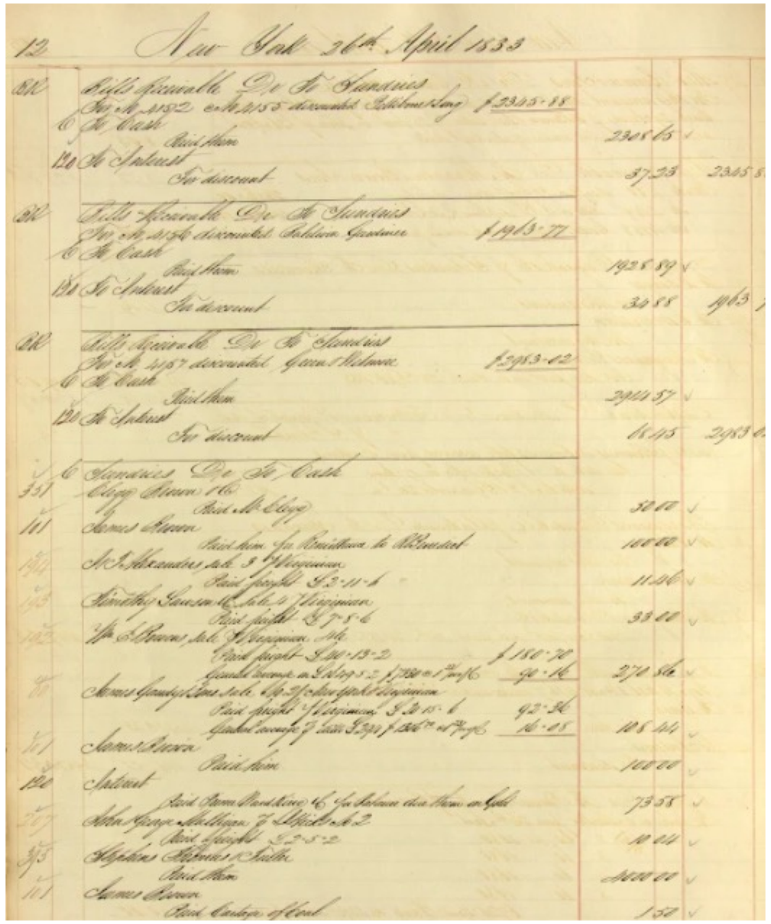
RENDERING LEGIBLE THE HISTORY OF AMERICAN CAPITALISM

SEQUENTIAL HANDWRITTEN TEXT RECOGNITION WITH CRNN-CTC NETWORK

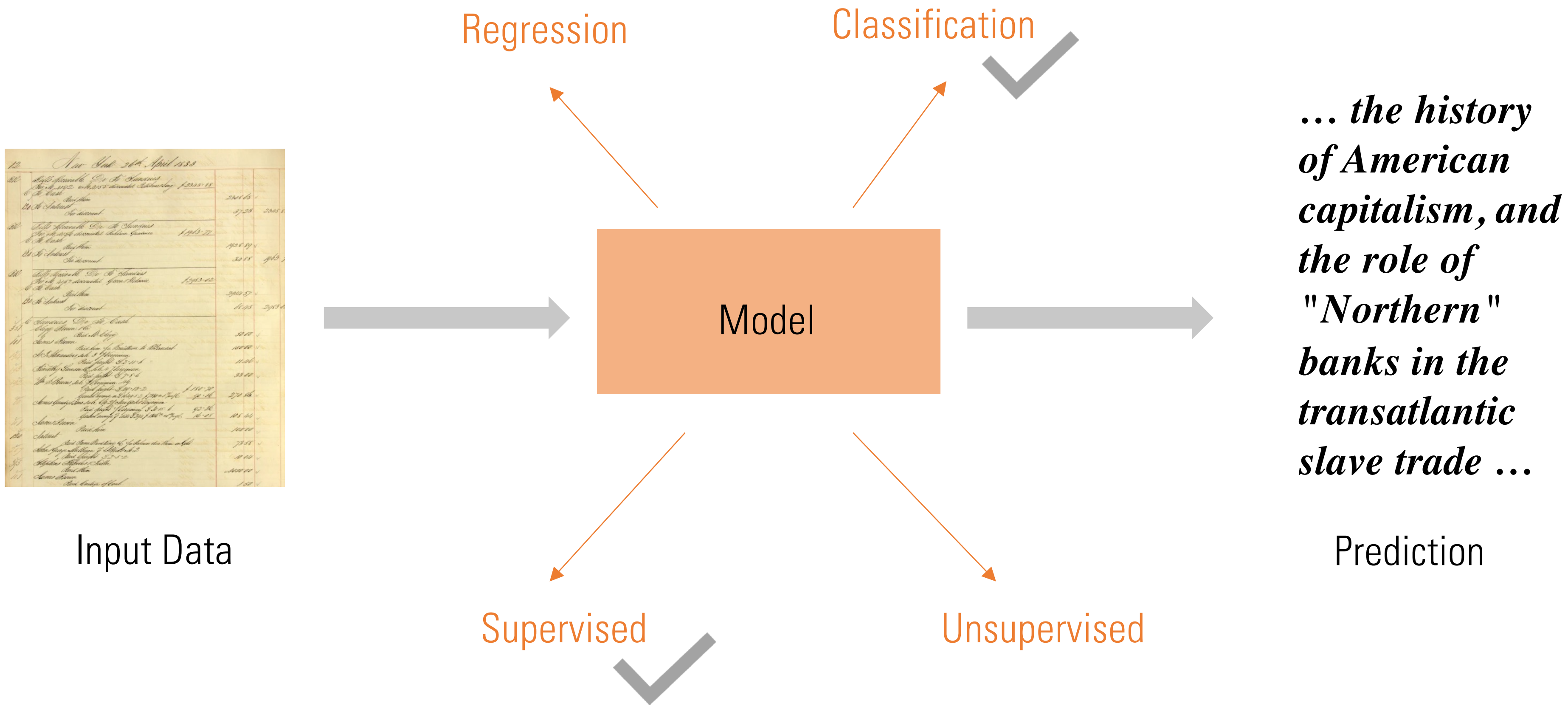
MOTIVATION

- In 2019, the New York Public Library digitized over 110,000 pages from the Brown Brothers & Company papers. A rich source of information on the workings of New York City finance, the history of American capitalism, and the role of "Northern" banks in the transatlantic slave trade, the Brown Brothers Collection has been vastly underutilized, in part because it is written in nineteenth-century longhand.



