Interpreting Deep-learning Models: An Exploratory Study with TypeNet for Keystroke Dynamics

Mia Onodera[1], Charles Devlen[2], Daqing Hou[2]

[1]Department of Computer and Electrical Engineering, University of Washington

[2]Department of Computer and Electrical Engineering, Clarkson University

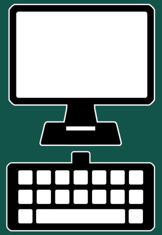
Behavioral Biometrics

As an avenue of cybersecurity, behavioral biometrics are a way to authenticate a user.

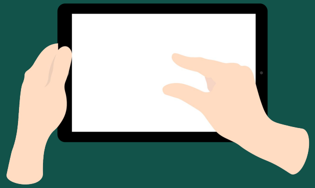
There are many methods of authentication:

Voice

Mouse

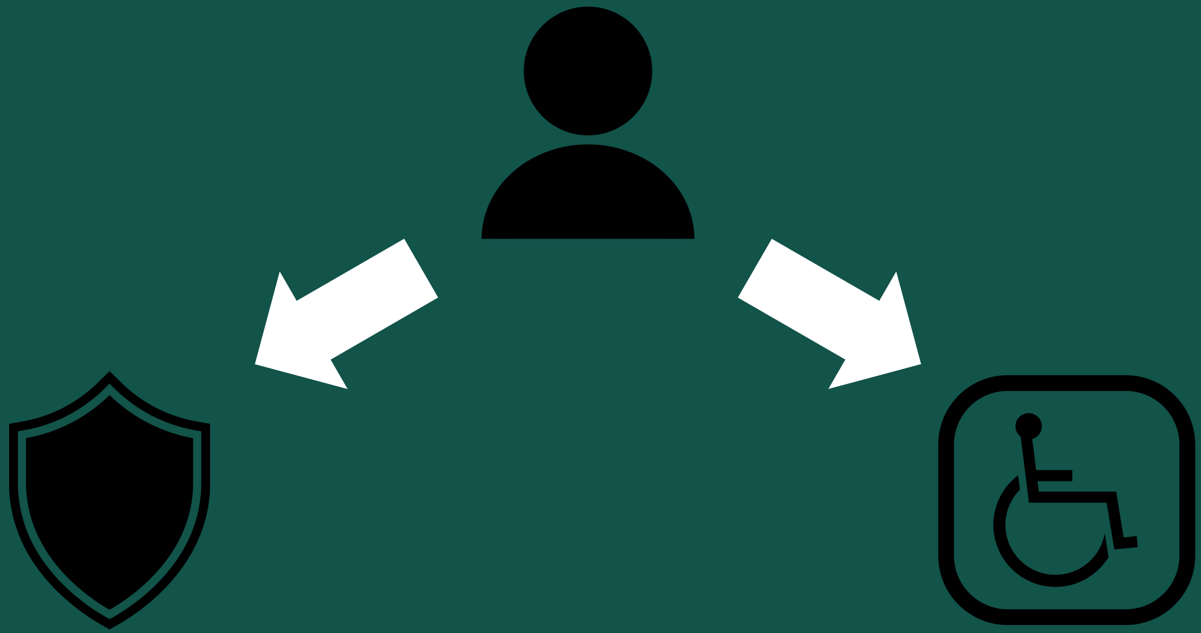


Keystroke



Touch

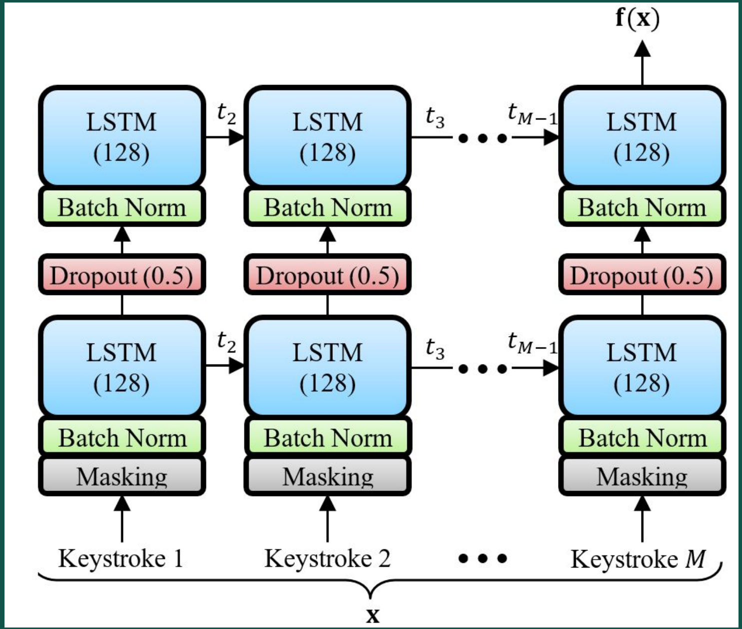
Why Behavioral Biometrics?



TypeNet

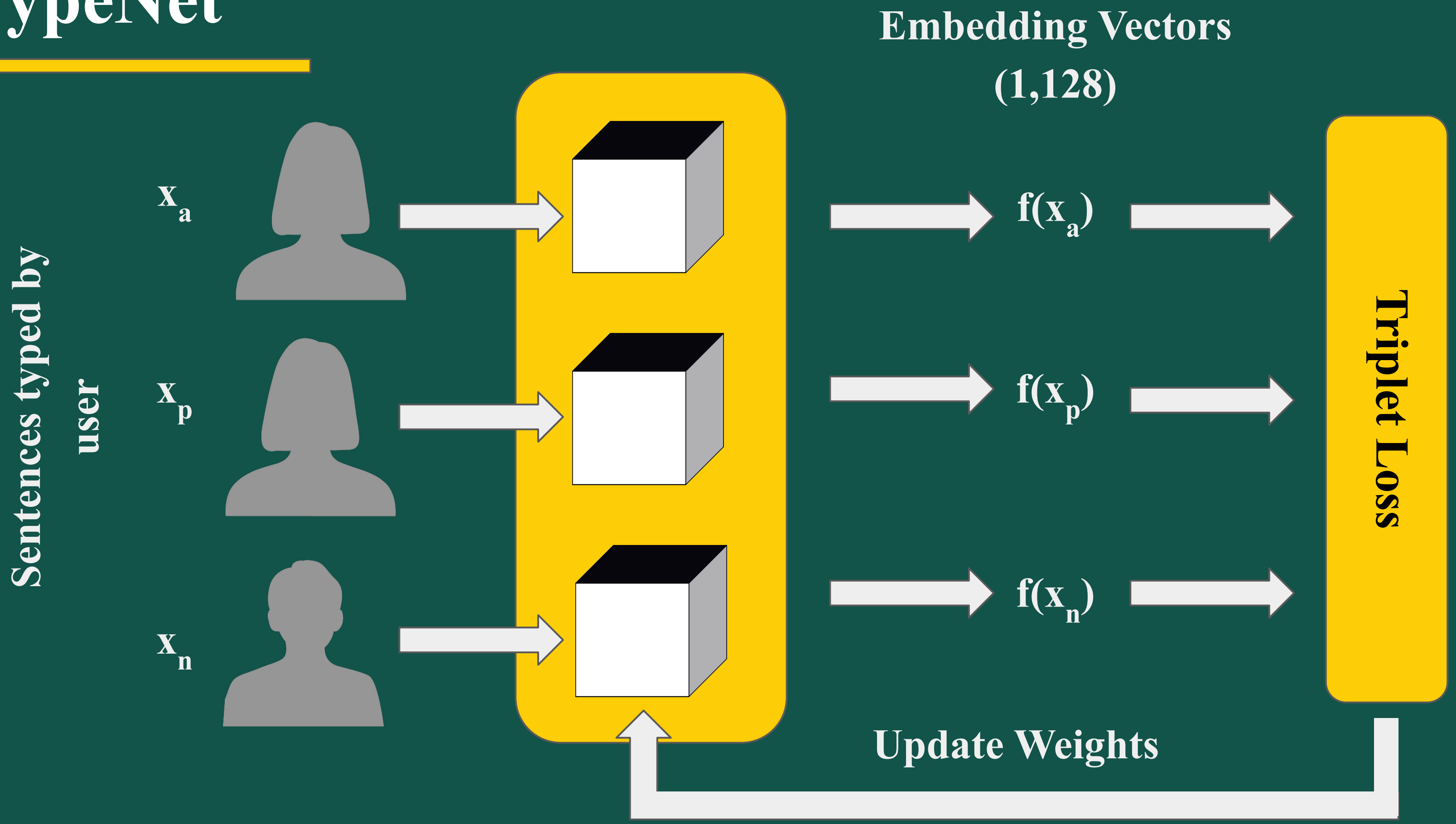
TypeNet is a type of Siamese Neural Network capable of capturing temporal (time-ordered) data, best for free-text sequences.

