Rochester Institute of Technology
REU
Human-Centered
Computational Sensing

Deep Multimodal Learning for Confusion Detection in Complex Collaborative Tasks

Sarah Smith, Tristan Chavez, Shweta Wahane
Advisors: Alex Ororbia, Reynold Bailey, Cecelia O. Alm Technical Assistance: Rajesh Titung

Introduction

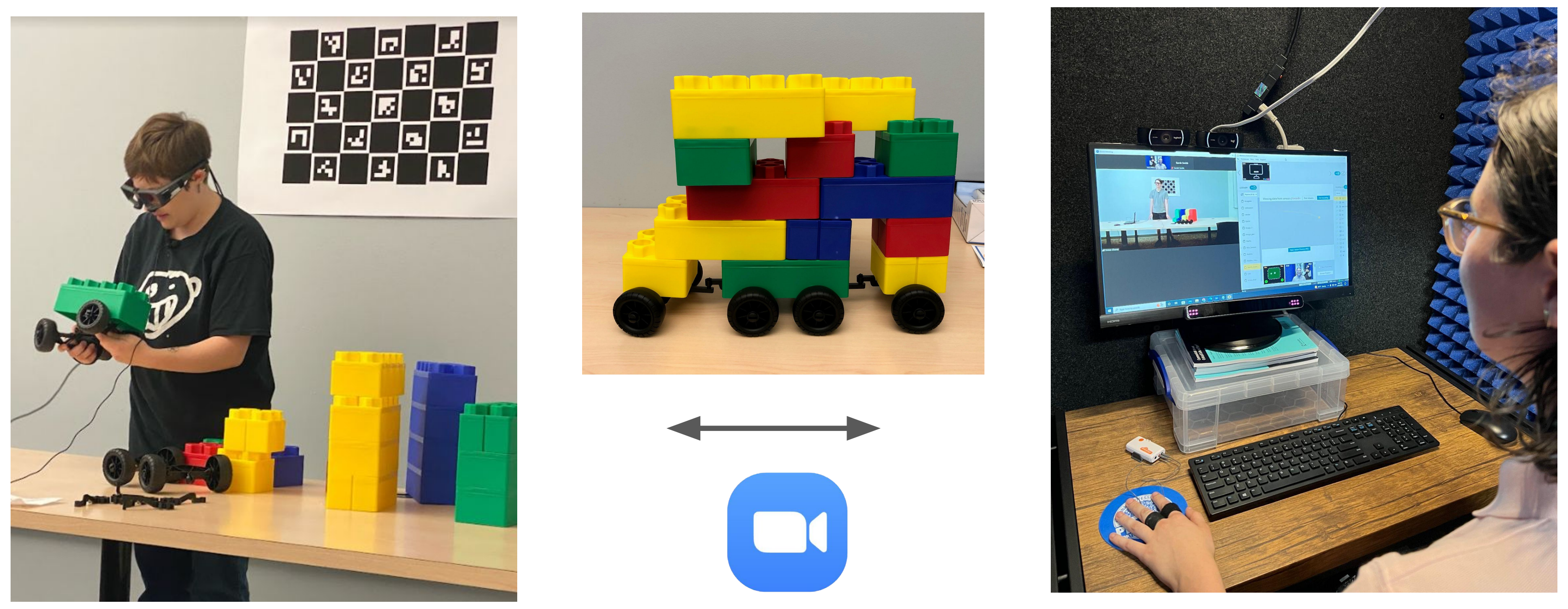
Confusion is a complex emotional affective state that can impact anyone in everyday life, but is especially common now with the rise of online learning and its lack of individualized attention. It is a worthwhile effort to create computer devices that can detect confusion in order to provide effective real-time interventions. Our study aims to create a machine learning model that can accurately detect and prediction confusion in the context of a complex collaborative task. We expand on previous research on the subject by devising an extensively multimodal Recurrent Neural Network (RNN) model that is trained to recognize confusion patterns.

RQ1: Which features from each type of data are most relevant to detecting confusion?