example ledger image

MACHINE LEARNING PIPELINE



Regression

Classification ✓

Input Data

Model

… the history of American capitalism, and the role of "Northern" banks in the transatlantic slave trade …

Prediction

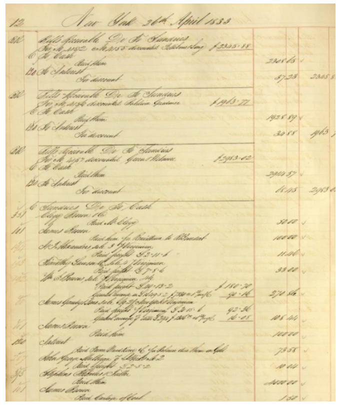
Supervised ✓

Unsupervised

DATASET

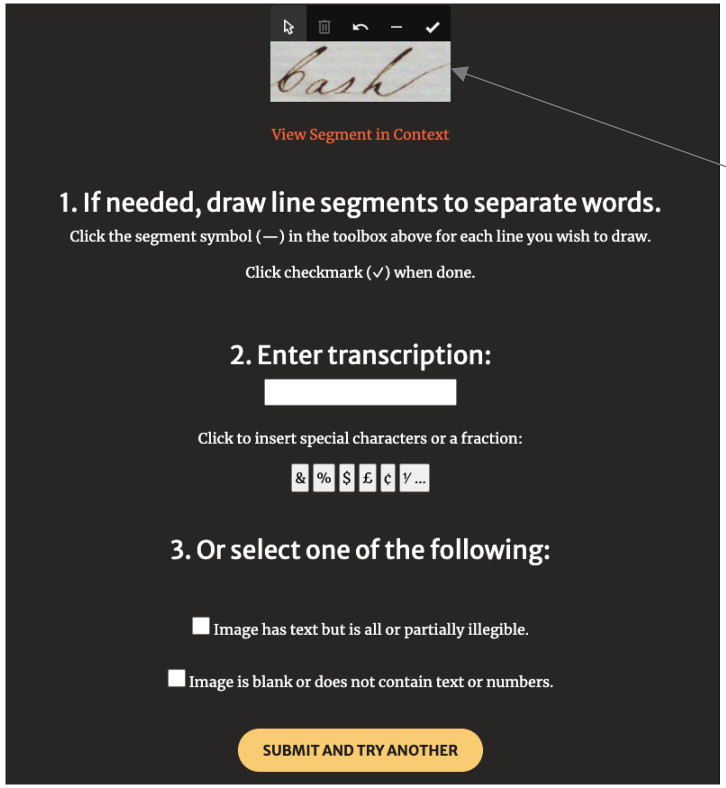
1. How to obtain the dataset?
2. How to convert the dataset into something that the machine can better understand?
3. What if the dataset is not large enough?