It has achieved an Equal Error Rate (EER) of 2.2% [1] on the Aalto Desktop dataset.



[1] A. Acien, A. Morales, J. V Monaco, R. Vera-Rodriguez, and J. Fierrez, “TypeNet: Deep Learning Keystroke Biometrics,” *IEEE Trans Biom Behav Identity Sci*, vol. 4, no. 1, p. 57, 2022, doi: 10.1109/TBIOM.2021.3112540.

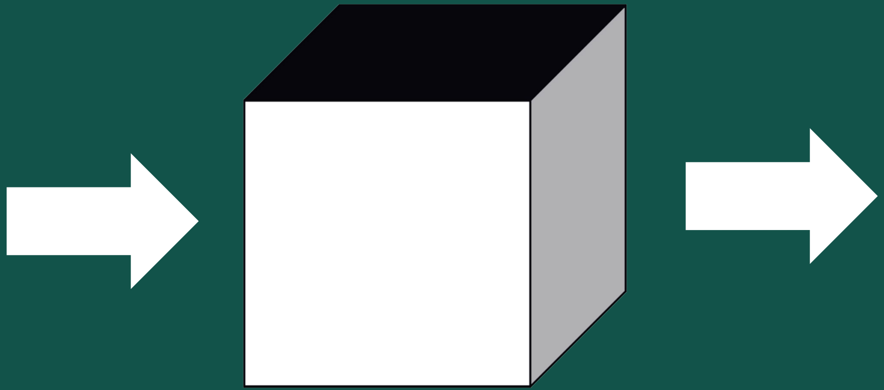
TypeNet


Embedding Vectors
(1,128)
Sentences typed by user
x_a
x_p
x_n
$f(x_a)$
$f(x_p)$
$f(x_n)$
Triplet Loss
Update Weights

Problem Statement

"ECE is the best major."

"Do you agree?"

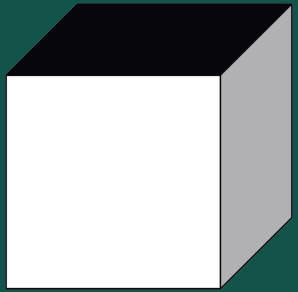


Yes, the input behavior matches the user.

How did this "black box" come to this conclusion?

Why should I trust this model?

