Multimodal Multi-Task Stealth Assessment for Reflection-Enriched Game-Based Learning

Anisha Gupta¹, Dan Carpenter¹, Wookhee Min¹, Jonathan Rowe¹, Roger Azevedo², and James Lester¹

¹ North Carolina State University, Raleigh, NC 27695, USA {agupta44, dcarpen2, wmin, jprowe, lester}@ncsu.edu ² University of Central Florida, Orlando, FL 32816, USA roger.azevedo@ucf.edu

Abstract. Game-based learning environments enable effective and engaging learning experiences that can be dynamically tailored to students. There is growing interest in the role of reflection in supporting student learning in game-based learning environments. By prompting students to periodically stop and reflect on their learning processes, it is possible to gain insight into students' perceptions of their knowledge and problem-solving progress, which can in turn inform adaptive scaffolding to improve student learning outcomes. Given the positive relationship between student reflection and learning, we investigate the benefits of jointly modeling post-test score and reflection depth using a multimodal, multitask stealth assessment framework. Specifically, we present a gated recurrent unit-based multi-task stealth assessment framework that takes as input multimodal data streams (e.g., game trace logs, pre-test data, natural language responses to in-game reflection prompts) to jointly predict post-test scores and written reflection depth scores. Evaluation results demonstrate that the multimodal multi-task model outperforms single-task neural models that utilize subsets of the modalities, as well as non-neural baselines such as random forest regressors. Our multi-task stealth assessment framework for measuring students' content knowledge and reflection depth during game-based learning shows significant promise for supporting student learning and improved reflection.

Keywords: Multimodal learning analytics, multi-task learning, natural language processing, reflection, game-based learning

1 Introduction

Recent years have witnessed growing interest in game-based learning environments due to their potential to foster student learning and create personalized learning experiences [1-3]. By embedding story-based problems within interactive virtual environments, game-based learning environments can promote student interest and provide engaging opportunities for problem solving. An important factor in game-based learning environments is students' ability to self-regulate their learning processes [4, 5]. Game-

based learning environments with robust student performance assessment capabilities show significant promise for identifying gaps in student knowledge and informing adaptive scaffolding designed to support self-regulated learning and problem solving.

A key challenge is how to reliably model student knowledge without disrupting engagement during learning activities. Stealth assessment addresses this challenge by analyzing fine-grained sequences of game interaction data to automatically infer students' knowledge and learning competencies without disrupting students' game-based learning experiences [6, 7]. Following an evidence-centered assessment approach [8], stealth assessment can be formulated as a regression task to predict students' post-test scores based on the observed sequences of students' problem-solving interactions during game-based learning. Accurate predictions of post-test scores could inform instructors about students in need of help. The predictions can also be utilized in-game to provide scaffolds to dynamically remediate gaps in knowledge and support problem solving.

While in-game action sequences have been investigated as a traditional source of predictive features for stealth assessment [7], students' written reflections also hold significant potential to improve prediction of student competencies when combined with in-game action sequences [5]. Moreover, supporting effective reflection is itself an important goal in game-based learning due to its central role in self-regulated learning. To identify whether students are reflecting effectively, reflections are often evaluated in terms of their depth (e.g., non-reflective, slightly reflective, and highly reflective [9]). However, analyzing written reflections is typically a manual process that is time and resource-intensive, so a key question is how to automatically model student written responses in order for stealth assessors to make robust predictions of learning outcomes.

This paper presents a multimodal multi-task machine learning framework for stealth assessment that involves generating sequential predictions of students' post-test and reflection depth scores based upon students' pre-test scores, game trace logs, and written reflections. The multimodal multi-task stealth assessment models utilize features extracted from students' game trace logs, including in-game events and dynamically changing learning goals, as well as linguistic features extracted from written reflections using a pre-trained ELMo language model based on a stacked bidirectional recurrent neural network [10]. We train stealth assessment models based on gated recurrent unit (GRU) neural architectures [11] that employ hybrid fusion to model multimodal data consisting of game trace logs and written reflection responses. We evaluate the predictive performance of multimodal, multi-task stealth assessment models against a random forest regressor as well as variants of unimodal and single-task GRU models.