RQ2: Does a multimodal dataset perform better than a unimodal one in detecting confusion? How many and which modalities are ideal?

Methods

Participants worked together over Zoom to complete two building block tasks, with only the Instructor having a picture of the goal structure.

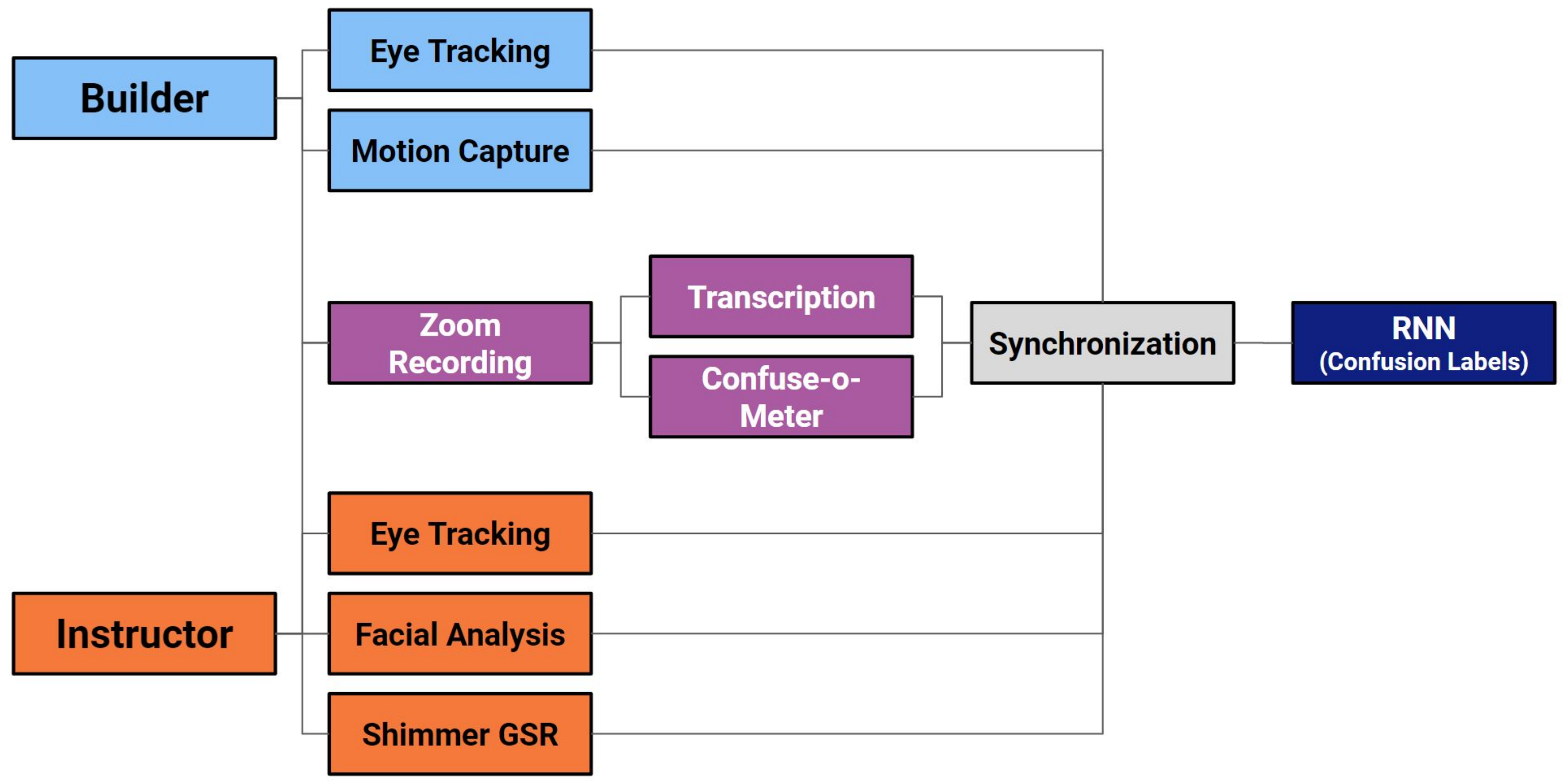


View of each participant's setup, along with the complex goal structure to be communicated.

