# GestureGenius: Bringing Classroom Interactions to Life

Austin Kwan
Jessica Luong
Chanson Tang
aka120@sfu.ca
jla715@sfu.ca
cta111@sfu.ca
Simon Fraser University
Burnaby, BC V5A 1S6, Canada

## ABSTRACT

In a classroom setting, fostering an optimal learning environment relies not only on effective teaching methodologies, but also on the ability to understand and respond to the dynamic behaviours of students in real time. Much of the current works on classroom behaviour are based around recognition, but do not provide interaction and response. This paper proposes a system for real-time classroom behaviour recognition and interaction, leveraging advancements in machine learning and computer vision technologies. We achieve this through extraction of human landmarks through live feed, classifying a classroom behaviour with said landmarks, and providing a suitable response to the user. This system has promising results and can be refined for reduced latency and better responses. By integrating these technologies with traditional classroom settings, educators can possibly reduce their workload and also gain valuable insights into student engagement.

## 1 INTRODUCTION

An optimal learning environment is a combination of effective teaching methodologies and effective, efficient responses to students. Taking inspiration from current projects on systems that already do classroom behaviour recognition. We wanted to expand on top of this by providing interaction alongside it. Existing work related to classroom behaviour recognition include [1], [2], [3]. Currently, there is little work on using the behaviour recognition to provide interaction to users.

### 1.1 Literature Review

Lin et al. [1] introduce a system for recognizing student behaviors in classrooms, addressing the limitation of single-behavior focus in similar studies. Their method employs consecutive frames from a classroom camera, OpenPose framework for skeleton data collection, and an error correction scheme to refine the data. Feature extraction is utilized to generate feature vectors representing human postures, followed by behavior classification using a deep neural network (DNN) model, resulting in improved performance compared to analogous methods.

Zhang et al. [2] focus on hand-raising as a common behavior in classrooms. By improving feature expression and utilizing techniques like Spatial Context Augmentation and Multi-Branch Dilated Convolution, they enhance the accuracy of hand-raising detection. Extensive experiments demonstrate the effectiveness of their approach, showing good performance compared to similar methods across different datasets, affirming the algorithm's versatility.

Wang, Jiang and Shen [3] introduce a method for detecting students' yawning behavior in classroom settings, aiming to analyze their learning states and engagement. Challenges such as occlusion, low-resolution faces, and interference from similar behaviors like talking are addressed by building a yawn gesture dataset and implementing an improved R-FCN coupled with a mouth fitting technique. This method achieves a high mean average precision (0.90 mAP) in classroom scenarios, thus meeting real-world application requirements.

Despite these works being adept at recognition for classroom behaviour, they do not fit with our goal of providing users with interaction.

### 1.2 Motivation and Goal

Our goal is to provide a system that can recognize classroom behaviours and provide an appropriate response to the identified behaviour. The main issues that prevent current approaches from reaching our goal are the lack of interest in interaction. Furthermore, existing datasets are mostly image-based and not video-based, which is what we are training on. Having a video-based dataset allows for more natural and accurate recognition of common classroom behaviours. The lack of interest in interaction is mainly due to the focus on recognition, which indirectly affects the datasets related to this topic. What we are contributing is an interactive application that recognizes common classroom social signals and a virtual agent that responds to said signals in real-time. This system will help educators with managing their students in an online classroom setting by providing a sense of entertainment and interaction in real-time.

## 2 APPROACH

### 2.1 System Design and Architecture

Our project integrates a backend powered by Python and a frontend developed using Unity. This integration allows for the efficient processing of social signals captured through a webcam and a response displayed by a virtual agent. The communication between Python and Unity is handled by TCP for its reliability, and separate threads are used to ensure real-time responsiveness.

*2.1.1 Python Components.* In the backend, Python handles recognition of social signals. It does so by first capturing video directly through a webcam. This data is then processed through MediaPipe for feature extraction and LSTM networks for classifying these features into the predefined social signals. Python additionally acts as a TCP client, managing the communication by sending processed signals to the Unity application.