2 Related Work

Multimodal machine learning techniques leverage information obtained across multiple data channels and have been found to show improvement in predictive performance as compared to unimodal models across a range of tasks [12]. These techniques have been widely investigated for a range of learning environments in the context of multimodal learning analytics [13, 14]. Aslan et al. developed a real-time multimodal analytics framework for measuring student engagement using appearance and context-based

features in an authentic classroom setting [15]. Emerson et al. analyzed multi-channel data including gaze, facial expression, and posture for predicting learner engagement in interactive museum exhibits [16, 17]. To effectively deal with heterogeneous streams of data, various data fusion methods have been explored along with multimodal learning analytics approaches. Zheng et al. [18] categorized data fusion techniques as being feature-based [19], stage-based [20] and semantic meaning-based [21]. Henderson et al. used multimodal data fusion for affect detection in game-based learning environments exploring feature-level as well as decision-level multimodal data fusion approaches [22]. Their results demonstrate that different data fusion approaches are suited to different affect detection tasks.

Another strand of related work is research on stealth assessment [6]. Stealth assessment is an application of evidence-centered design (ECD) [8], an assessment design framework that utilizes task-level evidence to infer learners' competencies on higherlevel concepts in game-based learning environments. Stealth assessment often utilizes game trace logs (e.g., sequence of events, locations visited) to model student knowledge, skills, and abilities using machine learning techniques [7, 23]. A promising approach for augmenting stealth assessment models is using data from student reflections [24]. Game-based learning environments can be designed to ask students to submit written reflections detailing what they have learnt and how their past experiences might shape their future learning plans, providing an important source of evidence for stealth assessment. Carpenter et al. analyzed students written reflections in a gamebased learning environment by rating students' reflections on a continuous scale based on their reflection depth [25]. They evaluated several machine learning techniques for their predictive performance on reflection scores. They reported the best performance using an SVM model utilizing ELMo embeddings. Geden et al. examined modeling approaches for predicting students' post-test scores using game-based features, pre-test scores and word embedding representations of reflection responses [5]. Our work leverages feature representations extracted from game interaction data and students' written reflections to predict a sequence of reflection scores and post-test scores cast as a time-series regression task in the context of multi-task learning [26].

3 Dataset

In this work, we use data from a pair of classroom studies conducted in 2018 and 2019 with CRYSTAL ISLAND, a game-based learning environment for middle school microbiology education [5, 25, 27]. The objective of the game is to find a disease that is spreading among a group of scientists on a remote island research station. In the game, the students explore the virtual environment, talk to non-player characters, read science books, articles and posters, test objects in a virtual laboratory, and submit a final diagnosis to the camp nurse. In both studies, a 17-item microbiology content knowledge pre-test (M = 6.78, SD = 2.75) was administered a week before the students first interacted with the game. Prior to game-based learning activities, researchers briefly introduced the students to the game, and then students interacted with the game until they solved the mystery, or approximately 100 minutes of gameplay time had elapsed. As

students interacted with CRYSTAL ISLAND (Figure 1, left), their game trace logs and written responses were recorded, providing a detailed account of each students' actions in the learning environment. After completion of the game, students completed a 17-item microbiology content knowledge post-test (M = 7.36, SD = 3.36), among other post-game measures. Our dataset consists of data from 119 students. Of these students, 51% identified as females, and their ages ranged from 13 to 14 (M = 13.6, SD = 0.51).



Fig. 1. (left) CRYSTAL ISLAND game-based learning environment, (right) In-game reflection prompt.