Confusion was induced through the complexity of the second task's structure, as well as half of Builders being intentionally uncooperative with their partners. Participants then watched a recording and rated their confusion continuously over the course of the previous tasks on a 4-point Likert scale by using the Confuse-o-Meter review tool, which shows the group's recording and confusion scale radio buttons.

Data Sets

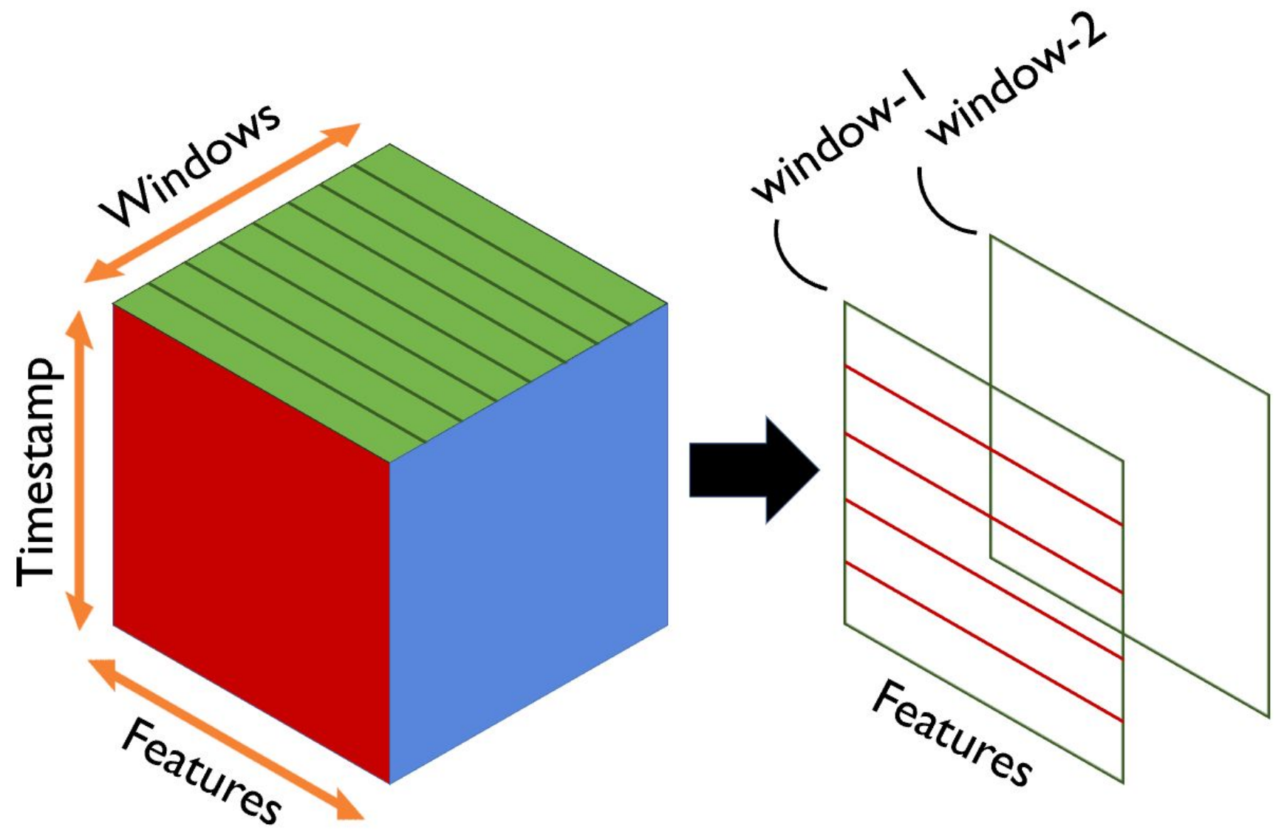
Data was acquired from 32 participants coming in groups of two, split equally into Instructor and Builder roles. The dataset was constructed from several modalities, some split by role and some with input from both participants. All modalities were synchronized after collection using one-second granularity. A participant-reported review task using the Confuse-o-Meter tool applied confused/not confused labels to the data.



