

Exploring the use of probabilistic latent representations to encode the students' reading characteristics

Erwin D. Lopez Z.¹, Tsubasa Minematsu¹, Taniguchi Yuta¹, Fumiya Okubo¹ and Atsushi Shimada¹

¹ Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan

Abstract

The emergence of digital textbook reading systems such as Bookroll, and their ability of recording reader interactions has opened the possibility of analyzing the students reading behaviors and characteristics. To date, several works have conducted compelling analyses characterizing the different types of students with the use of clustering ML models, while others have used supervised ML models to predict their academic performance. The main characteristic these models share is that internally they simplify the students' data into a latent representation to get an insight or make a prediction. Nevertheless, these representations are oversimplified, otherwise difficult to interpret. Accordingly, the present work explores the use of Variational Autoencoders to make more interpretable and complex latent representations. After a brief description of these models, we present and discuss the results of four explorative studies when using the LAK22 Data Challenge Workshop datasets. Our results show that the probabilistic latent representations generated by the proposed models preserve the student reading characteristics, allowing a better visual interpretation when using 3 dimensions. Also, they allow supervised regressive and classification models to have a more stable and less overfitted learning process, which also allows some of them to make better score predictions.

Keywords

Digital textbook, educational data encoding, academic performance prediction, educational data visualization, Variational Autoencoder

1. Introduction

The Bookroll application is an e-book reading system for distributing class reading materials that also provides students with a variety of digital interactions, such as adding memos, jumping to an arbitrary page, highlighting text, etc. [1,2]. Moreover, all these interactions are recorded in a local database, which leaves open the possibility to analyze these stream data to better understand the students reading behaviors and characteristics [1].

Different works have carried out compelling analyses by processing these data and identifying students' clusters and tendencies [3,4,5,6,7]. While all of them have used clustering techniques, the chosen students' data representation varies from one work to another. For example, works [3,4] selected time-related features obtained from cross-referencing the Bookroll timestamps and the LMS schedule information (e.g., the total time spent on content during the class), whereas works [5,6,7] selected some Bookroll events frequency values as features (e.g., the number of NEXT events) along with indirect observable features (e.g., the ratio between NEXT and PREV events, the total READING TIME).

Meanwhile, other works have attempted to use the Bookroll data to predict students' low academic performance [8,9,10,11]. For this purpose, they have made use of ML supervised learning models (e.g., SVM, RNN models) to map student data representations to their most probable score. From these data representations, we could identify that works [9,10] incorporated some Bookroll events frequency

values along the reading time, whereas works [8,11] used a vector representation known as ALP (Active Learner Point), which is based not only on Bookroll but also on Moodle LMS data.

From the first type of works, it is clear that an ML unsupervised learning model such as a clustering model can explore the students' characteristics and provide us with interesting insights. However, while clustering methods are useful for analyses and possess good interpretability, they assign the same latent representation to several different students (students in the same cluster are represented with the same label). On the other hand, ML supervised models such as Neural Networks assign different latent representations to each student (hidden layers representations) but usually, these representations are difficult to interpret, and their spatial locations do not provide any additional information (two close latent representations do not necessarily correspond to two similar student reading characteristics).

In this context, the present work aims to explore the latent representations generated by other ML unsupervised models, more specifically Variational Autoencoders [12], when processing Bookroll data from two different contexts, specifically the Kyoto and Kyushu University datasets provided by the organizers of the LAK22 Data Challenge Workshop.

2. Variational Autoencoders

In the present study, we propose two models based on the Variational Bayesian approach for latent representation presented in [12]. Since a variational autoencoder is a specific form of an autoencoder we will briefly describe the autoencoder model before presenting the two used models.

2.1. Autoencoders

As shown in Figure 1, an autoencoder is a neural network with a special layer distribution that is trained to attempt to copy its input to its output [13]. Its special layer distribution can be decomposed into two parts: an encoder and a decoder. Since the encoder translates the input into a latent low-dimensional vector, the model can't learn the identity function. This means that cannot copy the input values to the output values for any given input, but only approximate the input when it preserves the same characteristics of the inputs used for training the model.

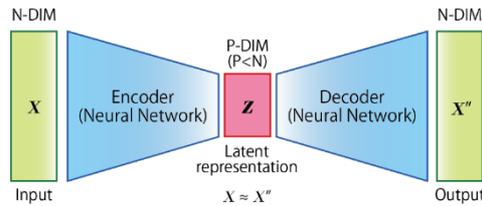


Figure 1: Autoencoder model architecture.

