



RENDERING LEGIBLE THE HISTORY OF AMERICAN CAPITALISM

SEQUENTIAL HANDWRITTEN TEXT RECOGNITION WITH
CRNN-CTC NETWORK

MOTIVATION

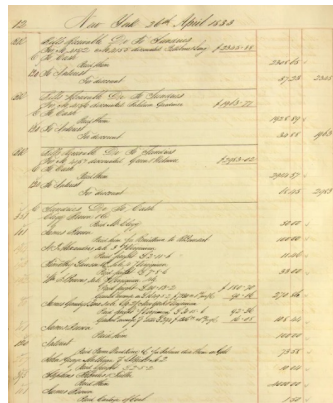
- In 2019, the New York Public Library digitized over 110,000 pages from the **Brown Brothers & Company** papers. A rich source of information on the workings of New York City finance, the history of American capitalism, and the role of "Northern" banks in the transatlantic slave trade, the Brown Brothers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

New York 26th April 1833

12	Dr. To Henry Jones		
13	Dr. To Henry Jones		
14	Dr. To Henry Jones		
15	Dr. To Henry Jones		
16	Dr. To Henry Jones		
17	Dr. To Henry Jones		
18	Dr. To Henry Jones		
19	Dr. To Henry Jones		
20	Dr. To Henry Jones		
21	Dr. To Henry Jones		
22	Dr. To Henry Jones		
23	Dr. To Henry Jones		
24	Dr. To Henry Jones		
25	Dr. To Henry Jones		
26	Dr. To Henry Jones		
27	Dr. To Henry Jones		
28	Dr. To Henry Jones		
29	Dr. To Henry Jones		
30	Dr. To Henry Jones		
31	Dr. To Henry Jones		
32	Dr. To Henry Jones		
33	Dr. To Henry Jones		
34	Dr. To Henry Jones		
35	Dr. To Henry Jones		
36	Dr. To Henry Jones		
37	Dr. To Henry Jones		
38	Dr. To Henry Jones		
39	Dr. To Henry Jones		
40	Dr. To Henry Jones		
41	Dr. To Henry Jones		
42	Dr. To Henry Jones		
43	Dr. To Henry Jones		
44	Dr. To Henry Jones		
45	Dr. To Henry Jones		
46	Dr. To Henry Jones		
47	Dr. To Henry Jones		
48	Dr. To Henry Jones		
49	Dr. To Henry Jones		
50	Dr. To Henry Jones		
51	Dr. To Henry Jones		
52	Dr. To Henry Jones		
53	Dr. To Henry Jones		
54	Dr. To Henry Jones		
55	Dr. To Henry Jones		
56	Dr. To Henry Jones		
57	Dr. To Henry Jones		
58	Dr. To Henry Jones		
59	Dr. To Henry Jones		
60	Dr. To Henry Jones		
61	Dr. To Henry Jones		
62	Dr. To Henry Jones		
63	Dr. To Henry Jones		
64	Dr. To Henry Jones		
65	Dr. To Henry Jones		
66	Dr. To Henry Jones		
67	Dr. To Henry Jones		
68	Dr. To Henry Jones		
69	Dr. To Henry Jones		
70	Dr. To Henry Jones		
71	Dr. To Henry Jones		
72	Dr. To Henry Jones		
73	Dr. To Henry Jones		
74	Dr. To Henry Jones		
75	Dr. To Henry Jones		
76	Dr. To Henry Jones		
77	Dr. To Henry Jones		
78	Dr. To Henry Jones		
79	Dr. To Henry Jones		
80	Dr. To Henry Jones		
81	Dr. To Henry Jones		
82	Dr. To Henry Jones		
83	Dr. To Henry Jones		
84	Dr. To Henry Jones		
85	Dr. To Henry Jones		
86	Dr. To Henry Jones		
87	Dr. To Henry Jones		
88	Dr. To Henry Jones		
89	Dr. To Henry Jones		
90	Dr. To Henry Jones		
91	Dr. To Henry Jones		
92	Dr. To Henry Jones		
93	Dr. To Henry Jones		
94	Dr. To Henry Jones		
95	Dr. To Henry Jones		
96	Dr. To Henry Jones		
97	Dr. To Henry Jones		
98	Dr. To Henry Jones		
99	Dr. To Henry Jones		
100	Dr. To Henry Jones		