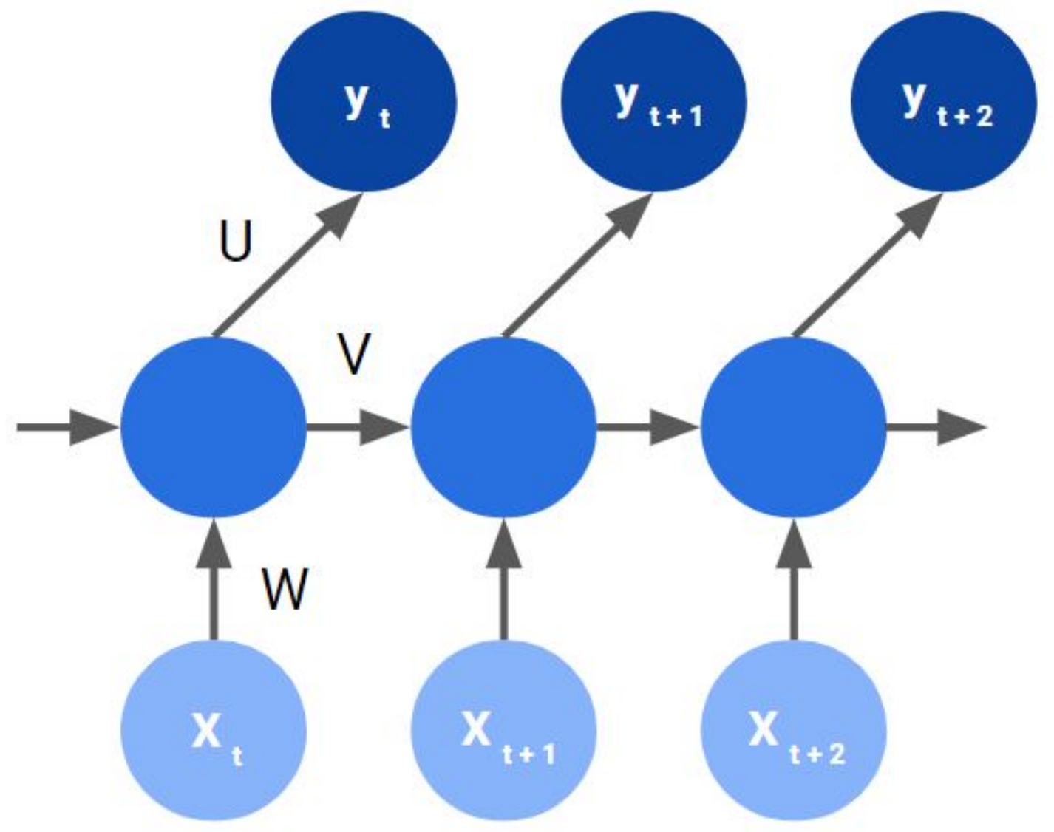
Data collection to processing workflow.

Neural Network

Windows are randomly sampled along with corresponding labels to be broken into training and testing sets. Labels are predicted using the softmax activation function and categorical cross entropy compares output with actual labels to calculate gradients. An Adam optimizer function applies gradients to weights.