2.2. First model: β -Variational Autoencoders (β -VAE)

While an autoencoder maps its input to a fixed latent representation, the Variational Autoencoder (VAE) is a model that maps its inputs to a probability distribution of possible latent representations. Since attempting to learn the encoder and decoder probabilistic distributions under this consideration leads to intractability problems, we can use the variational Bayesian approach [12]. To this purpose, we consider $q_\phi(z|x)$ as an approximation of $p_\theta(z|x)$ and define this new distribution to be a known probability distribution (usually a normal distribution) before attempting to learn all the distributions.

Therefore, as shown in Figure 2, the VAE can be trained just as an autoencoder that learns the mean and standard deviation of the latent representations' probabilistic distribution while assuring they follow the normal distribution. This can be achieved by using the Loss function presented in equation 1 where D_{KL} is the KL divergence that decreases when $q_\phi(z|x)$ is near to be a normal distribution.

$$Loss = MSE_{autoencoder} + D_{KL}, \quad (1)$$

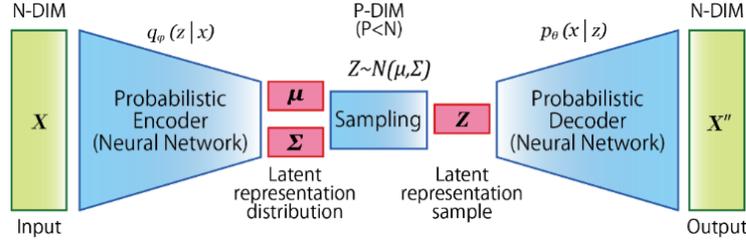


Figure 2: Variational Autoencoder model architecture.

Moreover, a β -VAE model is a VAE where the D_{KL} loss takes on greater importance after being multiplied by a hyperparameter β as shown in equation 2. This modification allows the model to assure that the latent space dimensions are not entangled and have better interpretability [14].

$$Loss = MSE_{autoencoder} + \beta D_{KL}, \quad (2)$$

Finally, since by definition the probabilistic latent representation “z” generates the input data with a certain likelihood distribution $p_{\theta}(x|z)$, when using Bookroll data as inputs, “z” could be representing some students' psychological characteristics that influence their decisions and actions when interacting with the Bookroll system (interactions that lastly generate the data).

2.3. Second model: Supervised β -Variational Autoencoders (β -SVAE)

Considering that the latent representations could be regarded as some students' psychological characteristics related to the Bookroll dataset generation process, we could consider that these representations are also related to the students' final score. In this case, there would exist a probability distribution $p_{\theta}(s|z)$ that could be modeled by a Deep Neural Network.

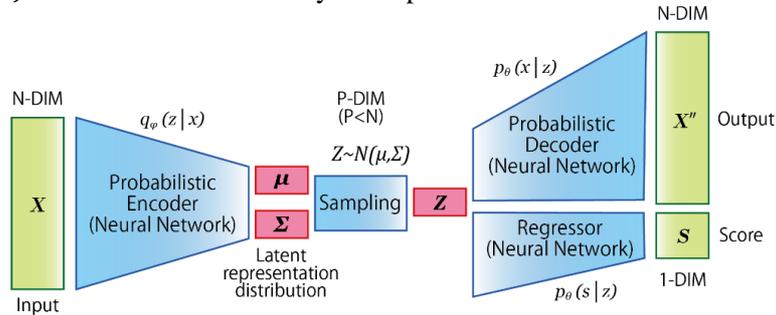


Figure 3: Supervised Variational Autoencoder model architecture.

From a similar intuition, the work [15] proposed the SVAE model shown in Figure 3. As the reader may note, the model is considerably similar to the VAE model but additionally considers an internal regressor model, requiring labels data for its training. In our case, this regressor model will oversee the mapping from the latent representations to their corresponding test scores. Therefore, the training loss function should be defined as shown in equation 3, where in addition to β , we also considered two additional weighing hyperparameters α and γ to overcome the scaling differences between the decoder reconstructing and the internal regressor errors.

$$Loss = \alpha MSE_{autoencoder} + \gamma MSE_{regressor} + \beta D_{KL}, \quad (3)$$

3. Kyoto Dataset Work

Since Kyoto and Kyushu University datasets were collected from different educational environments and exhibit different characteristics, we have employed different methodologies and carried out different studies for each educational context.

3.1. Method

The Kyoto dataset was composed of 11,091,211 logs from 10,001 secondary school students' stream data and their corresponding final total scores. This dataset recorded 15 different event actions, along with user device information, timestamps, and other special event characteristics (e.g., the length of the memo that was written on the page).