DATASET



amazon

segmentation tool



View Segment in Context

1. If needed, draw line segments to separate words.

Click the segment symbol (—) in the toolbox above for each line you wish to draw.

Click checkmark (✓) when done.

2. Enter transcription:

Click to insert special characters or a fraction:

& % $ £ ¢ ⅟…

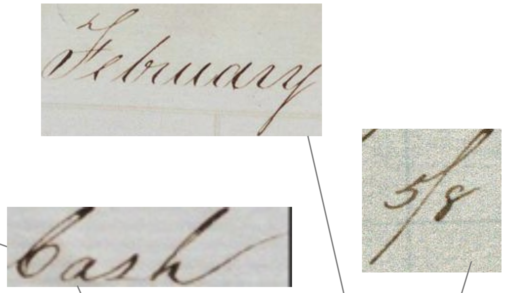
3. Or select one of the following:

□ Image has text but is all or partially illegible.

□ Image is blank or does not contain text or numbers.

SUBMIT AND TRY ANOTHER

transcribers

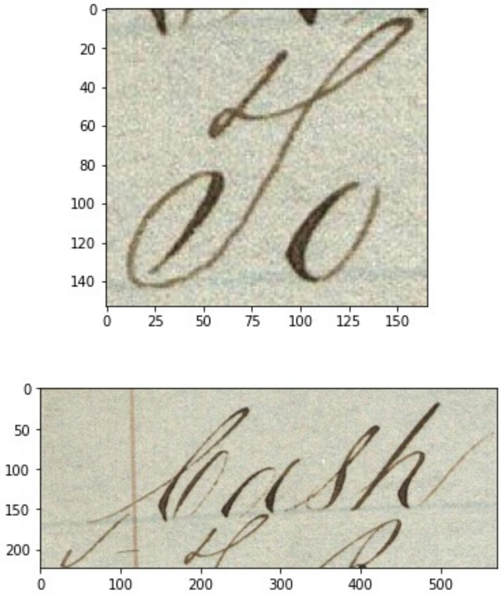


id	name	label
001	xxx.jpg	Feburary
002	xxx.jpg	5/8
...	...	...