While playing the game, students were prompted at several checkpoints (Figure 1, right) to reflect on their progress in the game and state their plans moving forward. At the end of the game, two additional prompts appeared, requiring students to detail their approach towards solving the current problem and propose a way to solve a similar problem in the future. Across all types of reflection, there were 729 written reflection responses. The average length of a reflection response was approximately 20 words. A rubric was formulated for evaluating the reflection depth of written responses on a scale of 1 to 5 [25]. Two researchers individually rated the written responses following the rubric, and the final reflection rating was calculated by averaging the rating values for each written response (M = 2.41, SD = 0.86). An intraclass correlation of 0.669 was achieved indicating moderate inter-rater reliability.

4 Stealth Assessment Frameworks

4.1 Data Representation and Preprocessing

This work focuses on two game trace log features: events and completed plot points. We only consider 8 distinct game event types in CRYSTAL ISLAND—conversing with a virtual character, reading books and articles, filling out the diagnosis worksheet, encountering a written response prompt, accomplishing a goal, reading a poster, testing a virtual object, and submitting a diagnosis—and 20 distinct plot points, which are ingame milestone events important to successfully completing the game. The event feature is a cumulative count vector, with each element representing the number of occurrences of a specific event from the start of the game. We apply the standard score (z-score) for the count of each event in the event vector to standardize count-based representations. The plot point vector is a binary representation with each element

representing a unique gameplay milestone, set to zero by default and changed to one once the milestone is accomplished.

In addition to game trace log features, we also use ELMo embeddings [10] of students' written reflections. These written reflections are the responses provided by the students to the in-game prompts. ELMo sentence embeddings have the benefit of encoding contextual information. We use an ELMo model pre-trained with the 1 Billion Word Benchmark comprising approximately 800M tokens of news crawl data from WMT 2011 [10]. We consider the average of the ELMo word embeddings in each of the written responses to construct a single embedding of length 1,024 for each response. Since the embedding size is prohibitively large to effectively model our dataset, we reduce the dimensionality of the embeddings by applying principal component analysis (PCA) and preserving the 32 features that have the most variance. In our stealth assessment models, the most recent reflection embedding is passed as input, along with the game trace logs, at each time step. For predictions prior to a student's first reflection, an embedding with uniform random noise is used to indicate that no information is yet available. We also include students' standardized pre-test scores as input to our models.

We subsample each student's trace data into overlapping subsequences of 20 consecutive actions, such that each action taken by a student has a corresponding subsequence, comprising the current and past 19 actions. Both cumulative and binary vectors in these subsequences are based on actions performed from the start of gameplay until the current timestamp. The first few entries whose lengths are shorter than 20 are padded with zeros to keep the length of each subsequence to exactly 20 timesteps. The task of predicting the series of reflection ratings is formulated as sequential inferences on the rating of the next written response based on the current subsequence. We model our problem as a regression task. For each in-game event, we predict the post-test score as well as the reflection rating. The reflection rating label corresponds to the next written response that the student is expected to submit at a later time step.

4.2 Modelling Techniques

In our current work, we experiment with early, late, and hybrid fusion of multimodal data. We hypothesize that how to fuse different modalities plays a crucial role to obtain high-fidelity predictive models, particularly because each modality has a different nature, number of features, and representation method. We compare performance achieved using the three data fusion techniques and identify the best performing technique for each task. Multi-task learning (MTL) allows a model to share intermediate latent representations between related tasks, enabling the model to leverage information across related tasks to achieve better generalization performance [26]. We evaluate MTL focusing on whether shared representations for modeling related learning outcomes of post-test scores and written reflection ratings can benefit predictive performance for one or both tasks. For training each model, we used Adam optimizer [28] with mean squared error (MSE) as the loss function. The models are detailed as follows:

Early fusion model. This model takes a concatenation of pretest scores, written reflection embeddings and game trace logs as input, which is then passed through a GRU layer (128 hidden units, 0.1 dropout, 0.01 L2 kernel/recurrent/bias regularization factor), followed by a dense (64 units, ReLU activation) and a dropout layer (0.1). The hidden representation is passed through a dense layer and a separate dense layer for predicting post-test score and time-series reflection rating (each with 8 shared hidden units followed by 1 output unit). (Please note that the following late and hybrid fusion methods use the same hyperparameter configurations reported above.)