*2.1.2 Unity Components.* On the frontend, Unity provides interaction to the user. It operates as the TCP server, receiving classified signals from the Python backend. It then displays the virtual agent, which responds to the signals with appropriate animations and eye states, providing an interactive user experience. For example, if the signal received from Python is 'clap', the agent responds with clapping and 'happy' eyes as shown in Figure 1.

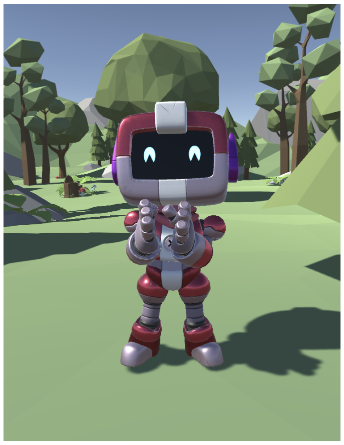


**Figure 1: Virtual agent response to clapping by user.**

## 2.2 Dataset

Our dataset was carefully constructed using live-streamed video recordings of our team members, tailored specifically to the project's requirements. Each sequence consisted of 30 frames featuring a single social signal with only one person visible. We incorporated variations within each gesture to enhance the dataset's robustness. For example, the gesture 'thumbs up' was captured with variations such as the left hand only, right hand only, and both hands. The dataset of all our actions comprises 1620 instances (810 seconds at 60 frames/second) for training and 20 videos of varying length for testing.

*2.2.1 Training Set.* The training set comprises 540 sequences per team member. The breakdown of sequences for each social signal category, with the specific signal in parentheses, is as follows: Attention (raising hand) - 60 sequences, Approval (nodding) - 90 sequences, Disapproval (shaking head) - 90 sequences, Celebratory (cheering) - 60 sequences, Questioning (crossing arms) - 120 sequences, Appreciation (clapping) - 90 sequences, and Neutral (sitting still) - 30 sequences. This distribution is shown in Figure 1.

Some signals have more sequences than other due to the different variations in how people can perform each signal. We picked these categories to cover the range of nonverbal classroom communications one may express. Additionally, since the data is collected and performed by our team in controlled conditions, minimal preprocessing was required.

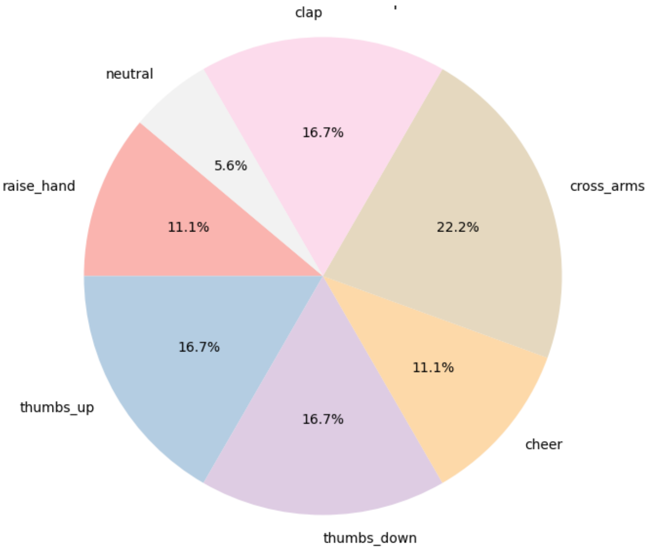


**Figure 2: Distribution of social signals in training data.**

During data collection, we capture 30 frames for each sequence, with each sequence representing a specific social signal. This structured collection ensures that each signal is accurately labelled and corresponds directly to the actions being performed.

*2.2.2 Test Set.* Our test set comprises publicly available YouTube clips, selected to evaluate the model's generalization capabilities with individuals not part of our team. The test set includes three sequences for each signal category, except for 'neutral', which has only two. Our team manually annotated these sequences, effectively creating a set of ground truth labels for testing.