As shown in Figure 4, we converted each event action and device type categorical included in one student's stream data to one-hot encoding representations and then added all of them up, resulting in 18 columns that preserve the number of occurrences of each event and type of device. Additionally, we processed the timestamps column with an algorithm capable of calculating the student's total reading time, as well as the number of times that the Bookroll platform was automatically closed by the system time-out feature. Thus, our student vector representation was composed of 20 different features.

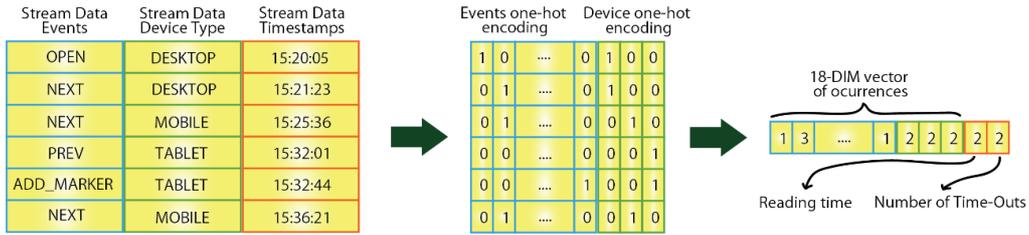


Figure 4: Student's stream data representation method.

We then iterated this process for each student included in the dataset and created a matrix consisting of 20 columns and 10,001 rows that we define as our non-labeled dataset X. Next, we cross-referenced each student in this dataset with his/her corresponding final total score and got a column of 10,001 rows that represents our labels dataset Y. Then, we split these datasets into the X_{train} , X_{test} , Y_{train} , and Y_{test} subsets, where X_{train} , Y_{train} contains data from 6,600 randomly selected students, and X_{test} , Y_{test} include the complementary data (the 3,401 students not included in X_{train} and Y_{train}).

Finally, we conducted two exploratory studies with these datasets, which are described as follows.

3.1.1. First exploratory study: Data visualization

This first study aims to explore the visual characteristics of the vector representations generated by a β -SVAE. Since the Kyoto dataset includes score labels for all of their 10,001 students, it meets the labels requirement of this supervised model.

For this study, we trained a β -SVAE with an encoder of 2 layers of 19 units each, two additional layers of 3 units for the mean and standard deviation encodings, a decoder of 4 layers of 19, 19, 20, and 20 units; and a regressor of 2 layers of 3 and 1 unit. This optimal architecture was found after exploratory experiments with the support of the optimization framework Optuna [16].

Thereby, we obtained an encoder able of reducing our 20-dimensional student's representation to a probable 3-dimensional vector. To visualize these probable vectors, we generated a set of 25 samples drawn from a normal distribution with the means and standard deviations predicted by the β -SVAE encoder. Then we plotted the 25 vector representations of all the students who belong to the test set.

The main purpose of this study was to visualize how are distributed the students in the latent space, whether the students with similar characteristics are closely distributed, and whether exists a tendency that represents visually the relationship between the students' spatial distribution and their final scores.