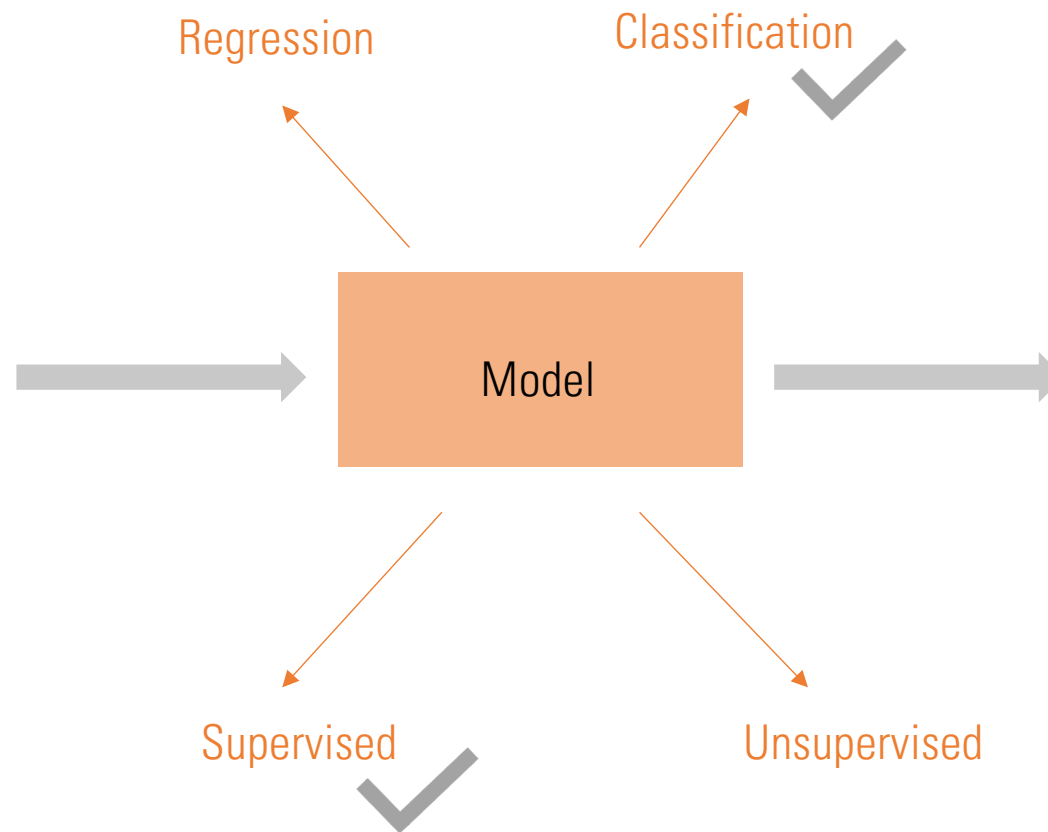
example ledger image

MACHINE LEARNING PIPELINE



The image shows a page from a historical ledger, likely from the 18th or 19th century. It features handwritten entries in cursive script, organized into columns. The entries appear to be financial or administrative records, possibly related to trade or shipping. The paper is aged and yellowed.

Input Data



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

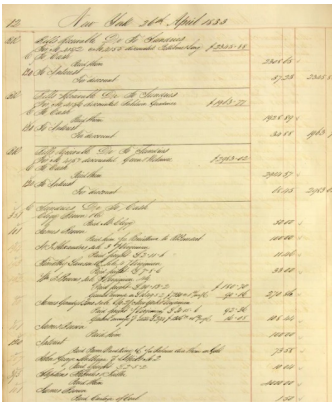
Prediction

The background features several thin, light orange lines that intersect and cross each other in a non-orthogonal manner, creating a modern, geometric aesthetic. These lines are scattered across the white background, with some extending from the edges towards the center.

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:
[text input field]

Click to insert special characters or a fraction:
& % \$ £ ¢ ¥ ...

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.

