Early Prediction of Children's Task Completion in a Tablet Tutor using Visual Features (Student Abstract)

Bikram Boote [1]*, Mansi Agarwal[2]*, Jack Mostow[3]
[1] Jadavpur University, India
[2]Delhi Technological University, India
[3]Carnegie Mellon University, USA
bikram.boote@gmail.com, mansiagarwal_bt2k16@dtu.ac.in, mostow@cs.cmu.edu

Abstract

Intelligent tutoring systems could benefit from human teachers' ability to monitor students' affective states by watching them and thereby detecting early warning signs of disengagement in time to prevent it. Toward that goal, this paper describes a method that uses input from a tablet tutor's user-facing camera to predict whether the student will complete the current activity or disengage from it. Training a disengagement predictor is useful not only in itself but also in identifying visual indicators of negative affective states even when they don't lead to non-completion of the task. Unlike prior work that relied on tutor-specific features, the method relies solely on visual features and so could potentially apply to other tutors. We present a deep learning method to make such predictions based on a Long Short Term Memory (LSTM) model that uses a target replication loss function. We train and test the model on screen capture videos of children in Tanzania using a tablet tutor to learn basic Swahili literacy and numeracy. We achieve balanced-class-size prediction accuracy of 73.3% when 40% of the activity is still left.

Introduction

Analyzing the dynamics of students' affective states over the course of a learning process is important in order to create a more engaging environment. Human teachers can monitor students' facial expressions, behavior, and performance to detect disengagement and address it. Ideally, intelligent tutors should likewise detect and respond to early warning signs of disengagement (Dewan, Murshed, and Lin 2019).

This paper describes a method to monitor affective state, and evaluates its ability to predict disengagement. We investigate this problem in the context of RoboTutor (McReynolds et al. 2020), a tablet tutor that teaches basic Swahili literacy and numeracy through a series of educational activities. We train and test a method that analyzes screen capture videos that include tablet-camera input and predicts whether the child will complete the current activity.

Dataset Description

The data for this study consists of screen-capture recordings of 200 RoboTutor sessions of children aged 6-12. Each video has temporal resolution of 48 frames per second and spatial resolution of 1024×720 pixels. The 192×136 pixel window at the top right of the screen displays the input from the tablet's user-facing camera, including the user's face.

We segmented each video into one clip per activity and labeled its outcome as *Complete* if the child completed the activity or *Bailout* if the child tapped on the Back button to stop the activity. The 200 videos yielded 1195 clips, 803 labeled *Complete* and 392 labeled *Bailout*.

Methodology

We now describe our automated process for segmenting and labeling the screen videos, the features we extracted from them to train on, and the model we trained.

Automatic Segmentation and Labeling: To segment a video into activities, we needed to know where each activity started and ended. Unfortunately, the videos were from a version of RoboTutor that did not yet log this information. Instead, we inferred activity boundaries from the videos themselves by detecting the selector screen displayed before an activity and the rating screen displayed afterwards, as follows. To detect if a video frame shows one of these screens, compare it to a reference image of the screen type and decide if fewer than 5% of their pixels differ. This threshold is low enough to detect a screen type accurately but flexible enough to tolerate normal variations in its appearance, such as which item in a menu is highlighted.

The screen videos show taps as small white dots. To detect taps on the Back button, we used OpenCV's HoughCircles (Yuen et al. 1990) method to look for a white circle in the 64×36 pixel area at the top left of the screen where the Back button is located. Our automated labeller segmented the videos into clips, each starting at an activity selector screen and ending either at the activity rating screen (labeled *Complete*) or tap on the Back button (labeled *Bailout*).

It is important to emphasize that we used this information about RoboTutor only to segment and label the videos, which could be done instead from timestamped log entries if available. The subsequent process of training and testing a disengagement prediction model on the labeled data did not use any tutor-specific information.

Feature Engineering: We used visual features computed by OpenFace (Baltrušaitis, Robinson, and Morency 2016), a facial behavior analysis tool. We used the same set of static

*Both student authors contributed equally. We thank the team that developed RoboTutor and the children who used it.