## 2.3 Feature Extraction

We extracted landmarks from frames using the MediaPipe Holistic model. The features extracted include:

- 33 pose landmarks, each with coordinates (x, y, z) and a visibility score. These landmarks capture the overall body movements and are illustrated in Figure 3.
- 21 hand landmarks for each hand, also with coordinates (x, y, z), which add small gestures and hand movements to the data set, as shown in Figure 4.

This results in a total of 258 coordinates per frame, which are then saved in NumPy arrays and stored in .npy files. These files serve as the data input for training our model to classify various social signals.

## 2.4 Model Choice

Our model architecture is shown in Figure 5. It uses three long short-term memory (LSTM) layers stacked together to analyze on temporal data. Each layer potentially captures different aspects of temporal dependencies in the data, which is crucial for recognizing the sequential nature of human gestures. Next, we use dense layers with the ReLU activation function to introduces non-linearity into the model, allowing it to learn more complex patterns in the data.

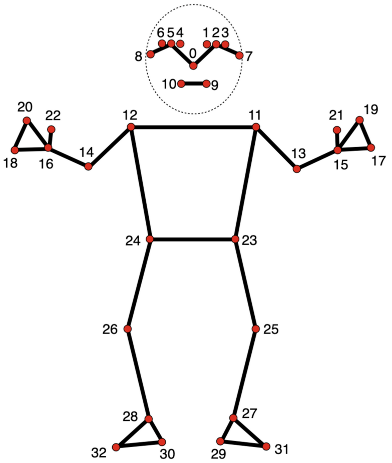


**Figure 3: Body landmark locations, [4].**

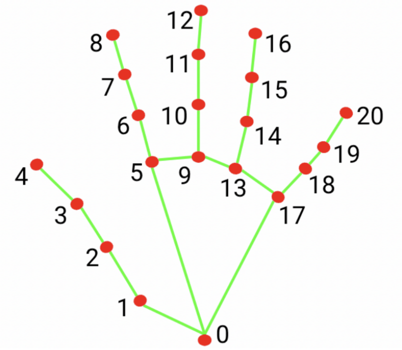


**Figure 4: Hand landmark locations, [5].**

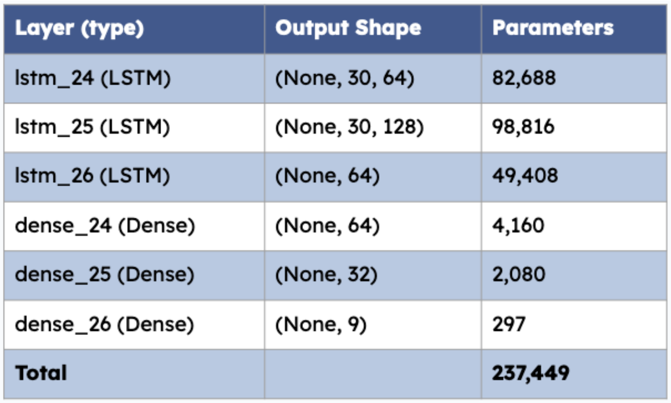
The final output layer uses a softmax activation for classification. As for the libraries and frameworks used for our model, we used the Keras and TensorFlow libraries for our model. We chose them for their efficiency and our familiarity with them from previous projects and assignments.

# 3 EXPERIMENTS AND RESULTS

## 3.1 Recognition Evaluation

*3.1.1 First Approach.* The data for our first approach consisted of approximately 1200 data points. Each data point is annotated on their visual landmarks by their x, y, and z coordinate from an image. This data is then stored into a comma-separated values file and read in a NumPy array. The array is then preprocessed through normalization and label encoding. The data was split into training and test points with a split of 80/20 respectively. A sequential model which included a 1D convolution layer and multiple dense layers was then used for training and testing over 100 epochs. The model's accuracy on the test set was 46%.