Proposed Work



Feature Analysis

Embedding (Output) Analysis

Protect Sensitive Information

Build user trust

Potential Commercialization

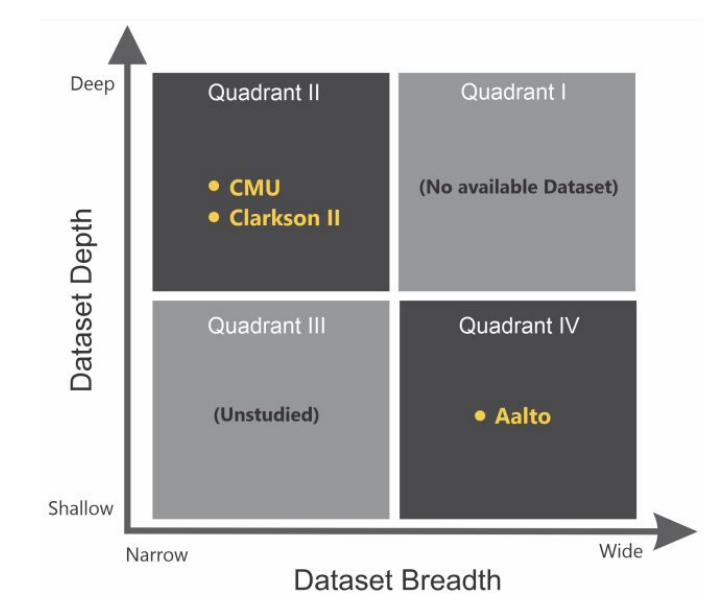
Dataset

Depth -> Amount of data collected per subject

Breadth -> Number of subjects

Both greatly impact model performance [2].

I will opt to use the Clarkson II dataset due to its depth and free-text collection which applies real world situations.



[2] A. A. Wahab and D. Hou, "Impact of Data Breadth and Depth on Performance of Siamese Neural Network Model: Experiments with Two Behavioral Biometric Datasets," 2023 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 2023, doi: 10.1109/BIOSIG58226.2023.10345993.