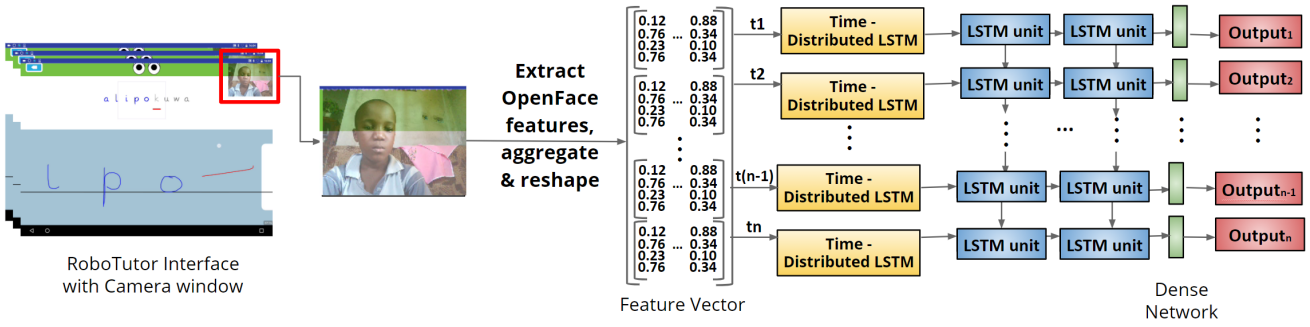

Figure 1: Model Architecture

features as (Agarwal and Mostow 2020), namely head proximity, pitch, yaw, roll, eye gaze, blink, pupil dilation, and Facial Action Units. To smoothe noisy measurements, we averaged static features over 4 frames, and normalized each feature to the interval $[0, 1]$.

Model Architecture and Training: To predict whether a child will complete an activity, we train a Long Short Term Memory (LSTM) based deep learning model using the target replication technique (Lipton et al. 2015) to reward early prediction. The model consists of a Time Distributed LSTM layer with 4 units followed by 2 regular LSTM layers each of 64 units and finally a dense (fully connected) network. The dense network has 3 layers, with 200 neurons in the first layer, 50 neurons in the second layer, and a single output neuron in the final layer to represent a probabilistic binary output. Fig. 1 shows the model architecture.

The target replication objective function is:

$$\alpha.\frac{1}{T}\sum_{t=1}^{T} loss(y'(t), y(t))) + (1-\alpha).loss(y'(T), y(T)) \quad (1)$$

Here T represents the total number of timesteps and α is a hyperparameter in the range $[0, 1]$ to weight the relative importance of errors on prediction at the last time step T versus at the individual time steps t. Our loss function is binary cross-entropy given by:

$$-(y(i).log(y'(i)) + (1 - y(i)).log(1 - y'(i))), \quad (2)$$

where $y(i)$ is the true label (*Complete* or *Bailout*) and $y'(i)$ is the predicted probability of completing the activity.

We use Stochastic Gradient Descent optimization with a learning rate of 0.001 to train the model for 500 epochs. The clips vary in length, so each epoch updates the gradient separately for one clip at a time, rather than zero-padding shorter clips to a common clip length so as to update gradients in batches of clips.

Results

We tested our model on a random test set of 90 clips with 47 *Complete* instances and 43 *Bailout* instances, verifying that the children in the training and test sets did not overlap. With 40% of each clip left, the model had accuracy 73.3% at α = 0.6. Table 1 reports the variation in accuracy with different values of alpha (α).

alpha (α)	0.1	0.2	0.3	0.4	0.5	**0.6**	0.7	0.8	0.9
Accuracy (%)	52.2	61.1	66.7	65.6	71.1	**73.3**	66.7	67.8	68.9

Table 1: α = 0.6 got the top accuracy of the values tried.

Conclusion

This paper presents a deep learning model to predict task completion as an indicator of disengagement in children using a tablet tutor. We segmented and labeled data automatically using pixel difference and dot detection. Our model only uses visual cues extracted from the tablet's user-facing camera input, so it could potentially be generalized to other tutors more easily than methods that rely on tutor-specific features. This work contributes to the automated identification of early visual harbingers of disengagement. It should help improve tutors at design time and eventually guide pedagogical decisions in real time.

References

Agarwal, M.; and Mostow, J. 2020. Semi-Supervised Learning to Perceive Children's Affective States in a Tablet Tutor. In *AAAI*, 13350–13357.

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. IEEE.

Dewan, M. A. A.; Murshed, M.; and Lin, F. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6(1): 1.

Lipton, Z. C.; Kale, D. C.; Elkan, C.; and Wetzel, R. 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* .

McReynolds, A. A.; Naderzad, S. P.; Goswami, M.; and Mostow, J. 2020. Toward Learning at Scale in Developing Countries: Lessons from the Global Learning XPRIZE Field Study. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 175–183.

Yuen, H.; Princen, J.; Illingworth, J.; and Kittler, J. 1990. Comparative study of Hough transform methods for circle finding. *Image and vision computing* 8(1): 71–77.