We believe that the model performed poorly due to how the data was collected and how the model was trained. As the model used the image's coordinates rather than a sequence of data, the model relied too heavily on the user's location.



| Layer (type) | Output Shape | Parameters |
|---|---|---|
| lstm_24 (LSTM) | (None, 30, 64) | 82,688 |
| lstm_25 (LSTM) | (None, 30, 128) | 98,816 |
| lstm_26 (LSTM) | (None, 64) | 49,408 |
| dense_24 (Dense) | (None, 64) | 4,160 |
| dense_25 (Dense) | (None, 32) | 2,080 |
| dense_26 (Dense) | (None, 9) | 297 |
| **Total** | | **237,449** |

**Figure 5: Overview of model architecture.**

*3.1.2 Second Approach.* The second model uses the architecture described in the Approach section, which includes three stacked LSTM layers followed by dense layers. This configuration was chosen to capture sequential data, which is inherent in human movement patterns. However, the model was trained on a limited dataset consisting of only 30 instances for each social signal, sourced from a single team member with minimal variation within each category. This lack of diversity in the training data likely contributed to overfitting, as the model was not exposed to a broad range of examples.

We trained our second model over 100 epochs. We performed 5-fold-cross-validation and the accuracy for each of the folds is: 70.45%, 95.35%, 90.70%, 86.05%, and 86.05%. The average accuracy from the cross-validation was approximately 85.72%, indicating a strong model performance during training.

However, when evaluated on the independent test set, the model's accuracy dropped significantly to 35%. This suggests potential overfitting to the training data. The confusion matrix in Figure 6 shows that the model is unable to label 'thumbs up', 'thumbs down' or 'cheers' correctly. The model is able to identify all three instances of 'cross arms'. One instance of misclassification is the model labels a 'raise hand' sequence as 'cheer'. This may be due to how both hands are lifted to similar positions. We tried to address similar issues in the third approach by including more variations in the training dataset for each social signal.

*3.1.3 Third Approach.* Our third and final model uses the same structure as the second approach, but used two methods to address overfitting. The first is early stopping to monitor the performance of our model in order to find the optimum number of epochs. Next, we expanded our dataset to include all three team members, and added variations for each social signal. For example, our dataset for the second model used only 30 sequences of 'cheer', but the third model uses 60 instances, which include cheering with hands above the head and cheering with hands below the head.

This model was trained over 7 epochs with a batch size of 32. We performed 5-fold-cross-validation and the accuracy for each of the

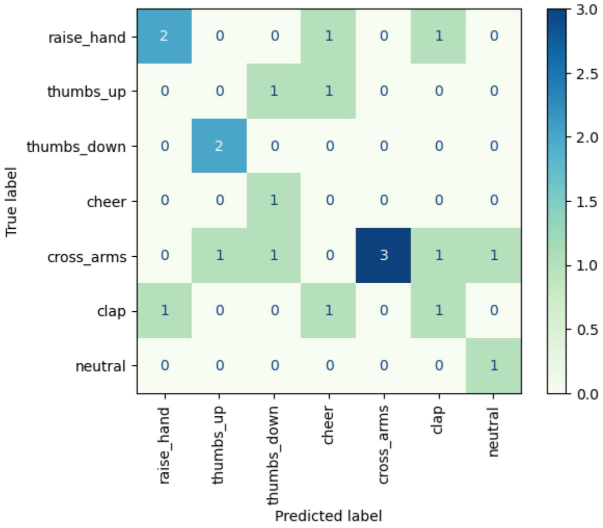


**Figure 6: Confusion matrix for second model evaluated on the test set.**

folds is: 77.31%, 90.74%, 97.22%, 94.91%, and 98.15%. The average accuracy from the cross-validation was approximately 91.67%, which is an improvement of 5.95% compared to the second approach.

Evaluating on the independent test set yielded an accuracy score of 55%, which also shows an improvement over the previous approach. Figure 7 shows that the model is still able to identify all instances of 'cross arms'. The model no longer mislabels 'raise hand' as 'cheer', but there is one case where it mislabels 'cheer' as 'raise hand'.