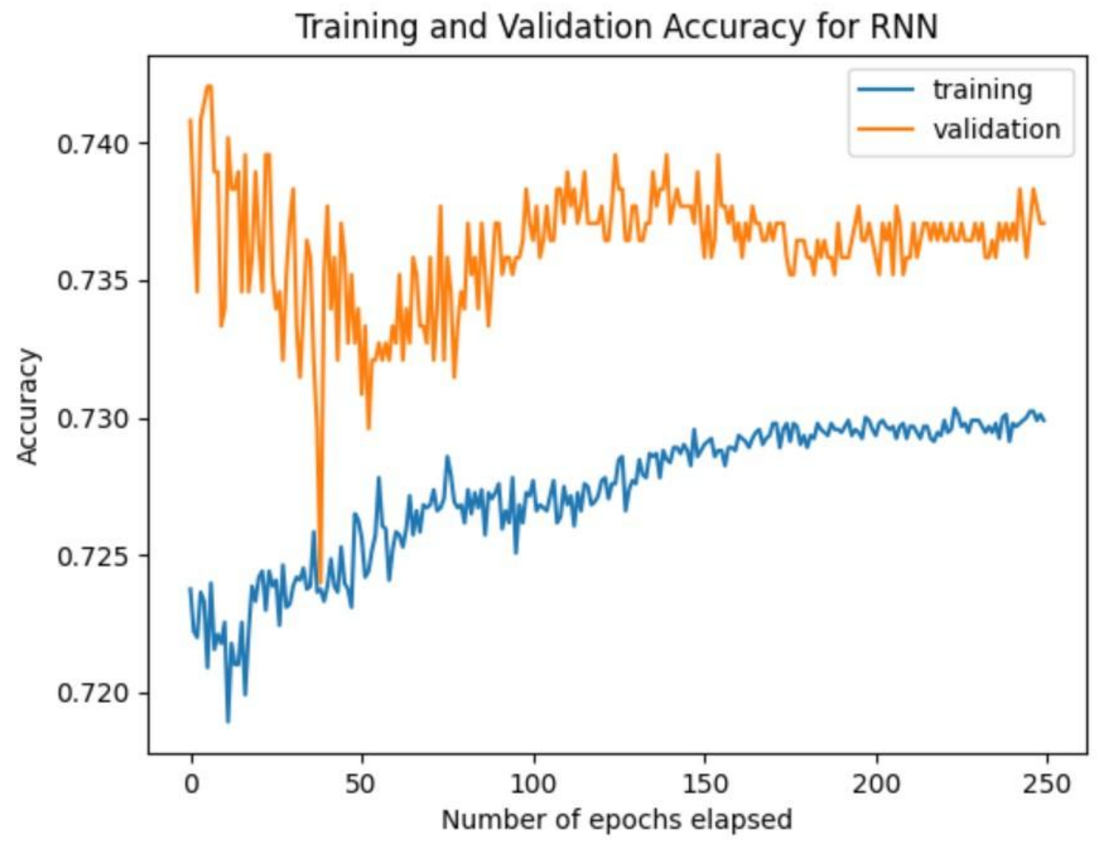
Tensor structure for data input and windowing..



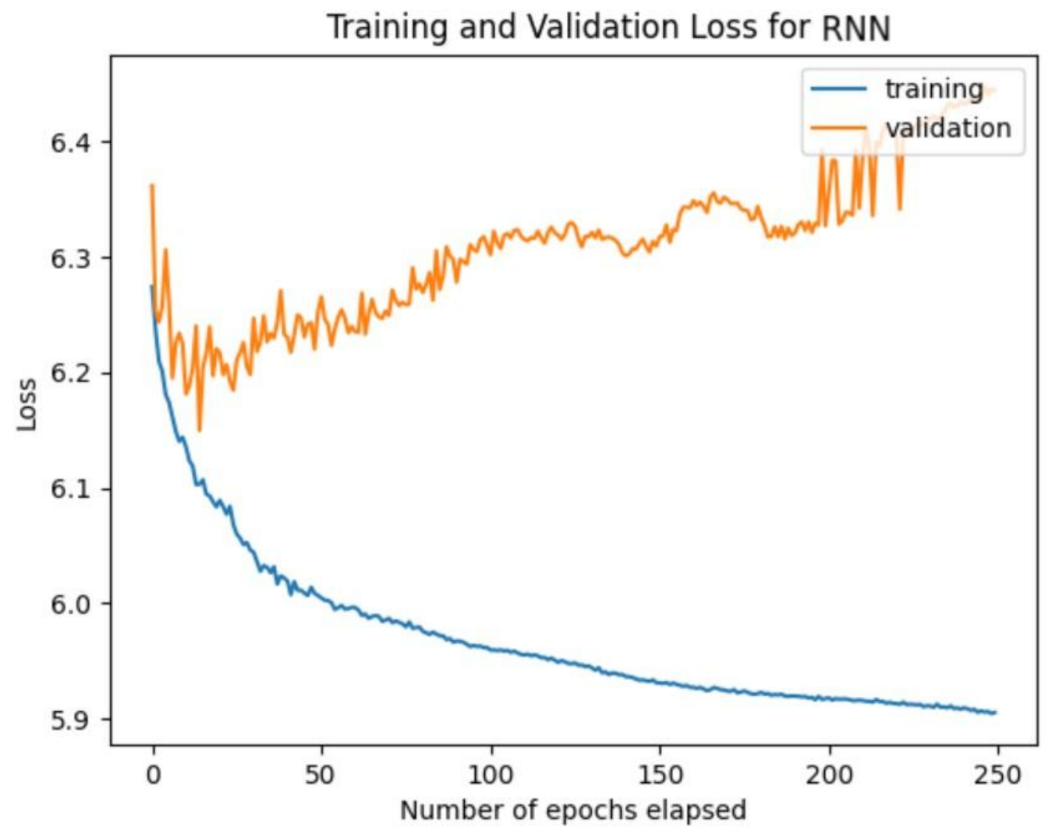
RNN structure. From bottom to top: Data Input, Hidden Nodes, Output at time *t*.