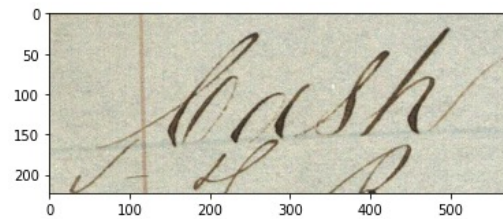
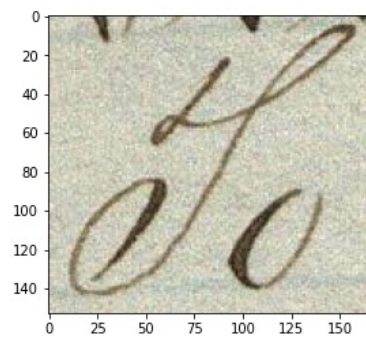
SUBMIT AND TRY ANOTHER

transcribers

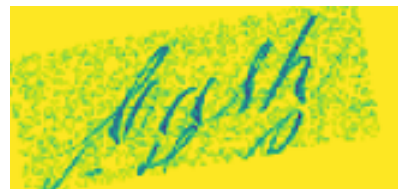
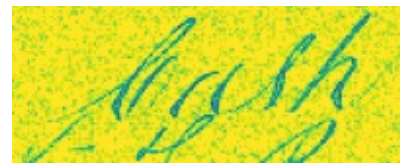
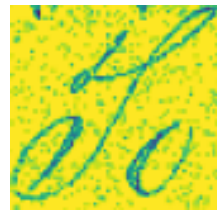
id	name	label
001	xxx.jpg	Feburary
002	xxx.jpg	5/8
...

<https://brownbros.newyorkscapes.org>
credit to Grace Afsari-Mamagani

DATA PREPROCESSING



Raw image



anti-clockwise
rotation

Thresholding +
Normalization



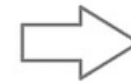
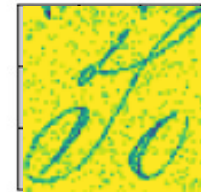
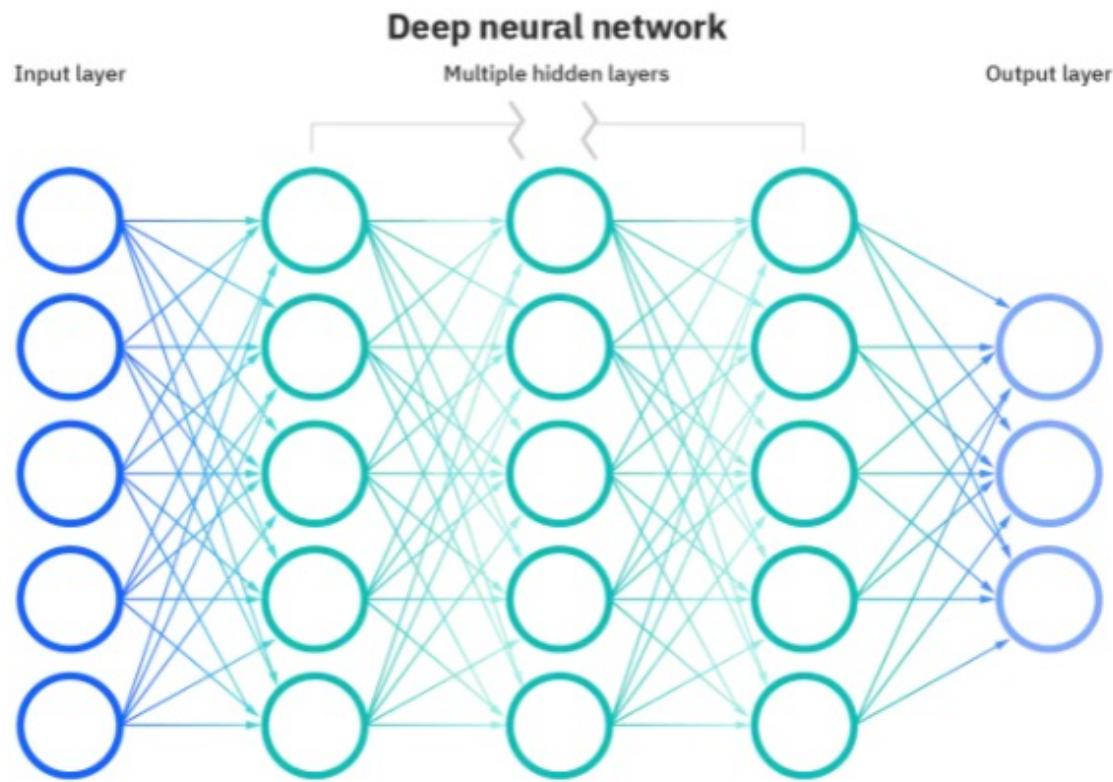
Synthetic image