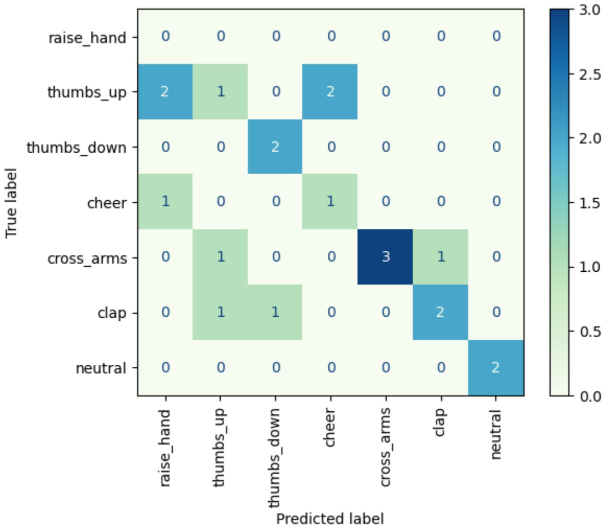


**Figure 7: Confusion matrix for the third model evaluated on the test set.**

## 3.2 Interaction Evaluation

For the interaction portion, the latency to recognize and perform the response was an acceptable length. It did not take an extended amount of time, but it can be improved for lower latency and higher accuracy.

As for user feedback, we tested on friends and family in a casual setting to gain their thoughts. Feedback was based on whether the agent's response was appropriate for the corresponding signal and usability of the interactive agent. Generally, the feedback was positive for the agent's response. At our current iteration, it is difficult to say our system is ready for practical use, but the feedback was that further development would help the usability of the system for our goal and proposed setting.

# 4 DISCUSSION

Examples of situations where our system did not perform as well are certain actions being confused with another one, recognition involving multiple people, and tracking social signals while switching focus between different people. When our system runs the test dataset, it will at times confuse actions like clapping and cross arms with each other. Other errors include misclassification of two 'thumbs up' as 'cheers'. We included examples of two 'thumbs up' in our dataset, but the model is still unable to identify this signal. Additionally, the model frequently misinterpreted hands held together as 'clapping'.

Possible routes for future work include more signal recognition, more responses, and further development to not just provide interaction but also useful insight. Our dataset could be more robust with more data and variations in the actions; this can be done with more subjects to produce more data. This project can be further worked on to include signals related to student engagement like boredom and slacking off; this can help give insight to the instructor on how to improve their delivery of material to keep the class engaged.

This project can be also be developed to provide useful insight to the users. This could be inquiries about assignment due dates, assignment details, and more engaging responses like asking why the user may be feeling a certain way or performing a certain action. In a sense, it would be similar to ChatGPT, but more geared towards reducing the workload of the instructor while efficiently providing insightful responses, so the users do not have to wait for extended periods of time.

# 5 CONCLUSION

Overall, our project has promising results in its capability to recognize classroom behaviour and interact with the users. We are hopeful that our system could be of use in an online classroom setting, where it may be difficult for educators to effectively help or tend to each student. This system shows the potential of transforming the traditional classroom setting to one that is more dynamic and interactive, that can also provide insight for improving student engagement and reducing the workload of educators.

# REFERENCES

[1] F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. Le, "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, vol. 21, no. 16, p. 5314, Aug. 2021.

[2] L. W. G. Zhang, L. Wang and Z. Chen, “Hand-raising gesture detection in classroom with spatial context augmentation and dilated convolution,” *Computers & Graphics*, vol. 110, pp. 151–161, Feb. 2023.

[3] F. J. Z. Wang and R. Shen, “An effective yawn behavior detection method in classroom,” *Neural Information Processing*, pp. 430–441, 2019.

[4] Google for Developers, “Pose Landmark Detection Guide,” https://developers.google.com/mediapipe/solutions/vision/pose_landmarker, [Accessed: 16-Apr-2024].