3.1.2. Second exploratory study: Score regression performance

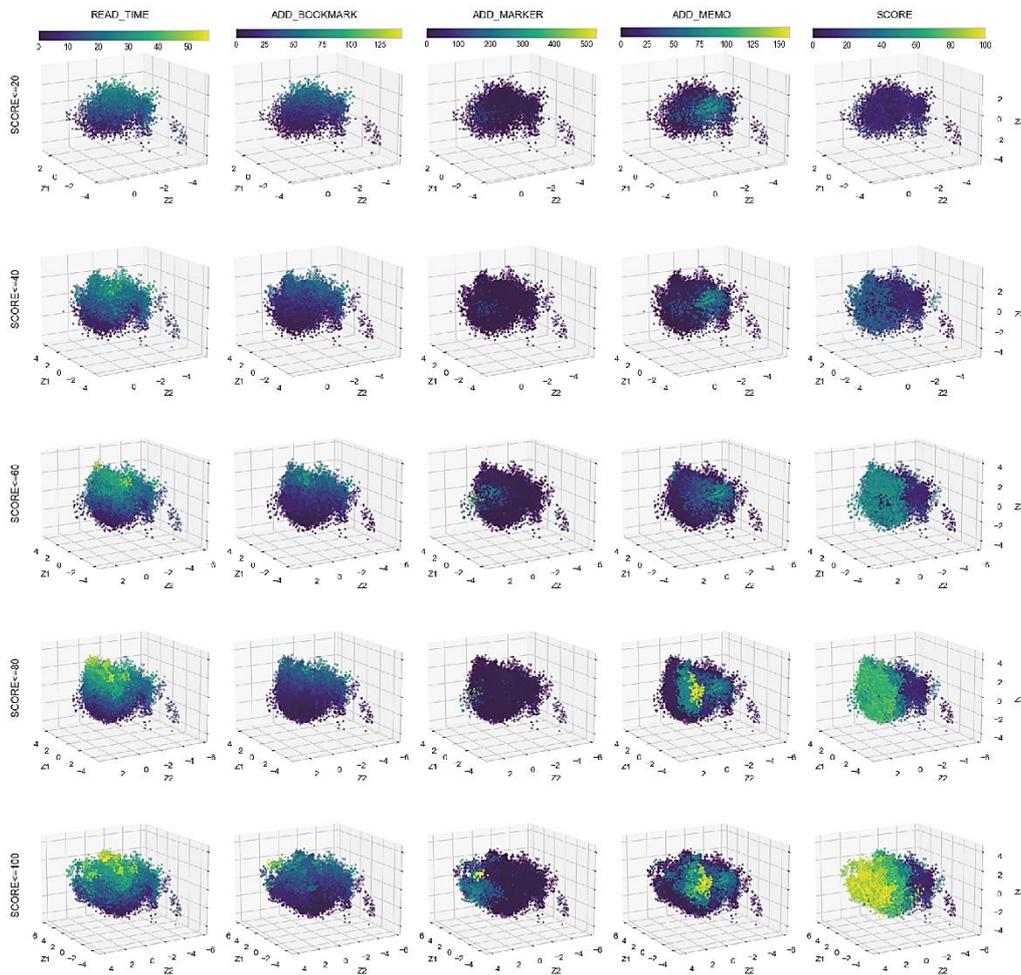
This second study aims to explore the loss or gain of score predictive performance when using latent encodings generated by the β -VAE and β -SVAE models. Since these representations are lower-dimensional, there is a part of the student data that has been lost. While this means an information loss, it could also mean an omission of unnecessary information and a more general representation.

Table 1

Deep Neural Networks architectures used in the second exploratory study

Model	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
DNN	20 units	20 units	3 units	3 units	1 unit
Shallow DNN	6 units	3 units	1 unit	-	-
SVAE DNN	3 units	3 units	2 units	1 unit	-
VAE DNN	8 units	2 units	1 unit	-	-

For this study, we employed the latent encodings generated by the model of the first study. Additionally, we trained a β -VAE with the same architecture, but with an 8-dimensional latent representation instead of 3. Then we generated the encodings by using the encoder part similarly to the first study. Finally, we split the test data into 4 subsets with a 4-fold cross-validation method and used each subset for training 4 different deep neural networks, which architectures can be found in Table 1. The first two models were trained with the 20-dimensional features representations, the third with the β -SVAE encodings, and the fourth with the β -VAE encodings. The first model architecture was based on the β -VAE encoder architecture to disregard the variational models' architecture influence, while the second model architecture was optimized via exploratory experimentation.

**Figure 5:** Students' 3-dimensional latent representations (Z1, Z2, and Z3: latent space axes).

3.2. Results

3.2.1. First exploratory study: Data visualization

We plotted the 25 samples from the 3,401 students of the test set with the use of the Matplotlib Python library [17]. A summary of the results can be seen in Figure 5. Here each point denotes a latent

representation sample from one student, while its color represents the number of times that the event indicated on the top of the column has occurred in the data of the respective student. Two special cases are the first and last columns, where the color represents the reading time and score itself respectively. On the other hand, the rows represent a condition for plotting the students' representations, meaning that in the first row we only plotted the encodings of the students with a score lower than 20, and so on.

Even though the best way to present this latent space is probably with an interactive application where the user can select the reading characteristic that want to be colored and which students want to visualize, within the limitations of this written work we decided to present these variants to facilitate the reader's visual inspection. From here the reader can observe in the last row that the students with similar reading characteristics (e.g., a similar number of ADD_MEMO) are nearly spatially located (the color tones change smoothly in space). Moreover, there exist some directions where a certain reading characteristic is incrementing (For example reading time in Z3 and score in Z2). Therefore, a teacher can evaluate the students' final score based on their position in the latent space and compare the reading characteristics differences between them. For example, in row 1 column 1, around 40% of the low-score students read around 10 hours, 20% around 20 hours, and the other 40% around 30 hours.

3.2.2. Second exploratory study: Score regression performance

We plotted the models' validation MSE during the training processes for each fold as shown in Figure 6. The logarithmic scale was chosen considering that the initial large MSE error did not contribute to a better comparison.