The slide features a minimalist design with several thin, light-orange lines that intersect to form a series of irregular, open rectangular frames around the text. The word 'MODEL' is centered in the left frame in a large, elegant, black serif font.

MODEL

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

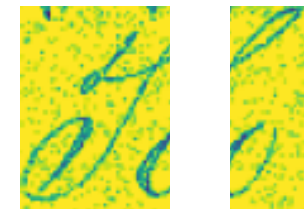
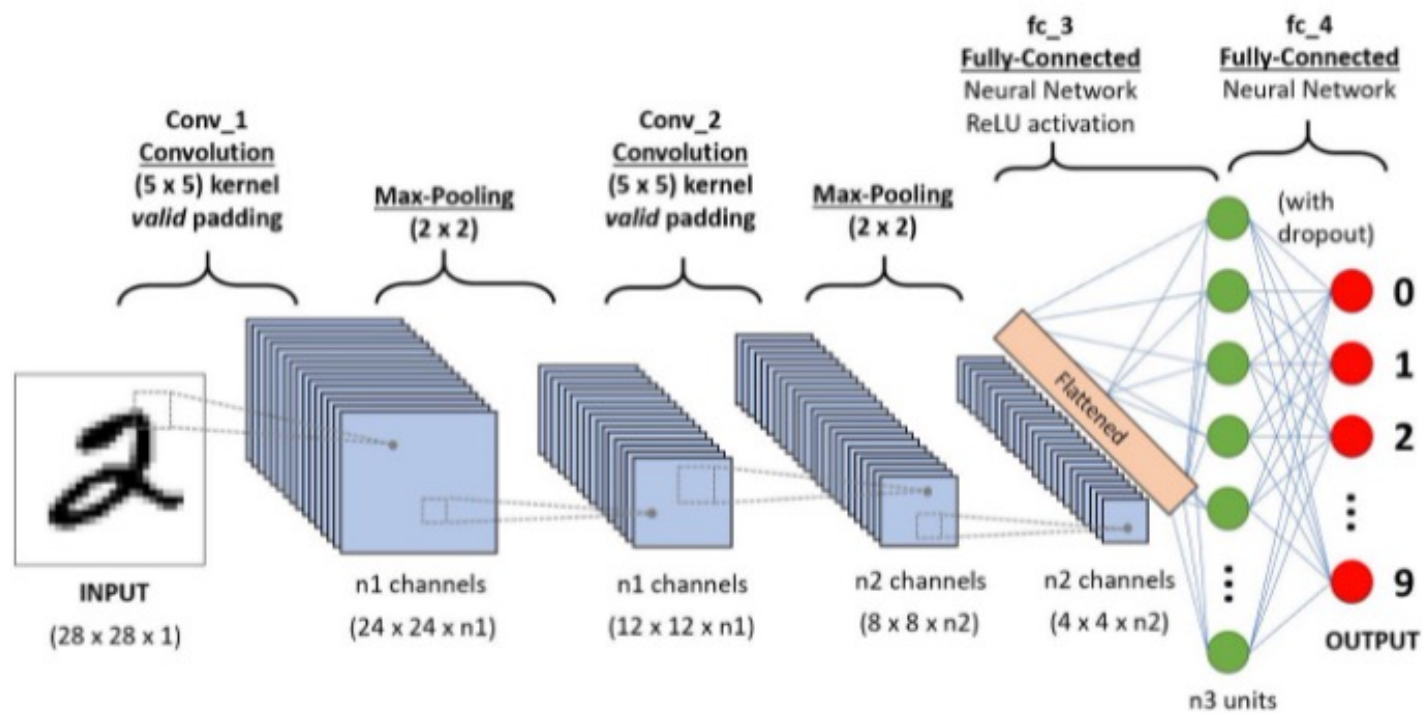
VANILLA NEURAL NETWORK