Results

With a window size of 5, learning rate of .05, 5 hidden layers, and 250 epochs, we found that accuracy did not improve as the model was trained and the loss increased, which means the model is not generalizing over time.

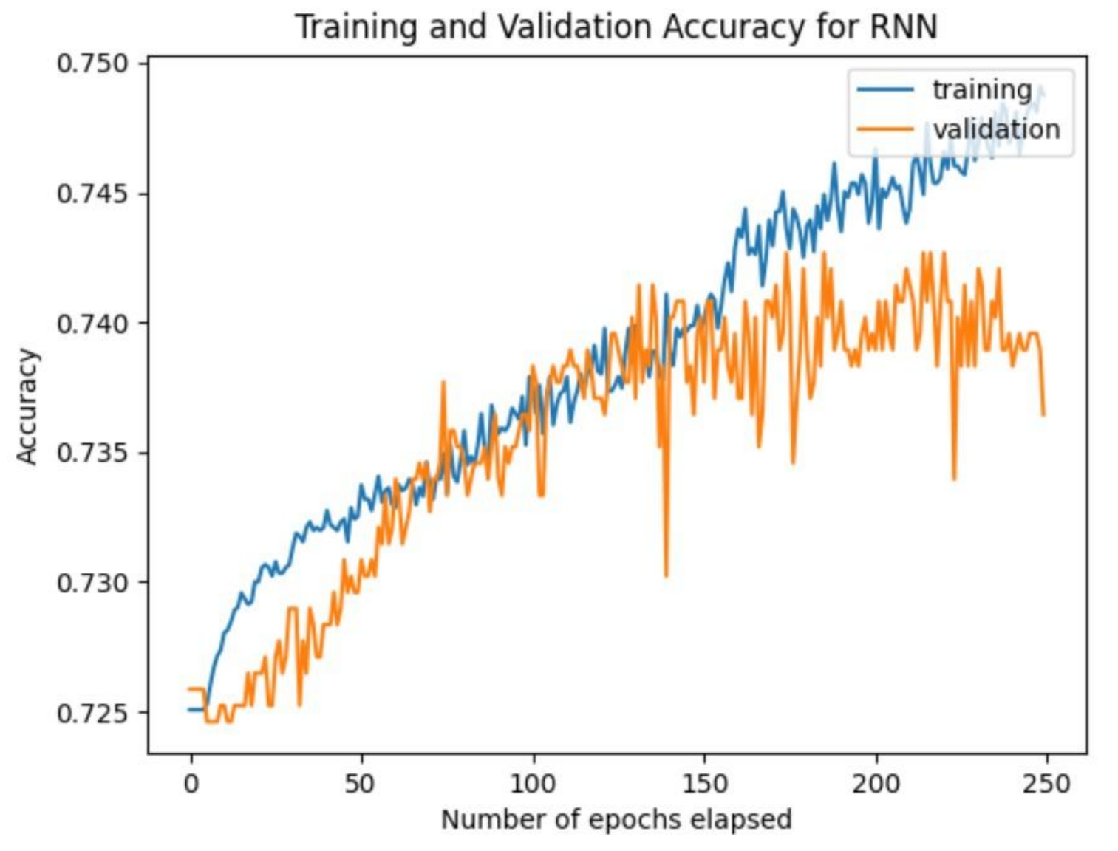


Accuracy over time - Experiment 1.

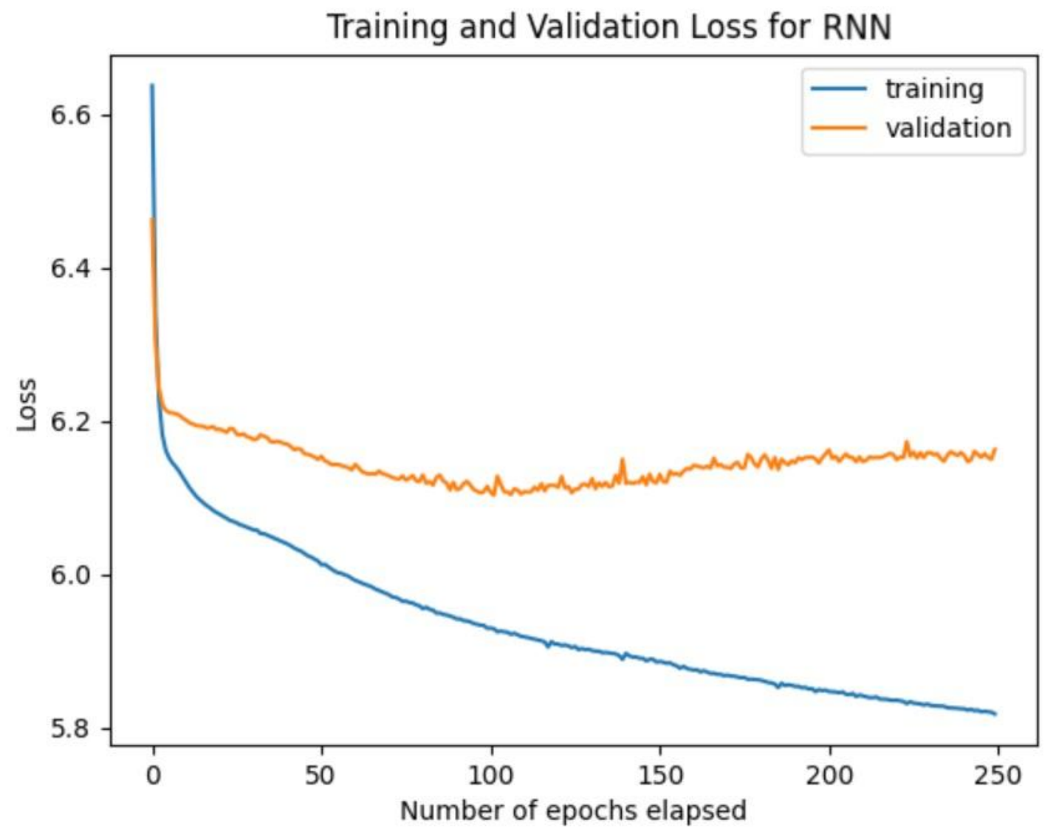


Loss over time - Experiment 1.

Accuracy over time did show improvement when the learning rate was changed to .001 and hidden layers to 20, and the loss was more stable, rather than increasing.



Accuracy over time - Experiment 2.



Loss over time - Experiment 2.

Future Work

We are in the process of optimization and improvement of our RNN accuracy through systematic adjustment of hyperparameters. We also intend to include the transcript data to the model, which was withheld due to time constraints.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Award No. IIS-1851591. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.