From these graphs, we can observe that the variational models slightly outperformed the Shallow DNN model in all the validation subsets, while the first DNN model was overfitted on the train subsets in all the cases. From these results, we can conclude that the variational models have compressed the students' data ensuring as much as possible that the lost information was not too important for characterizing the students reading behaviors or their relationship with the final score outcomes. Also, despite being necessary deeper analyses and explorative trials, these results suggest that the latent encodings are more informative than their higher dimensional counterparts allowing an improvement in the score regression performance.

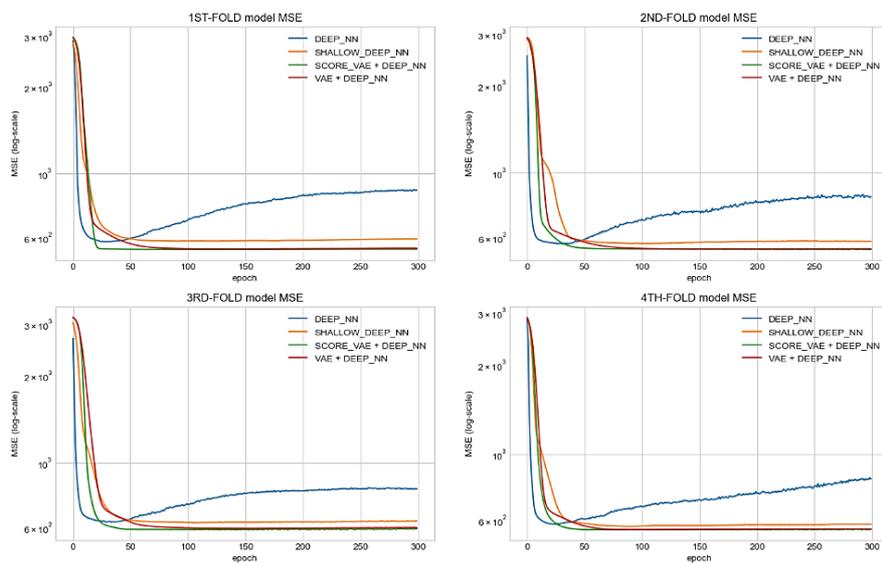


Figure 6: 4-Fold cross-validation results of the 4 proposed DNN for the score regression task.

4. Kyushu Dataset Work

4.1. Method

The Kyushu Dataset was composed of 4 datasets from two Kyushu University different courses, encrypted as A and B, and collected in the years 2019 and 2020. We identified from the 4 datasets the

presence of 826,874 logs in total, which represent approximately 7.45% of the number of logs in the Kyoto dataset. Besides, unlike the Kyoto dataset, this dataset recorded 17 different event actions and labeled their final scores as one of five letter grades (A, B, C, D, and F).

We converted students' stream data to 37-dimensional vectors with the same methodology adopted in section 3.1. The difference in dimensions is because in this case, we considered a 32-dimensional one-hot encoding of the event actions. This consideration was made to use a previously designed β -VAE architecture. Finally, we split each course dataset into the X_{train} , X_{test} , Y_{train} , and Y_{test} subsets, but due to the small amount of data, this time we preserved 50% of them in the test set.

Additionally, we created a dataset containing 9,885,286 logs of unlabeled data from 4,649 Kyushu University students enrolled in 47 different courses of the years 2018, 2019, and 2020. Then, we represented all students in this dataset as 37-dimensional vectors and used them to train a β -VAE with the next architecture: encoder with 2 layers of 36 units, two layers of 8 units for encoding the mean and standard deviation, and a decoder with 3 layers of 36, 36, and 37 units.

Finally, we have conducted two additional exploratory studies considering this model and datasets.

4.1.1. Third exploratory study: Score classification performance

This third study has a similar motivation to the second since it aims to explore the loss or gain of score predictive performance. However, in this case, the scores were represented as letter grades and consequently, the models were trained for a classification task.

First, we generated the Z_{train} , Z_{test} latent encodings by processing the X_{train} , X_{test} subsets using the β -VAE encoder and a normal sampling of 25 samples. Then, we trained one DNN model with the original X_{train} , Y_{train} and another with the Z_{train} , Y_{train} (Y_{train} was oversampled to match the Z_{train} dimensions). The architectures of these models are described in Table 2.

Table 2

Deep Neural Networks architectures used in the third exploratory study

Model	Layer 1	Layer 2	Layer 3
DNN	[12-20] units	5 units	Softmax
VAE DNN	8 units	5 units	Softmax

4.1.2. Fourth exploratory study: Reconstruction Loss distribution

This last exploratory study aims to point the principal differences between the original datasets and their corresponding reconstructions. By doing so, we can be aware of the encoding information loss and the main differences between the characteristics of the data used for training and validation. To this end, we just used the whole β -VAE model to generate reconstructions of the datasets of each course and calculated the average reconstruction difference on each feature. Finally, we plotted the results.