1
1
0
4
2
1
0
2
1

CONVOLUTIONAL NEURAL NETWORK

Capture the Spatial dependencies in an image

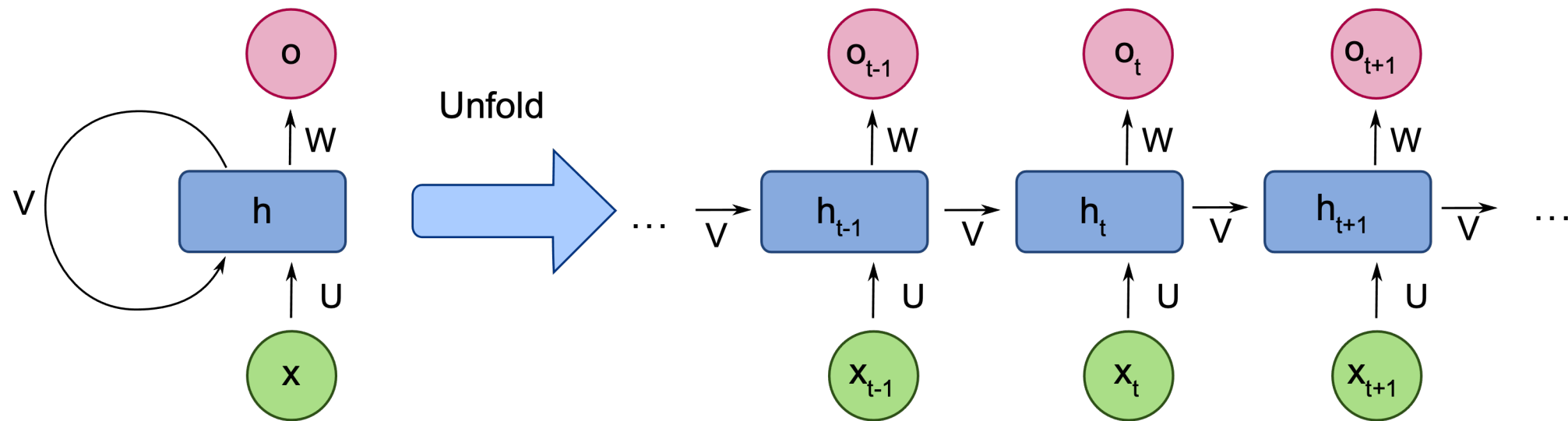


7

0

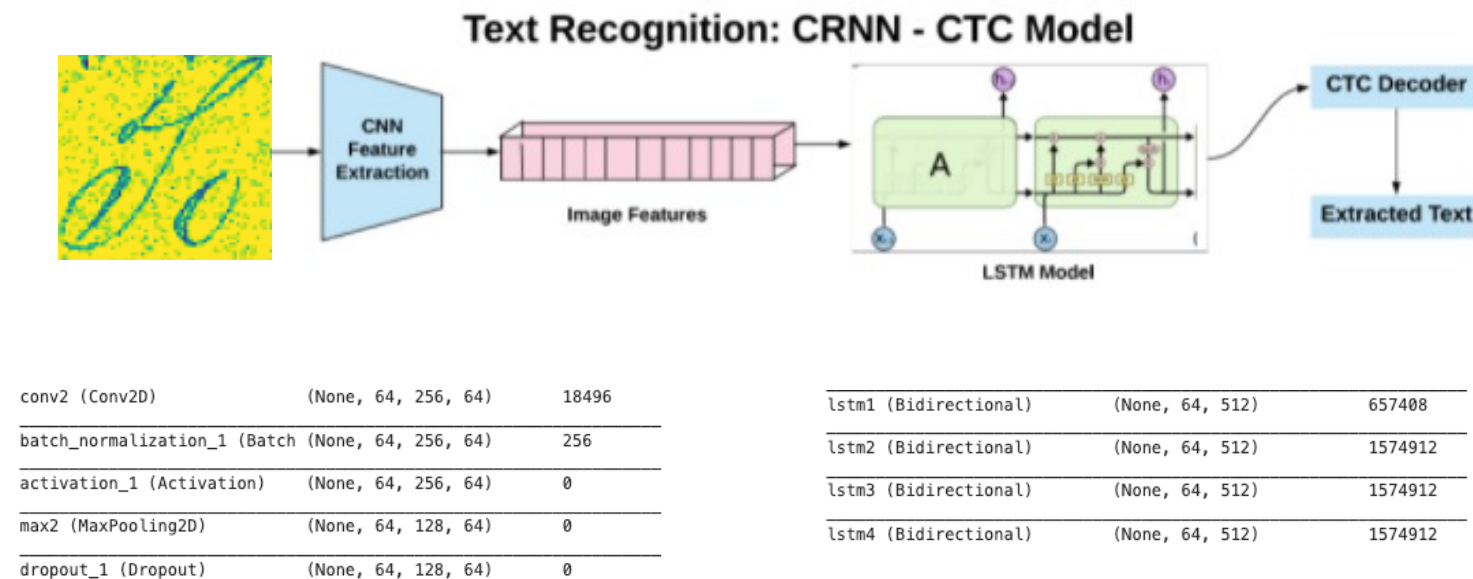
RECURRENT NEURAL NETWORK

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



CONNECTIONIST TEMPORAL CLASSIFICATION (CTC)

h h e € € l l l € l l o

h e € l € l o

h e l l o

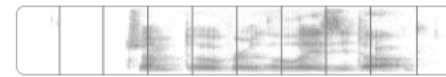
h e l l o

alignment

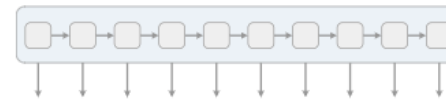
First, merge repeat characters.

Then, remove any € tokens.

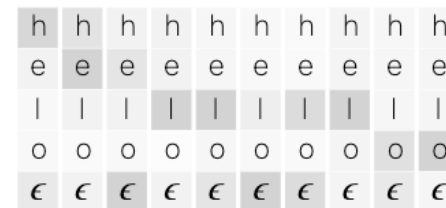
The remaining characters are the output.



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



The input is fed into an RNN, for example.



