“Quick, Draw!” Doodle Recognition

Xiaomeng Shen, xshen10@staford.edu, Fangqin Dai, fdai@staford.edu

Predicting:

- Building a classifier for doodle images in 340 classes; Help improve pattern recognition
- Best CNN model is ResNet50, acc 0.932, best RNN model ConvLSTM2D, acc 0.895

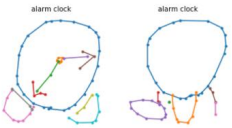
Data & Features

- Training data: 49 million doodling data with 340 classes generated by “Quick, Draw!”;
- Validation set: 5000 samples; Test dataset: 112k samples; Evaluation metrics: MAP@3
- Data format: time series vector drawing with label

	countrycode	drawing	key_id	recognized	timestamp	word
0	US	[[[111, 148, 161, 175, 199, 218, 231, 236, 234...	5159910851477504	TRUE	2017-03-21 13:02:16.246170	alarm clock
1	US	[[[154, 144, 129, 86, 66, 45, 45, 50, 76, 111,...	4608088873107456	TRUE	2017-03-01 21:42:04.745090	alarm clock

format of the drawing array is as following:

```
[
  [  // First stroke
    [x0, x1, x2, x3, ...],
    [y0, y1, y2, y3, ...],
    [t0, t1, t2, t3, ...]
  ],
  [  // Second stroke
    [x0, x1, x2, x3, ...],
    [y0, y1, y2, y3, ...],
    [t0, t1, t2, t3, ...]
  ],
  ... // Additional strokes
]
```



- Strokes converted to image to apply CNN

Models:

- CNN, loss function softmax

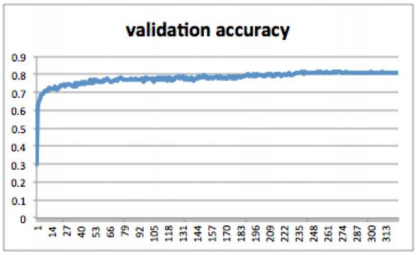
type	filter size	filter num	stride	pad
Conv1 + ReLU	7 x 7	16	(2, 2)	same
Conv2 + ReLU	7 x 7	32	(1, 1)	same
Conv3 + ReLU	7 x 7	48	(1, 1)	same
Maxpool	3 x 3		(2, 2)	same
Conv4 + ReLU	3 x 3	64	(1, 1)	same
Conv5 + ReLU	3 x 3	96	(1, 1)	same
Maxpool	3 x 3		(2, 2)	same
Conv6 + ReLU	3 x 3	128	(1, 1)	same
Conv7 + ReLU	3 x 3	256	(1, 1)	same
Conv8 + ReLU	3 x 3	512	(1, 1)	same
Maxpool	3 x 3		(2, 2)	same
Flatten + FC				

Training:

- Train by chunks (size=50k)
- Early stopping when overfit on one chunk, then move to next

Results:

- Val acc ~0.81; MAP@3: 0.85



- Comparison between state-of-the-art performance

model	input size	color	epochs num	MAP@3
MobileNet	128 x 128	gray	2	0.923
MobileNetV2	128 x 128	gray	3	0.922
ResNet50	128 x 128	RGB	4	0.932
DenseNet121	128 x 128	RGB	2	0.929
ConvSTM2D_MobileNet	64 x 64	gray	1	0.898
ConvLSTM2D_Plain	32 x 32	gray	1	0.895

Tricks:

- Encode as much info as possible into images (temporal, speed)
- cv2 is 8x faster than pillow, json.loads is 10 faster than ast.literal_eval
- Preprocess csv files to a single HDF5 file
- Generate images on the fly instead of saving to disk

Future:

- Sketch-R2CNN / Ensemble all the models

Reference: ResNet, ConvLSTM2D, Sketch-a-Net, Sketch-R2CNN