4.2. Results

4.2.1. Third exploratory study: Score classification performance

The third study's results are summarized in Figure 7. We selected the Precision, Recall, and F1-score metrics for this classification task. Since the validation data represented 50% of the total data in all the courses, we also plotted the training results for a better contextualization.

These graphs show a slight outperformance of the variational models in the A-2019 and B-2019 courses, while they do not present evidence of harmful information loss in the other 2 courses. Also, the models which used the latent representations have fewer problems in generalizing their prediction to the test set. These results align with the results of the second study, incrementing the amount of evidence that suggests that the probabilistic latent encodings are a more informative representation of the data.

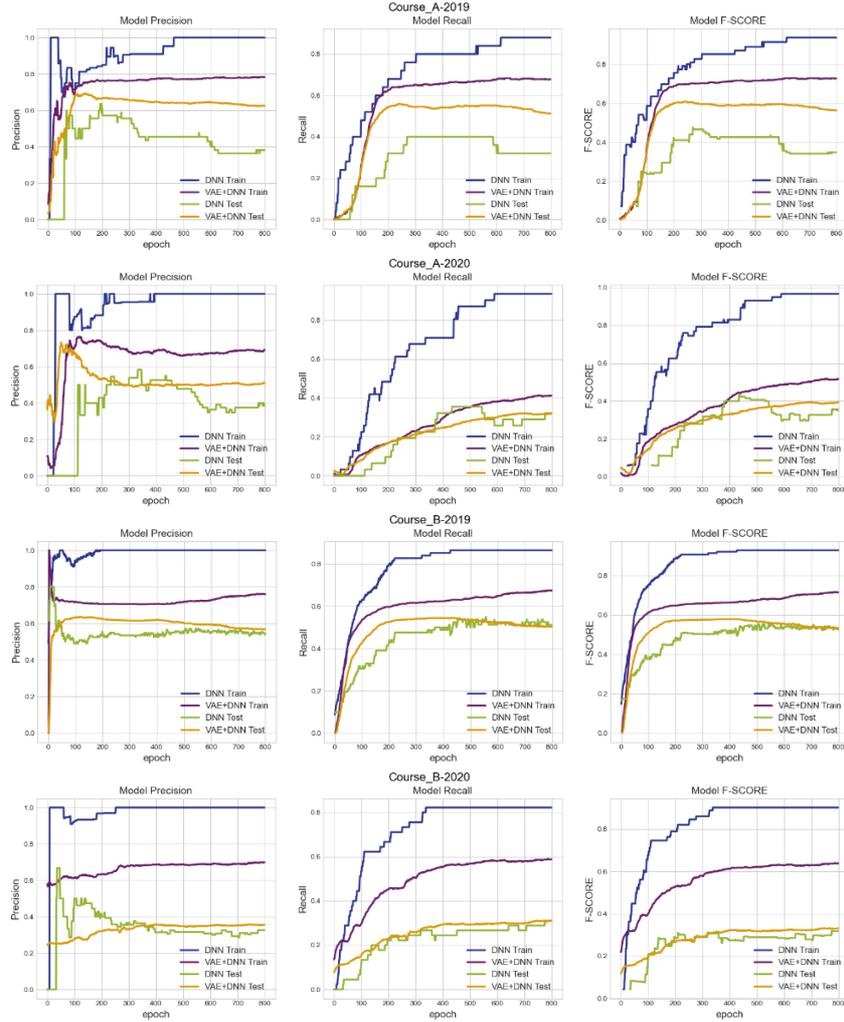


Figure 7: Train and test results of the 2 proposed DNN for the classification regression task.

4.2.2. Fourth exploratory study: Reconstruction Loss distribution

Finally, the results of the fourth study from Figure 8 show that in general, the features TIME_OUT and GET_IT suffered from a greater information loss. After considering the results from the third study where Course A-2019 exhibited the better outperformance, we presume that in this case, the lower number of TIME_OUT occurrences considered by the β -VAE model did not harm the classification performance, whereas the lower number of GET_IT occurrences considered by the model in the courses A-2020, B-2019 and B-2020 prevented itself to achieve a better classification performance.

Also, we can interpret the reconstruction difference values as the general differences between the training dataset and the Kyushu dataset. This interpretation is based on the autoencoders characteristic of reconstructing their inputs only when they share similar characteristics with the training data. It means that the reconstructions generated from inputs with some dissimilarities will exhibit a data coherence that would be present in the data if these dissimilarities did not exist. For example, without carrying out a deeper analysis, we can affirm that in the A-2019 course the number of bookmarks events was high relative to the other events if we consider the usual event relationships present in the other 46 courses of the training dataset.