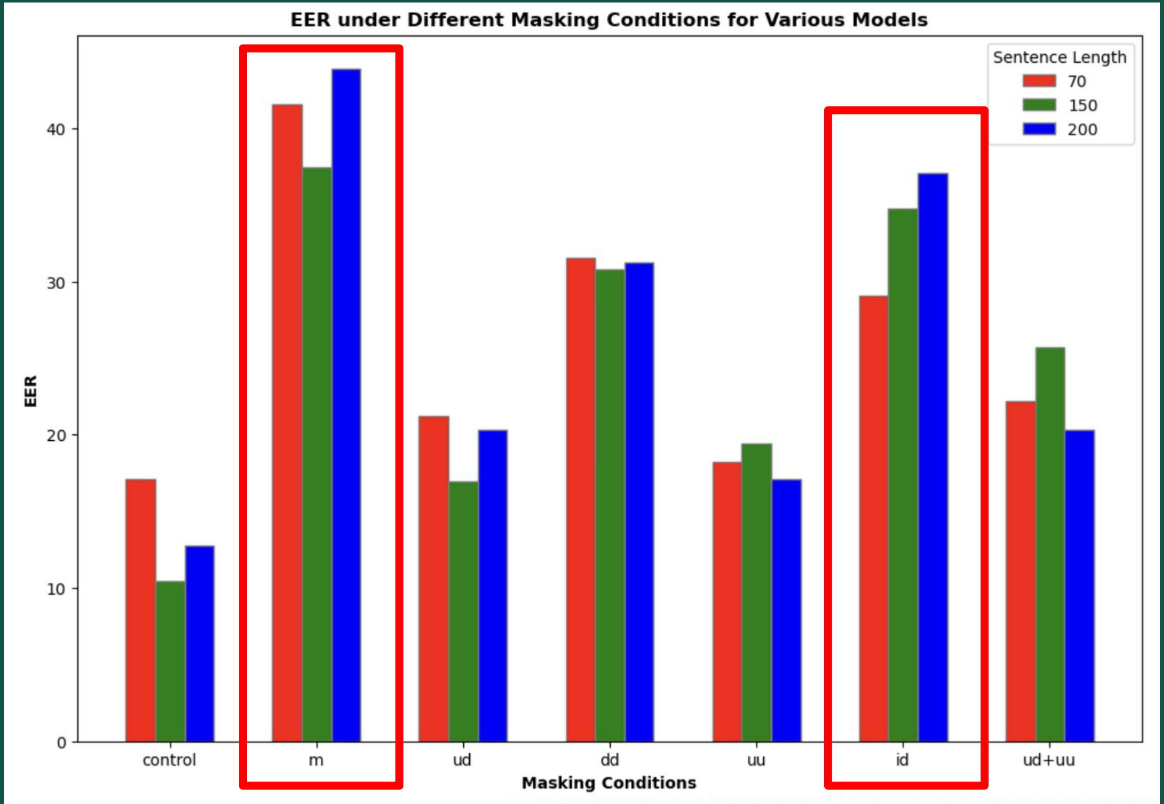
Clarkson II Dataset Preprocessing

Sentence: “The dog jumped”

Keystroke: “Th” + “he” + “e ” + “ d” + “do” + “og” …

- Dwell time (m) - the time spent lingering on the first key, from key press to release/
- Flight time Up-Down (ud) - The time between a key being released (up) until the next key being pressed (down)
- Flight time Down-Down (dd)- The time between one key being pressed (down) until the next key being pressed (down)
- Flight time Up-Up (uu) - The time between one key being released (up) until the next key being released (up)
- ID - ASCII of first key / 256 * ASCII of second key / 256

Feature Analysis



Approach: For every test trial, mask a specific feature / column to 0 then compare EER

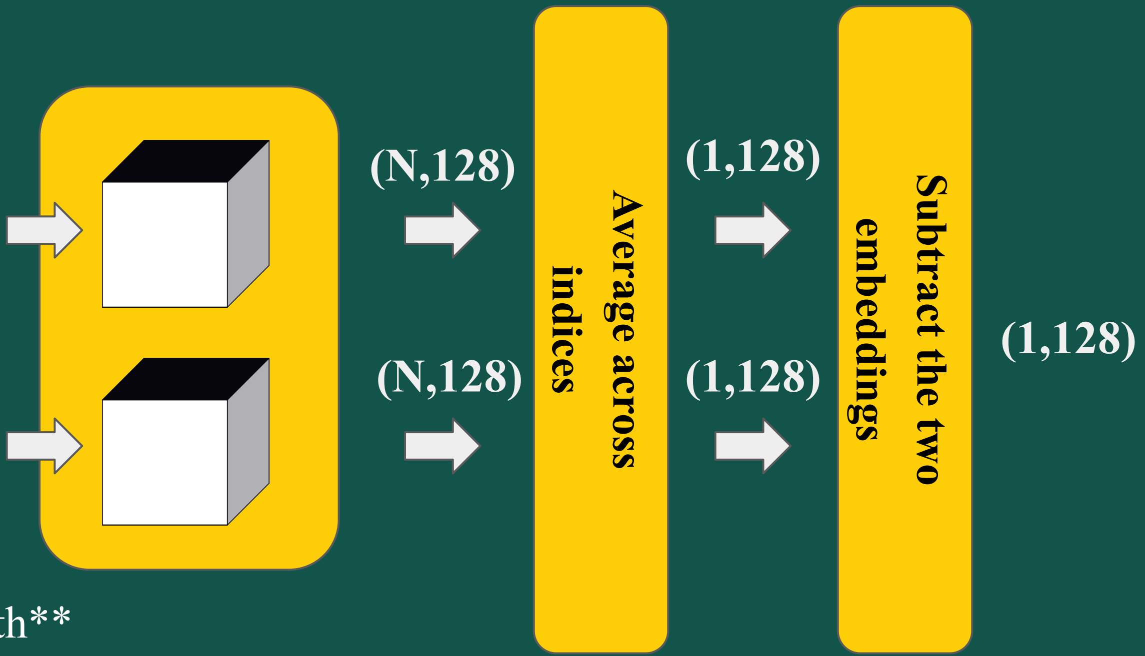
General Consensus: All features are significant.
Dwell time and id seem to be the most significant.