Late fusion model. This model takes written reflection embeddings and gameplay logs as inputs through separate GRU layers followed by dense and dropout layers. Pre-test scores are passed through a dense network. These three outputs are concatenated and passed to two dense layers predicting post-test score and time-series reflection rating.

Hybrid fusion model. This model (Figure 2) adopts a hybrid approach of early and late fusion. Written reflection embeddings and game trace logs are concatenated as input (i.e., early fusion) to a GRU layer. The output is passed through dense and dropout layers. The pre-test scores are combined with the output (i.e., late fusion) and passed through dense layers for predicting post-test score and time-series reflection rating.



Fig. 2. Hybrid fusion model for multi-task learning.

5 Results

To evaluate each model, we use student-level 5-fold cross validation. For the training set in each fold, we perform an 80-20 split to create a training and validation set, respectively. For evaluating model performance, we use R^2 score and MSE metrics. We use a random forest (RF) model as a baseline. This model takes one subsequence as input each time and predicts each label individually (i.e., single-task models only). We obtain one final post-test score by averaging multiple subsequence-level predictions for each student, since post-test score for a student is the same across all subsequences. On the other hand, number of reflection ratings varies per student. We thus average predictions across actions performed between consecutive reflection prompts.

The average cross-validation results comparing the performance of single and multitask models for different data fusion techniques can be seen in Table 1. For multimodal MTL experiments, the hybrid and late fusion GRU models have similar performance on post-test score prediction (R^2 score 0.30 and MSE 0.026), while the hybrid fusion GRU model outperforms both early and late fusion models for time-series reflection rating prediction (R^2 score 0.28 and MSE 0.004). For single-task models, we note that the best performing GRU models show similar performance on post-test score prediction as our RF baseline with R^2 score 0.20 and MSE 0.030. However, all single-task models perform poorly for reflection rating prediction, as they exhibit negative R^2 scores. This indicates that single-task learning is not an effective modeling approach for the task, fitting worse than the average value of unseen data. We further evaluate the contribution of each modality in the performance of the best performing models (i.e., GRUs with hybrid fusion) by testing different combinations of input modalities. We compare single and multi-task models' performance for post-test score prediction in Table 2 and for time-series reflection rating prediction in Table 3. A model has better performance if it has higher R^2 score and lower MSE. The results show that best performance is achieved with MTL models that utilize all three modalities.

 Table 1. Comparison of predictive performance of stealth assessment models using different data fusion techniques along with the random forest (RF) baseline (STL: single-task learning, MTL: multi-task learning, EF: early fusion, LF: late fusion).

Models	Post-test score prediction (R^2)	Post-test score prediction (MSE)	Time-series reflection rating prediction (R^2)	Time-series reflection rating prediction (MSE)
STL-RF	0.20	0.030	-0.41	0.008
STL-EF	0.20	0.030	-0.32	0.008
STL-LF	0.19	0.031	-0.20	0.007
STL-Hybrid	0.20	0.030	-0.33	0.008
MTL-EF	0.15	0.032	0.08	0.005
MTL-LF	0.30	0.026	0.13	0.005
MTL- Hybrid	0.30	0.026	0.28	0.004

 Table 2. Comparisons of single and multi-task GRU models with hybrid fusion for predicting post-test scores using different combinations of modalities.