Using a similar approach, we can compare the 2019 and 2020 versions of each course using the training data characteristics as a reference point. For example, in both courses but especially in Course A, we observe that, even though the average reading time is similar in both years, the students spent less time in proportion to their activity in the year 2020. In other words, they were more active on Bookroll during the online classes of the year 2020.

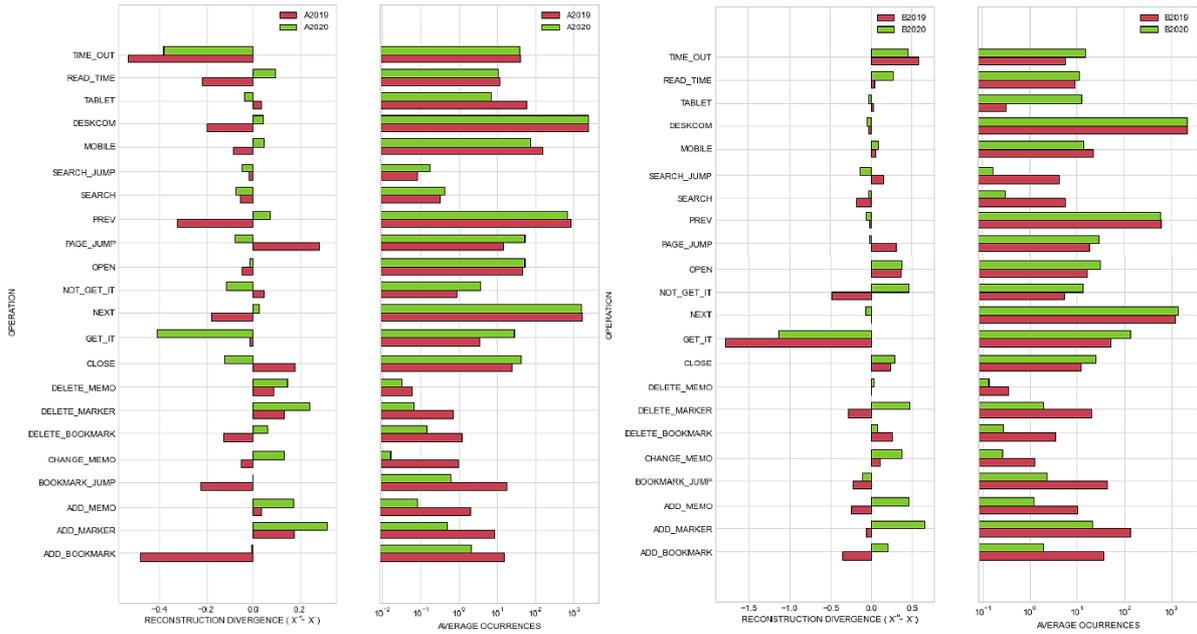


Figure 8: Average reconstruction difference values for each feature and their occurrence distribution.

5. Discussion

To the best of our knowledge, the probabilistic latent encodings from the first study represent the first form of visualizing Bookroll data in a human-interpretable 3-dimensional space. Also, these representations open new analysis possibilities to get better students' data insight. For example, using a similar approach of the work [3], researchers could identify the existent shades of the spectrum between two different study approaches. However, based on the fourth study results, we would like to point that in the case of conducting studies similar to [3,4,5,6,7], researchers will get better results if they consider data from similar sources (e.g., data from courses with the same professor). This is a result of the generalization effects of the autoencoder models and their ability of modeling data with similar characteristics. Nevertheless, this simplification is more regulated than the present in clustering models, which allows a less overfitted model and more general results (clustering works tend to conduct analyses only on the model training data).

Besides, in the same way is easier for a human to interpret the Bookroll data when using the latent encodings, we can expect a machine to have fewer problems learning to interpret and map the students' data to their corresponding score. The results from the third and fourth studies showed evidence that supports this intuition. For example, all the models trained with the latent representation samples developed a more stable and less overfitted learning process. We have also observed a similar or better performance of these models, suggesting that works such as [8,9,10,11] could improve their performance results with the use of these latent encodings. The intuition of this performance improvement would be that the non-labeled data preserve the general reading behaviors information and the β -VAE models trained with them store this information for enriching the students' data representation. Additionally, the presented models can also be used for binary classification, exhibiting the same benefits during the training process and validation results [18]. Since the educational datasets for this task present a data imbalance issue, the probabilistic modeling of the VAE models allows the design of different methodologies to address this problem, as the work done by [19].