[5] ——, “Gesture recognition task guide,” https://developers.google.com/mediapipe/solutions/vision/gesture_recognizer, [Accessed: 16-Apr-2024].

# A DATASHEETS FOR DATASETS

- For what purpose was the dataset created?
  - The dataset was created for our application to show its viability.
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  - Our group created this dataset on no one’s behalf.
- Who funded the creation of the dataset?
  - There was no funding involved in the creation of the dataset.
- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
  - The instances are in mostly .npy file format and a small portion of video in mp4 format.
- How many instances are there in total (of each type, if appropriate)?
  - 1620 NumPy files and 20 mp4 files.
- Does the dataset contain all instances, stances or is it a sample (not necessarily random) of instances from a larger set?
  - The dataset does not contain all possible instances, it can be improved with more data.
- What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features?
  - The NumPy files contains the x and y coordinates of the recorded landmarks in the live feed.
- Is there a label or target associated with each instance?
  - There are labels associated with all instances.
- Is any information missing from individual instances?
  - Face and lower body landmarks are missing, as they are not as relevant in our use case.
- Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?
  - No.
- Are there recommended data splits (e.g., training, development/validation, testing)?
  - Yes.
- Are there any errors, sources of noise, or redundancies in the dataset?
  - Yes, there are possible errors, sources of noise, or redundancies. Our data may overlap with one another and may not be as accurate as we think it is.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  - Yes, the dataset is self-contained.
- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)?
  - No.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  - No.

- Does the dataset identify any subpopulations (e.g., by age, gender)?
  - No.
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
  - Yes and no. The mp4 files are stock videos of individuals doing a certain action.
- Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometrics or genetic data; forms of government identification, such as social security numbers; criminal history)?
  - No.
- How was the data associated with each instance acquired?
  - The data were directly reported by the subjects, which is us, the authors.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
  - Using webcam and Mediapipe Pose Landmarker.
- If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
  - The dataset is not a sample from a larger set.
- Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?
  - The authors were the only ones involved in the data collection process, and we were not compensated for it.
- Over what timeframe was the data collected?
  - The timeframe in the data was collected is the same as the creation of the project.
- Were any ethical review processes conducted (e.g., by an institutional review board)?
  - No.
- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
- Were the individuals in question notified about the data collection?
- Did the individuals in question consent to the collection and use of their data?
  - No.
- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
  - No, as the individuals were ourselves.
- Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
  - No.
- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - Yes, there was labeling of the data.
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  - No.
- Is the software that was used to preprocess/clean/label the data available?
  - No.
- Has the dataset been used for any tasks already?
  - The dataset has been only used for our project.
- Is there a repository that links to any or all papers or systems that use the dataset?
  - Yes.
- What (other) tasks could the dataset be used for?
  - We have not thought of alternative tasks that the dataset could be used for.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  - No.
- Are there tasks for which the dataset should not be used?
  - Tasks that will personally benefit the teacher/instructor or developers. Tasks that can manipulate or control the users into thinking or acting in an unacceptable manner.
- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  - No, we currently do not have plans to distribute the dataset.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
  - If we do distribute it, it will most likely be through GitHub.
- When will the dataset be distributed?
  - Unknown at this moment.
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  - Unknown at this moment.
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  - Unknown at this moment.
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
  - Unknown at this moment.
- Who will be supporting/hosting/maintaining the dataset?
  - It will most likely be us, the authors.
- How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
  - We can be contacted at our respective emails at the beginning of this report.
- Is there an erratum?
  - No.
- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  - Maybe.
- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g.,

were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?
  - No.
- Will older versions of the dataset continue to be supported/hosted/maintained?
  - No.
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  - No mechanism to do so at this moment.

## B CONTRIBUTIONS

- Austin Kwan
  - First approach (research)
  - Data collection for test set
  - Report
  - Poster
- Jessica Luong
  - Unity and Python integration
  - Report
  - Poster
- Chanson Tang
  - First approach (code)
  - Second approach (code & research)
  - Third approach (code & research)
  - Report
  - Poster