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

Onodera, Mia C[1], Devlen, Charles[2], Daqing Hou[3]
1. Department of Electrical and Computer Engineering, University of Washington, Seattle, WA
2. Department of Software Engineering, RIT, Rochester, NY
3. Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY

Abstract

The TypeNet deep learning models have rapidly become a strong candidate for biometric authentication using keystroke dynamics. The model is structured with two LSTM layers for each keystroke, a type of Recurrent Neural Network (RNN) that captures dependencies in whole sequences. Additionally, a Siamese Neural Network (SNN) is used to perform triplet loss, aiming to minimize distances between genuine samples and maximize distances between imposter samples. The model has demonstrated impressive performance, with Equal Error Rates (EER) ranging from 2.2% to 26.8%, depending on the keystroke dataset.

However, a major challenge remains: trust. Users are unlikely to adopt a model they don't fully understand. Deep learning models, in particular, suffer from a lack of transparency, making it difficult to explain how they weigh specific inputs. Unlike traditional machine learning models, deep learning models learn hierarchical representations directly from raw data, contributing to their non-linear behavior. This opacity makes it essential to develop methods that allow users to better understand the model's decisions.

Additionally, keystroke dynamics bring privacy concerns. By identifying less critical keystrokes, we can reduce sensitive information storage without significantly impacting the model's performance. To improve transparency and address privacy concerns, this study introduces two explainability methods: feature masking and embedding analysis.

Feature masking involves systematically zeroing out specific input features, enabling us to observe changes in the model's performance, especially in terms of EER. The results of feature masking revealed the critical importance of some features. For instance, masking the **Digraph ID** increased the EER from 10.47 to 34.8, and masking **Dwell Time** raised it to 37.5, indicating that these features are crucial for accurate predictions. In contrast, other features like **Flight Time Up-Down** and **Flight Time Up-Up** had less influence on the EER, suggesting they play a minor role in the model's decision-making process.

Embedding analysis offers insights into how the model encodes various features by comparing the 128-dimensional output embedding before and after masking, revealing which indices are influenced by specific features. Our findings showed that certain indices were consistently activated by particular features; for instance, indices 75, 111, and 112 were frequently influenced by **Digraph ID**, while indices 41 and 56 were more sensitive to Dwell Time. Notably, letter keys were found to be more important than control keys. These results suggest that the TypeNet model encodes different features in distinct areas of the embedding space, enhancing our understanding of how it differentiates between genuine and imposter samples.

Acknowledgment: This work was supported by the NSF Award #2244049: *"REU Site: High Performance Computing with Engineering Applications"* at Clarkson University.