However, our results are not conclusive and further research is required. Also, despite a possible performance improvement, β -VAE encodings require an additional amount of data to train the model that could be used for the main training process. Yet, the results of the second study showed that the β -SVAE and β -VAE models allowed the same performance improvement, meaning that researchers that count with large amounts of unlabeled data could benefit from the use of these models.

6. Acknowledgments

This work was supported by JST AIP Grant Number JPMJCR19U1, and JSPS KAKENHI Grant Number JP18H04125, Japan.

7. References

- [1] B. Flanagan, H. Ogata, Integration of Learning Analytics Research and Production Systems While Protecting Privacy, in: Workshop Proceedings of ICCE2017, 2017, pp.333-338.
- [2] H. Ogata, M. Oi, K. Mohri, F. Okubo, A. Shimada, M. Yamada, J. Wang, and S. Hirokawa, Learning analytics for e-book-based educational big data in higher education, Smart Sensors at the IoT Frontier, Springer, Cham (2017): 327-350.
- [3] G. Akçapınar, A. C. Mei-Rong, R. Majumdar, B. Flanagan, and H. Ogata, Exploring student approaches to learning through sequence analysis of reading logs, in: Proceedings of the 10th LAK'20, 2020, doi:10.1145/3375462.3375492
- [4] G. Akçapınar, M.N. Hasnine, R. Majumdar, A. C. Mei-Rong, B. Flanagan and H. Ogata, Exploring Temporal Study Patterns in eBook-based Learning, in: Proceedings of ICCE2020, 2020, pp. 342-347.
- [5] C. Yang, B. Flanagan, G. Akçapınar and H. Ogata, Investigating Subpopulation of Students in Digital Textbook Reading Logs by Clustering, in: Companion Proceedings of the 9th LAK'19, 2019, 465-470.
- [6] G. Akçapınar, R. Majumdar, B. Flanagan, H. Ogata, (2018). Investigating Students' e-Book Reading Patterns with Markov Chains, in: Proceedings of ICCE2018, 2018, 310-315.
- [7] C. Yin, M. Yamada, M. Oi, A. Shimada, F. Okubo, K. Kojima, and H. Ogata, Exploring the Relationships between Reading Behavior Patterns and Learning Outcomes Based on Log Data from E-Books: A Human Factor Approach, International Journal of Human-Computer Interaction 35.4-5 (2019): 313-322
- [8] F. Okubo, T. Yamashita, A. Shimada, Y. Taniguchi, and S. Konomi, On the prediction of students' quiz score by recurrent neural network, in: CEUR Workshop Proceedings 2163, 2018.
- [9] M. N. Hasnine, G. Akçapınar, B. Flanagan, R. Majumdar, K. Mouri, and H. Ogata, Towards Final Scores Prediction over Clickstream Using Machine Learning Methods, in: Proceedings of ICCE2018, 2018, pp. 399-404.
- [10] C. Chen, S. J. Yang, J. Weng, H. Ogata, and C. Su, Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers, Australasian Journal of Educational Technology 37.4 (2021): 130-144. doi:10.14742/ajet.6116
- [11] R. Murata, T. Minematsu and A. Shimada, Early Detection of At-risk Students based on Knowledge Distillation RNN Models, in: Proceedings of the 14th International Conference on Educational Mining, 2021, pp. 699-703.
- [12] D. P. Kingma, and M. Welling, Auto-Encoding Variational Bayes, *CoRR*, *abs/1312.6114*, 2014.
- [13] I. J. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, Cambridge, MA, 2016.
- [14] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, in: Proceedings of the 5th ICLR, 2017.
- [15] T. Ji, S. T. Vuppala, G. Chowdhary, and K. Driggs-Campbell, Multi-Modal Anomaly Detection for Unstructured and Uncertain Environments, in: Conference on Robot Learning (CoRL), 2020.
- [16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama (2019). Optuna: A Next-generation Hyperparameter Optimization Framework, in: International Conference on Knowledge Discovery & Data Mining (KDD '19), 2019, doi:10.1145/3292500.3330701
- [17] J. D. Hunter, Matplotlib: A 2D graphics environment, Computing in Science & Engineering 9.3 (2007): 90-95. doi: 10.1109/MCSE.2007.55
- [18] E. D. Lopez, T. Minematsu, Y. Taniguchi, F. Okubo, and A. Shimada, Encoding students reading characteristics to improve low academic performance predictive models, in: Companion Proceedings of the LAK'22, (2022). To appear.
- [19] Du, X., Yang, J., and Hung, J. L. (2020). An Integrated framework based on latent variational autoencoder for providing early warning of at-risk students. IEEE Access, 8, 10110-10122.