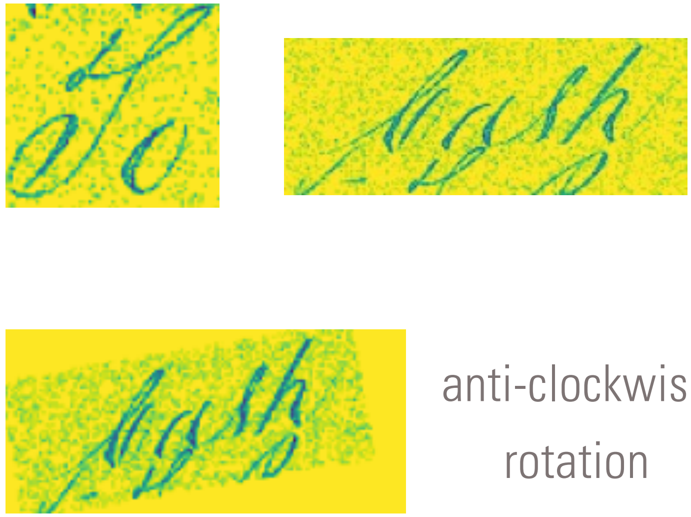
https://brownbros.newyorkscapes.org

credit to Grace Afsari-Mamagani

DATA PREPROCESSING



Raw image



anti-clockwise rotation

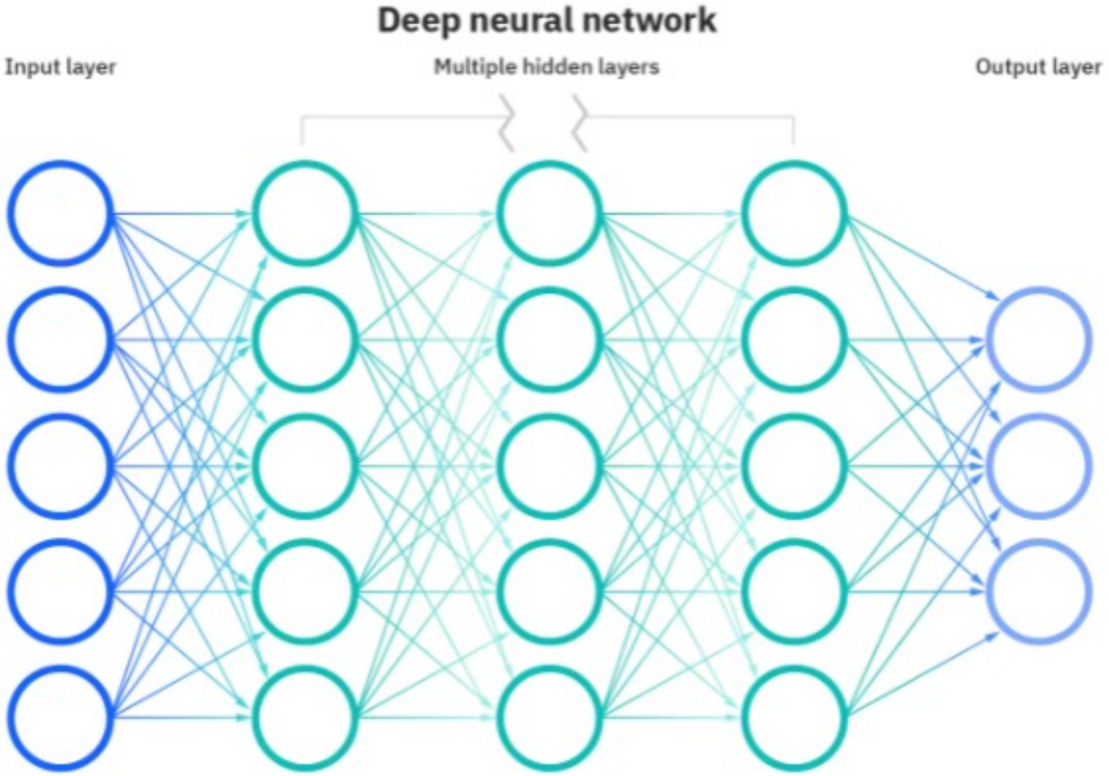
Thresholding + Normalization

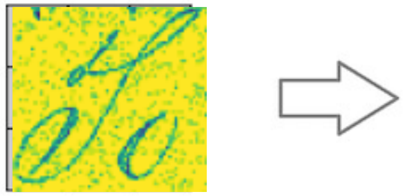
Synthetic image

MODEL

1. What is the model?
2. What is the loss function?
3. How to evaluate the model?

VANILLA NERUAL NETWORK

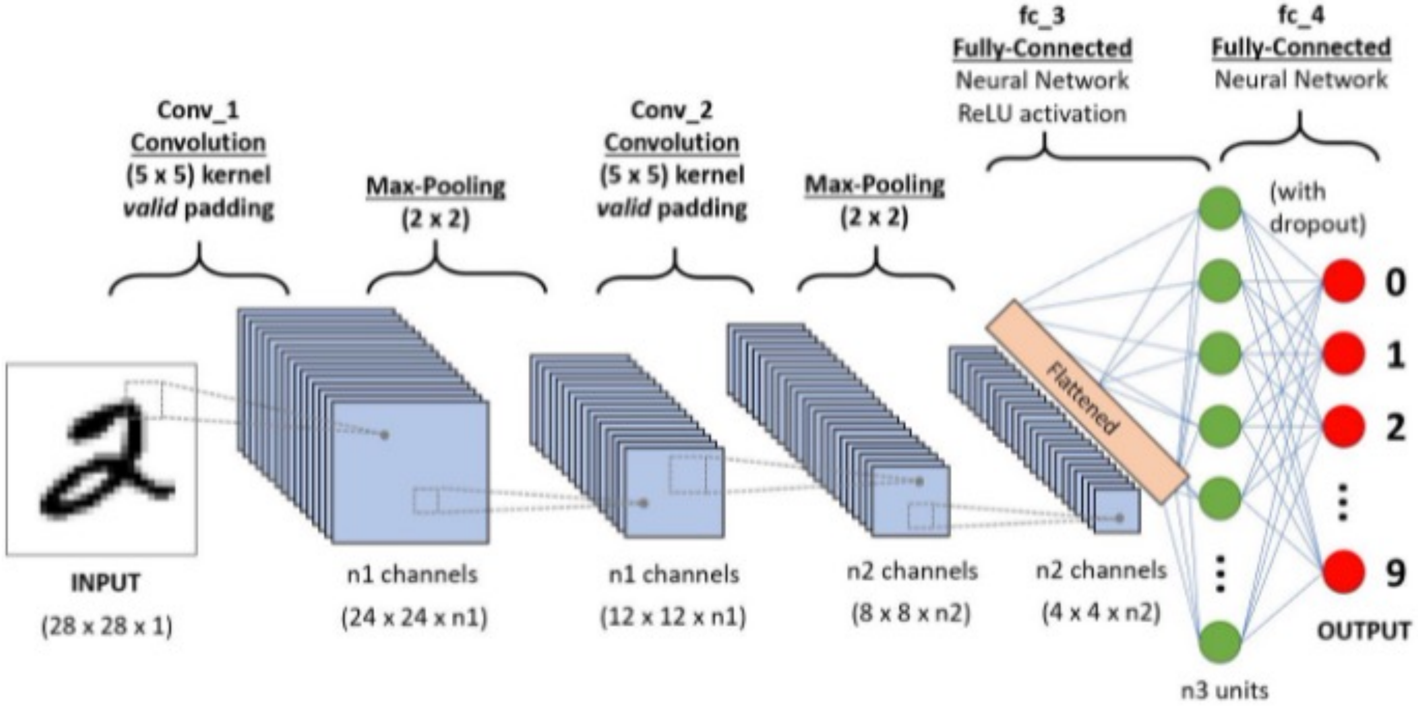


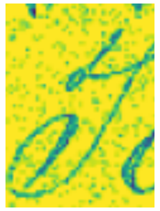


1
1
0
4
2
1
0
2
1

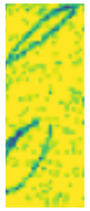
CONVOLUTIONAL NEURAL NETWORK

Capture the Spatial dependencies in an image





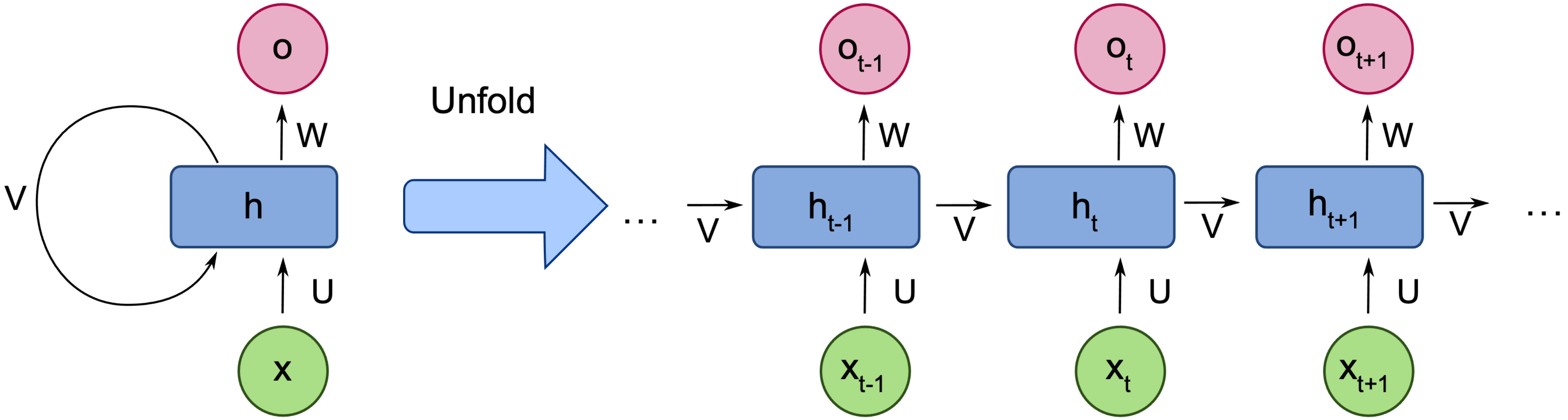
T



o

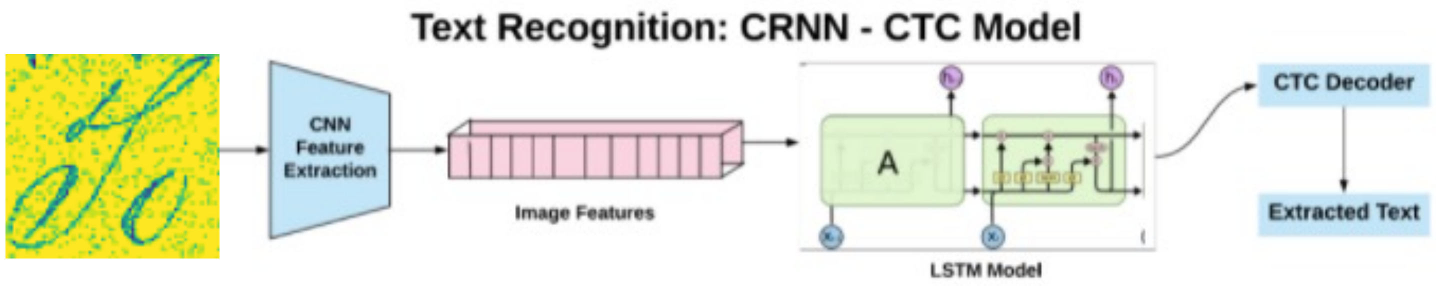
RECURRENT NEURAL NETWORK

