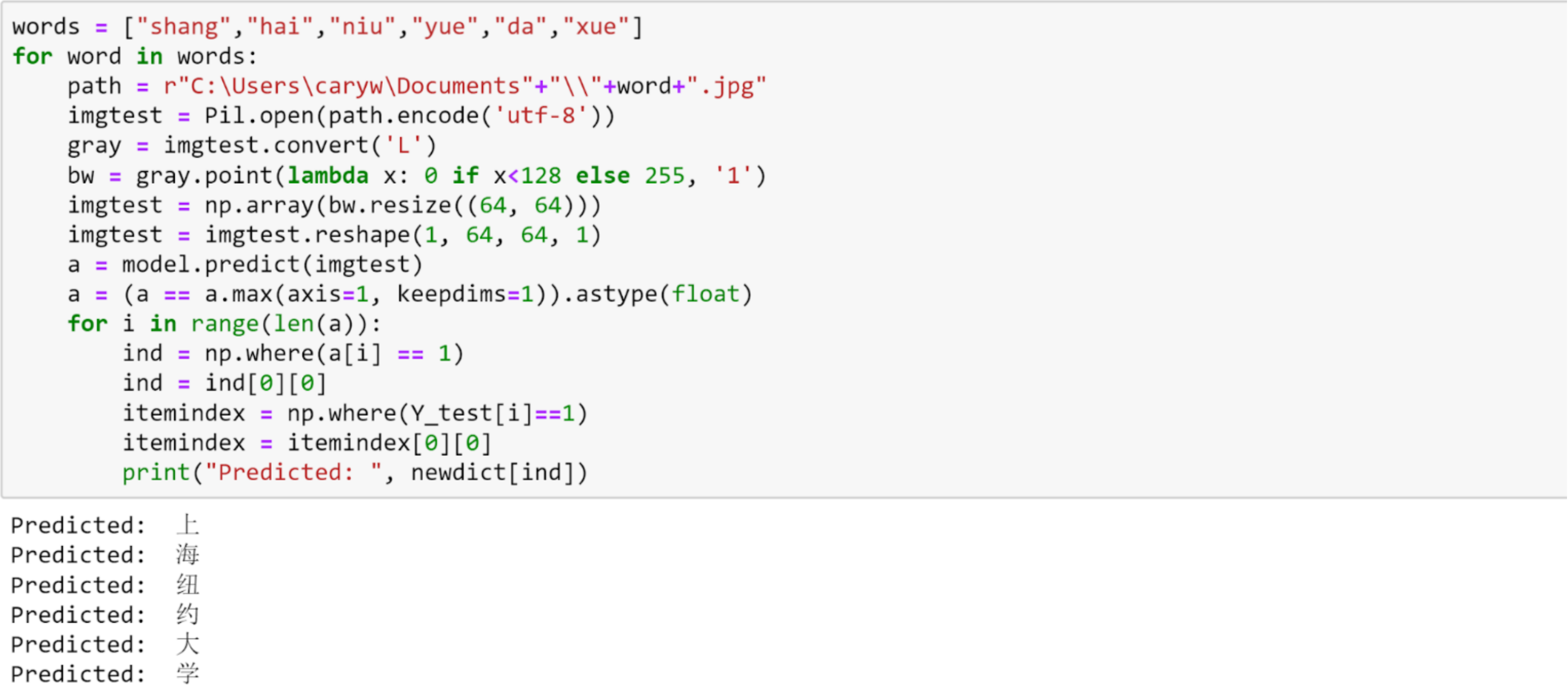


Accuracy for predicting the (己 Jǐ ，已 Yǐ，巳 Sì)has prediction accuracy of 63.8%

Predicting Our Own Handwriting







```python
words = ["shang","hai","niu","yue","da","xue"]
for word in words:
    path = r"C:\Users\caryw\Documents"+"\\"+word+".jpg"
    imgtest = Pil.open(path.encode('utf-8'))
    gray = imgtest.convert('L')
    bw = gray.point(lambda x: 0 if x<128 else 255, '1')
    imgtest = np.array(bw.resize((64, 64)))
    imgtest = imgtest.reshape(1, 64, 64, 1)
    a = model.predict(imgtest)
    a = (a == a.max(axis=1, keepdims=1)).astype(float)
    for i in range(len(a)):
        ind = np.where(a[i] == 1)
        ind = ind[0][0]
        itemindex = np.where(Y_test[i]==1)
        itemindex = itemindex[0][0]
        print("Predicted: ", newdict[ind])
```

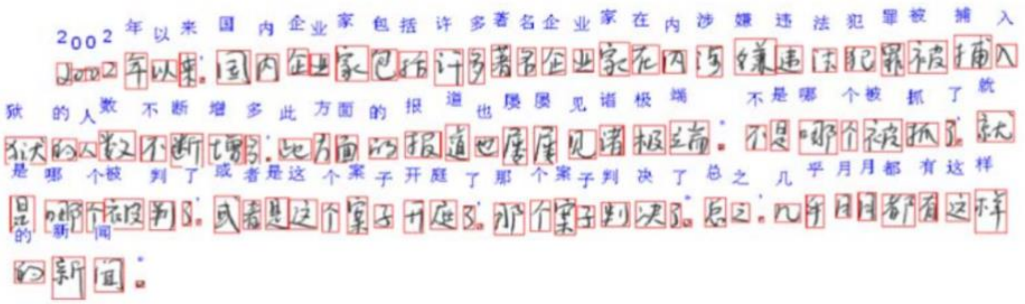
```
Predicted:  上
Predicted:  海
Predicted:  纽
Predicted:  约
Predicted:  大
Predicted:  学
```

If we had infinite time and a supercomputer...

Using more samples and predicting more characters

Open CV (Open Source Computer Vision Library)

Predicting sentences and even full on paragraphs of text


2002年以来国内企业家包括许多著名企业家在内涉嫌违法犯罪被捕入
狱的人数不断增多此方面的报道也屡屡见诸极端不是哪个被抓了就
是哪个被判了或者是这个案子开庭了那个案子判决了总之几乎月月都有这样
的新闻

References

http://cs231n.github.io/convolutional-networks/

https://computersciencewiki.org/index.php/Max-pooling_/_Pooling

www.nlpr.ia.ac.cn/databases/download/ICDAR2011-CASIA databases.pdf. (database)

https://en.wikipedia.org/wiki/Pleco_Software (pleco logo)

https://www.researchgate.net/figure/Three-Chinese-characters-written-in-different-styles_fig1_261499993

https://www.farmcottages.com/new-translator-widget/google-translate-logo/(google translate logo)

https://www.semanticscholar.org/paper/Recognition-of-Handwritten-Similar-Chinese-by-Fu-Xu/124812ab483bc718aedb4036f3f8595dc51b3ffe/figure/0