Modalities	Single-task (<i>R</i> ²)	Single-task (MSE)	Multi-task (R ²)	Multi-task (MSE)
Pre-test score	0.23	0.029	0.23	0.029
Reflection	0.08	0.035	0.11	0.033
Game trace	-0.10	0.041	0.001	0.040
Pre-test, reflect	0.28	0.027	0.27	0.028
Pre-test, game trace	0.12	0.033	0.17	0.031
Reflect, game trace	0.09	0.034	0.15	0.032
Pre-test, reflect, game trace	0.20	0.030	0.30	0.026

We observe that the hybrid fusion, multi-task GRU model outperforms the RF baseline by achieving an R^2 score of 0.30 and 0.28 and MSE 0.026 and 0.004 for post-test score and time-series reflection rating prediction, respectively. We also note that the best single-task GRU models' performance for post-test score prediction is achieved by using a combination of pre-test scores and written reflections (Table 2).

Modalities	Single-task (<i>R</i> ²)	Single-task (MSE)	Multi-task (R ²)	Multi-task (MSE)
Pre-test score	-0.29	0.008	-0.29	0.007
Reflect	-0.24	0.007	0.04	0.005
Game trace	-0.46	0.008	-0.22	0.007
Pre-test, reflect	-0.27	0.008	0.11	0.005
Pre-test, game trace	-0.23	0.007	-0.16	0.006
Reflect, game trace	-0.18	0.007	0.17	0.004
Pre-test, reflect, game trace	-0.33	0.008	0.28	0.004

 Table 3. Comparisons of single-task and multi-task GRU models with hybrid fusion for predicting time-series reflection ratings using different combinations of modalities.

The single-task learning results indicate that students' post-test scores primarily depend on prior knowledge about the subject matter and how well they reflect in the game, while game trace data added noise to the model by decreasing predictive performance. However, we observe that MTL (0.30 R^2 score, 0.026 MSE and 0.28 R^2 score, 0.004 MSE for post-test scores and time-series reflection ratings respectively) lends itself to successfully model the features, outperforming all single-task models for both tasks. Among the three fusion techniques, hybrid fusion outperforms both early and late fusion in MTL, suggesting that (1) combining gameplay logs with written response embeddings early facilitates modeling of complementary relationships between these two modalities, and (2) fusing intermediate representations learned from gameplay logs and written response features with pre-test features late is effective for both prediction tasks. Together with the significant improvement achieved by MTL over single-task learning utilizing multimodal data for time-series reflection rating prediction, the results suggest it is beneficial to simultaneously learn shared intermediate representations using labels from two related tasks, boosting predictive accuracies of both learning outcomes.

6 Conclusion

Stealth assessment in game-based learning environments shows significant potential to support effective learning experiences for students. Given the positive relationship between student reflection and learning, we introduced a multimodal, multi-task stealth assessment framework to dynamically infer student competencies on content knowledge and reflection. We investigated multimodal data fusion techniques for GRU models to predict post-test scores and a series of written reflection ratings utilizing pretest scores, reflection embeddings, and game trace logs. We derived ELMo embeddings to represent written responses and explored three multimodal data fusion mechanisms including early, late and hybrid fusion combined with single and multi-task learning architectures. The results suggest that the multi-task, hybrid fusion model significantly outperforms the RF baseline. The best predictive performance was achieved by combining all modalities in a multi-task setting using our hybrid fusion GRU model.

In future work it will be important to experiment with other contextual language embeddings to further improve models' generalization performance. Subsequences at later timestamps are expected to be better predictors of post-test scores. Early prediction measures for regression tasks could help us understand how early in the game we can accurately predict scores. It will be important to investigate additional data modalities to pick up cues that could further improve prediction of learning outcomes. Other measures including affect and engagement have important relationships to learning and could offer additional insight into learning processes and outcomes.