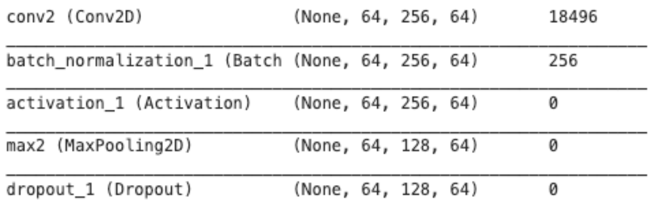
Capture the temporal dependencies in a sequence



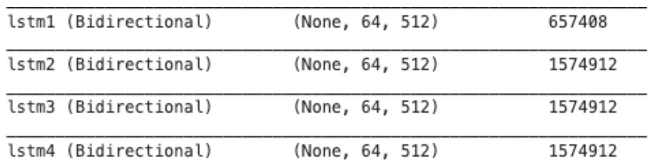
MODEL: CRNN-CTC

- Challenge: The output is a sequence of letters and the lengths may vary drastically.
- Use CNN to extract visual features + LSTM to capture sequential dependencies





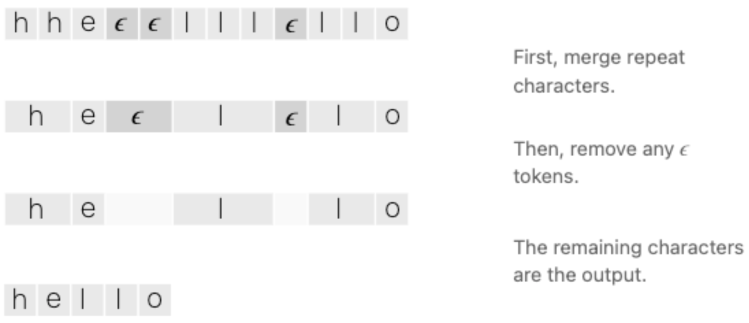
```
conv2 (Conv2D)               (None, 64, 256, 64)       18496
_________________________________________________________________
batch_normalization_1 (Batch (None, 64, 256, 64)       256
_________________________________________________________________
activation_1 (Activation)    (None, 64, 256, 64)       0
_________________________________________________________________
max2 (MaxPooling2D)          (None, 64, 128, 64)       0
_________________________________________________________________
dropout_1 (Dropout)          (None, 64, 128, 64)       0
```



```
_________________________________________________________________
lstm1 (Bidirectional)        (None, 64, 512)           657408
_________________________________________________________________
lstm2 (Bidirectional)        (None, 64, 512)           1574912
_________________________________________________________________
lstm3 (Bidirectional)        (None, 64, 512)           1574912
_________________________________________________________________
lstm4 (Bidirectional)        (None, 64, 512)           1574912
```

figure adapted from Weights & Biases Articles

CONNECTIONIST TEMPORAL CLASSIFICATION (CTC)