The network gives $p_t(a | X)$, a distribution over the outputs {h, e, l, o, €} 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

METRICS

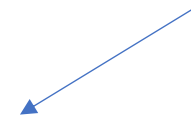
1. Word-level accuracy
2. Character-level accuracy
3. Edit distance

- 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.

1. **kitten** → **sitten** (substitute "s" for "k")

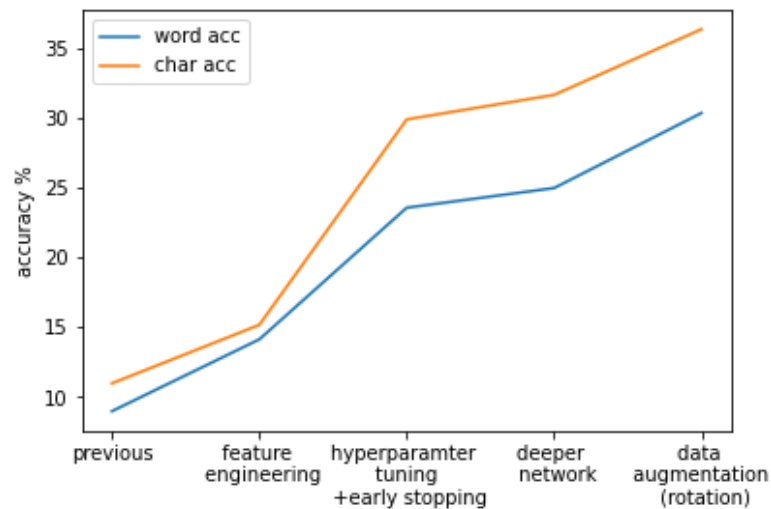
2. **sitten** → **sittin** (substitute "i" for "e")

3. **sittin** → **sitting** (insert "g" at the end)



RESULTS

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

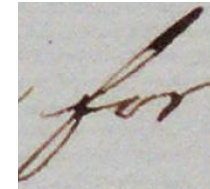


Correct examples

(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

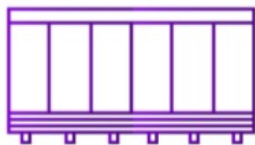
High Performance
Computing (NYU IT)

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!