Embedding Analysis

Say for digraph (‘t’, ‘h’), we take every instance of this digraph such as:
[m, ud, dd, uu, Digraph]

[0.048,0.071,0.12,0.15,0.18]

[0.048,0.071,0.119,0.151,0]



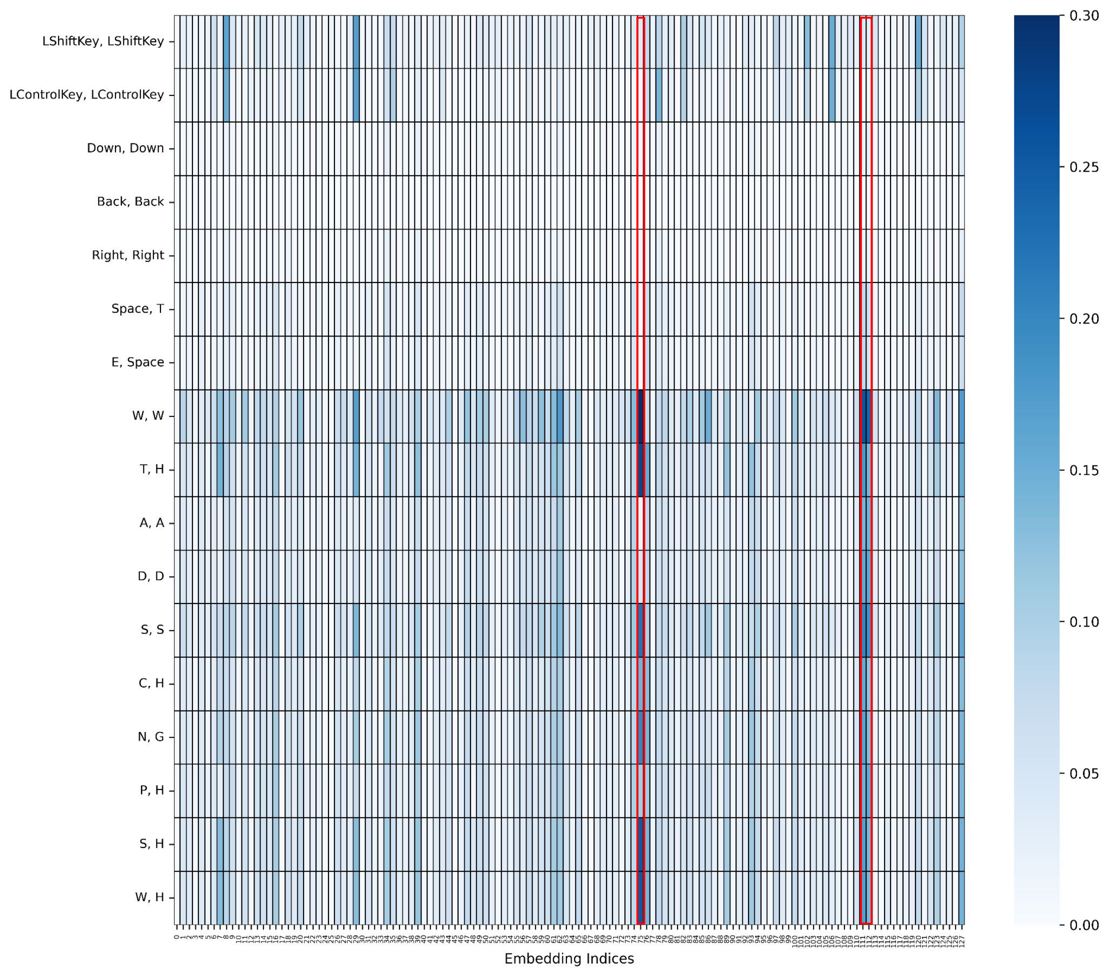
zero pad rest of sentence length**

ID

Heatmap of top 10 digraph embeddings:

Letter digraphs consistently activated compared to control keys.