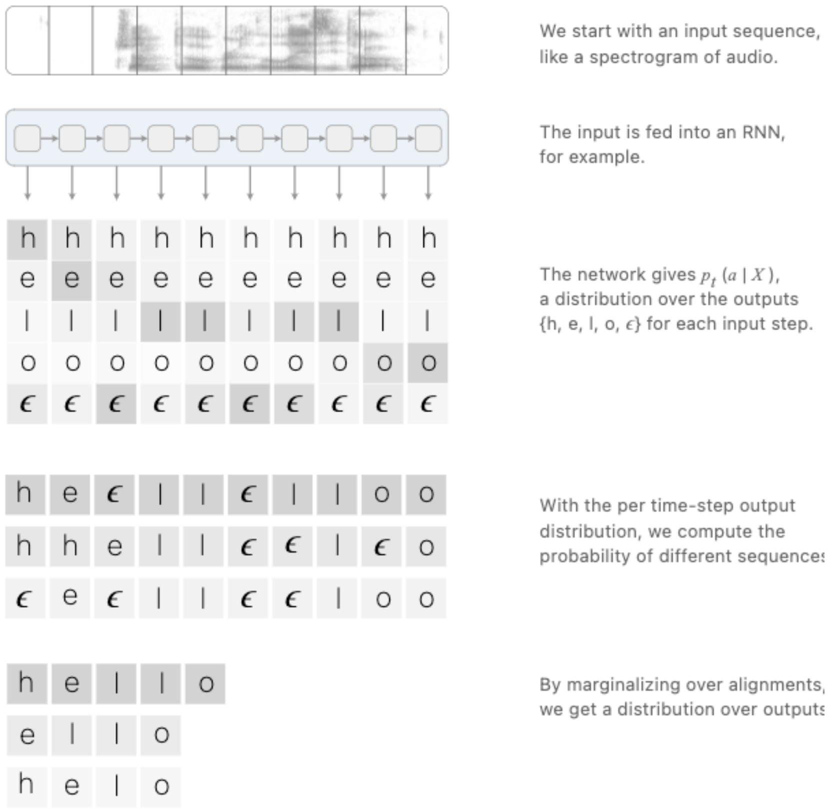


First, merge repeat characters.

Then, remove any ϵ tokens.

The remaining characters are the output.

alignment



We start with an input sequence, like a spectrogram of audio.

The input is fed into an RNN, for example.

The network gives $p_t(a \mid X)$, a distribution over the outputs {h, e, l, o, ϵ} for each input step.

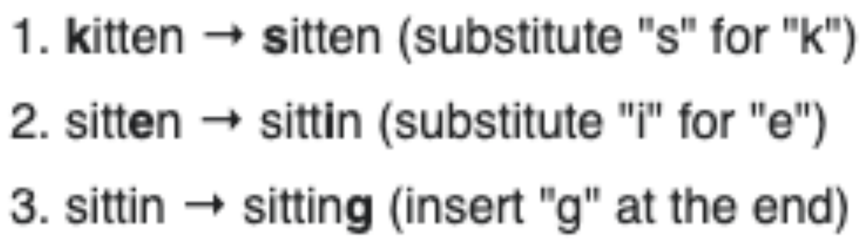
With the per time-step output distribution, we compute the probability of different sequences

By marginalizing over alignments, we get a distribution over outputs

loss

figure from Distill

METRICS

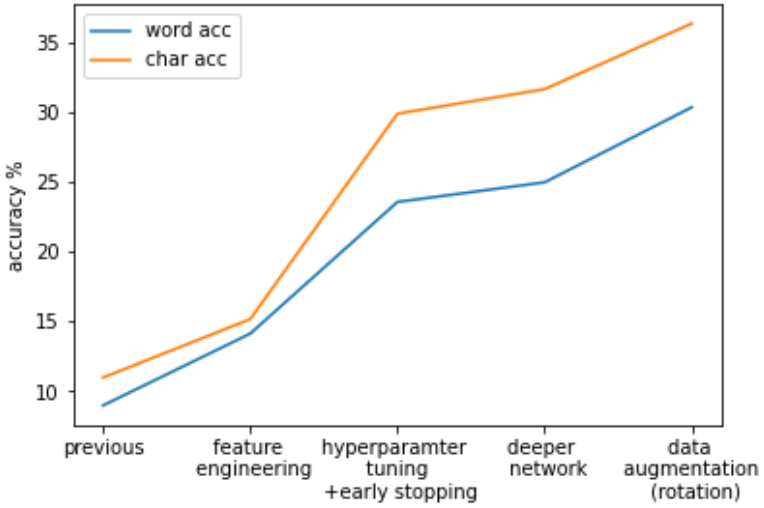


1. kitten → sitten (substitute "s" for "k")
2. sitten → sittin (substitute "i" for "e")
3. sittin → sitting (insert "g" at the end)

1. Word-level accuracy
2. Character-level accuracy
3. Edit distance
   - The Levenshtein distance allows deletion, insertion and substitution.
   - The Longest common subsequence (LCS) distance allows only insertion and deletion, not substitution.
   - The Hamming distance allows only substitution, hence, it only applies to strings of the same length.
   - The Damerau–Levenshtein distance allows insertion, deletion, substitution, and the transposition of two adjacent characters.
   - The Jaro distance allows only transposition.

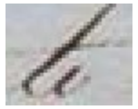
RESULTS

- Improve the word-level accuracy from 9% to 30.38%, char-level accuracy from 11% to 36.39%.