References

- Clark, D. B., Tanner-Smith, E. E., Killingsworth, S. S.: Digital games, design, and learning: A systematic review and meta-analysis. Review of Educational Research, 86(1), 79-122 (2016).
- Qian, M., Clark, K. R.: Game-based learning and 21st century skills: A review of recent research. Computers in Human Behavior, 63, 50-58 (2016).
- Plass, J. L., Mayer, R. E., Homer, B. D. (Eds.).: Handbook of Game-Based Learning. MIT Press (2020).
- Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., Lester, J. C.: Detecting Off-Task Behavior from Student Dialogue in Game-Based Collaborative Learning. In: International Conference on Artificial Intelligence in Education, pp. 55-66. Springer, Cham (2020).
- Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., Lester, J.: Predictive student modeling in game-based learning environments with word embedding representations of reflection. In: International Journal of Artificial Intelligence in Education, 31(1), 1-23 (2021).
- Shute, V. J.: Stealth assessment in computer-based games to support learning. Computer Games and Instruction, 55(2), 503-524 (2011).
- Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., Lester, J. C.: DeepStealth: Game-based learning stealth assessment with deep neural networks. IEEE Transactions on Learning Technologies, 13(2), pp. 312-325 (2019).
- Mislevy, R. J., Almond, R. G., Lukas, J. F.: A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1), i-29, (2003).
- Jung, Y., Wise, A. F.: How and how well do students reflect? multi-dimensional automated reflection assessment in health professions education. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 595-604. (2020).
- Peters, M. E.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018).
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).

- Baltrušaitis, T., Ahuja, C., Morency, L. P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443 (2018).
- Blikstein, P., Worsley, M.: Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. Journal of Learning Analytics, 3(2), 220-238 (2016).
- Oviatt, S., Grafsgaard, J., Chen, L., Ochoa, X.: Multimodal learning analytics: Assessing learners' mental state during the process of learning. In: The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2, pp. 331-374 (2018).
- Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., Arslan Esme, A.: Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-12 (2019).
- Emerson, A., Henderson, N., Rowe, J., Min, W., Lee, S., Minogue, J., Lester, J.: Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 107-116 (2020).
- Emerson, A., Henderson, N., Rowe, J., Min, W., Lee, S., Minogue, J., Lester, J.: Investigating Visitor Engagement in Interactive Science Museum Exhibits with Multimodal Bayesian Hierarchical Models. In: International Conference on Artificial Intelligence in Education, pp. 165-176. Springer, Cham (2020).
- Zheng, Y.: Methodologies for cross-domain data fusion: An overview. IEEE transactions on big data, 1(1), 16-34 (2015).
- Fu, Y.: Sparse real estate ranking with online user reviews and offline moving behaviors. In: 2014 IEEE International Conference on Data Mining, pp. 120-129. IEEE (2014).
- Zheng, Y., Liu, Y., Yuan, J., Xie, X.: Urban computing with taxicabs. In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 89-98. (2011).
- Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H.: Discovering urban functional zones using latent activity trajectories. IEEE Transactions on Knowledge and Data Engineering, 27(3), 712-725 (2014).
- Henderson, N.: Improving affect detection in game-based learning with multimodal data fusion. In: International Conference on Artificial Intelligence in Education, pp. 228-239. Springer, Cham, 2020.
- Kim, Y. J., Almond, R. G., Shute, V. J.: Applying evidence-centered design for the development of game-based assessments in physics playground. International Journal of Testing, 16(2), 142-163 (2016).
- 24. Luo, W., Litman, D.: Determining the quality of a student reflective response. In: The Twenty-Ninth International FLAIRS Conference, pp. 226-231. (2016).
- Carpenter, D., Geden, M., Rowe, J., Azevedo, R., Lester, J.: Automated Analysis of Middle School Students' Written Reflections During Game-Based Learning. In: International Conference on Artificial Intelligence in Education, pp. 67-78. Springer, Cham (2020).
- Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering. IEEE. (2021).
- Rowe, J. P., Shores, L. R., Mott, B. W., Lester, J. C.: Integrating learning, problem solving, and engagement in narrative-centered learning environments. International Journal of Artificial Intelligence in Education, 21(1-2), 115-133 (2011).
- Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

10