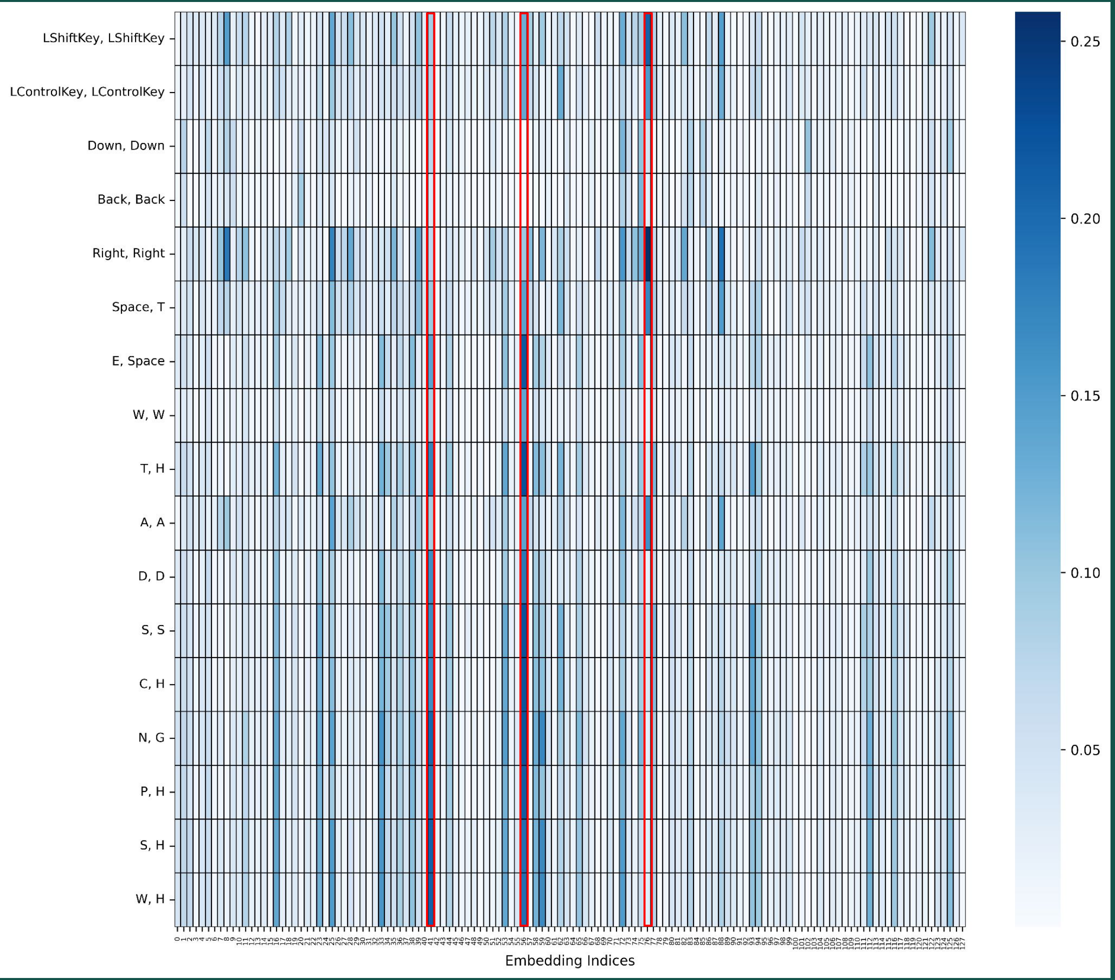
Indices 75, 111, 112 consistently activated.



Dwell Time

Indices 41, 56, 76 consistently activated.

Further solidifies that we have indices corresponding to specific features.



Conclusions and Future Work

Feature Analysis:

- ID then Dwell Time was shown to be the most significant feature within the model.

Embedding Analysis:

- We see consistent indices light up for specific features.
- Letter digraphs are more significant than control keys.

In the future I would like to determine specific indices quantitatively and do a in depth analysis of specific digraphs we can mask to ensure sensitive keystroke data remains secure.

Acknowledgements

This work was supported by the NSF Award #2244049