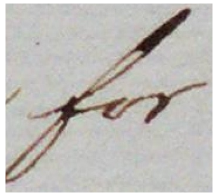


Correct examples

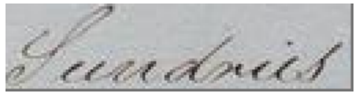
(short and/or frequent words)



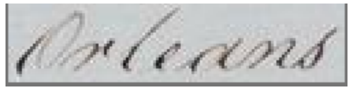
to



for



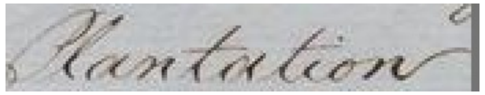
Sundries



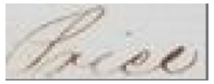
Orleans

Incorrect examples

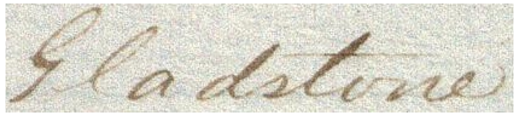
(Long, infrequent, blurry words)



Plantation (pred: lPlantation)



Price (pred: Brer)



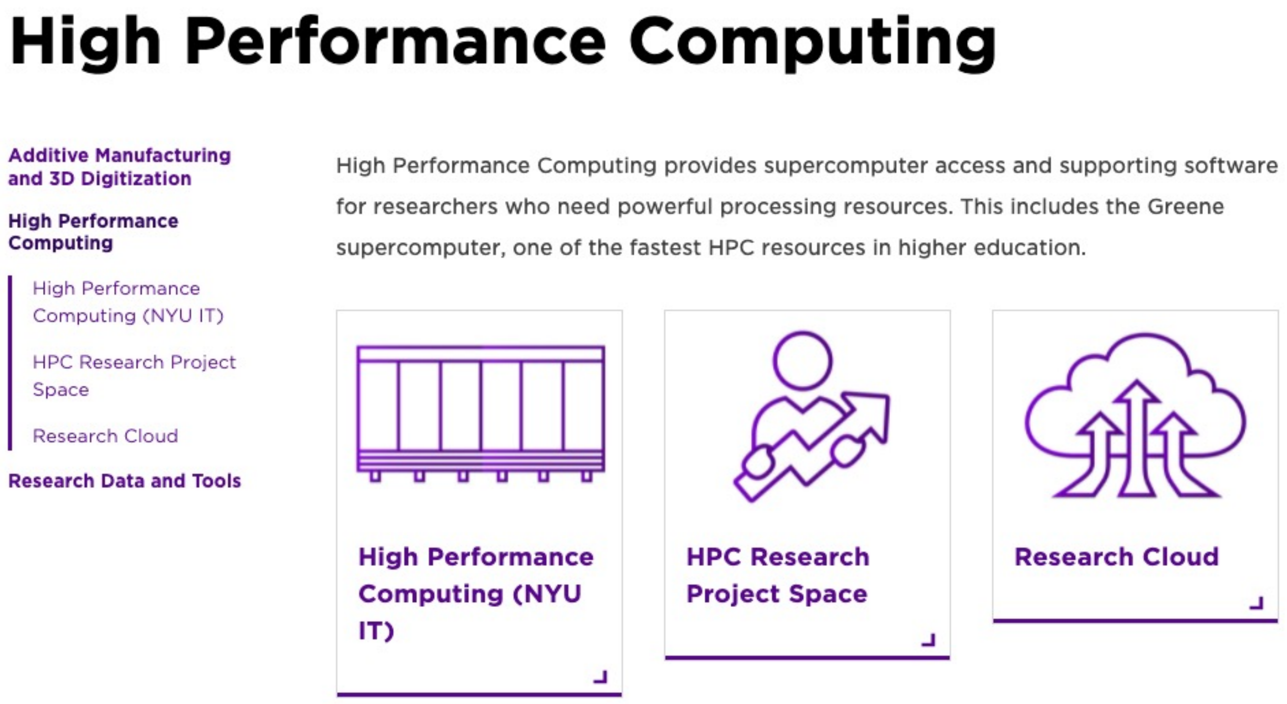
Gladstone (pred: Cawan)

CHALLENGE & FUTURE PLAN

1. The current preprocessed image still have a lot of noise
   - find a better preprocessing method to denoise, or
   - add some noise to the synthetic data to mimic the real-world images?
2. Numbers are important for analyzing financial activities, but the accuracy for them are low due to the lack of data (only 10% of the train set has numeric labels)
   - Only 10% of data has been transcribed now. Undergrad researchers are still working on transcribing more data this term!
   - Use synthetic images

- 3. Explore different models (e.g. Transformers)

RESOURCES



High Performance Computing

Additive Manufacturing and 3D Digitization

High Performance Computing

- High Performance Computing (NYU IT)
- HPC Research Project Space
- Research Cloud

Research Data and Tools

High Performance Computing provides supercomputer access and supporting software for researchers who need powerful processing resources. This includes the Greene supercomputer, one of the fastest HPC resources in higher education.

High Performance Computing (NYU IT)

HPC Research Project Space

Research Cloud

Both have GPUs!