Educational and Information Technologies
https://doi.org/10.1007/s10639-023-11592-0

Using attention-based neural networks for predicting student learning outcomes in service-learning

Eugene Yujun Fu[1,2] · **Grace Ngai**[1,3] · **Hong Va Leong**[1] · **Stephen C.F. Chan**[3] · **Daniel T.L. Shek**[4]

Received: 9 June 2021 / Accepted: 9 January 2023

Abstract

As a high-impact educational practice, service-learning has demonstrated success in positively influencing students' overall development, and much work has been done on investigating student learning outcomes from service-learning. A particular direction is to model students' learning outcomes in the context of their learning experience, i.e., the various student, course, and pedagogical elements. It contributes to a better understanding of the learning process, a more accurate prediction of students' attainments on the learning outcomes, and improvements in the design of learning activities to maximize student learning. However, most of the existing work in this area relies on statistical analysis that makes assumptions about attribute independence or simple linear dependence, which may not accurately reflect real-life scenarios. In contrast, the study described in this paper adopted a neural network-based approach to investigate the impact of students' learning experience on different service-learning outcomes. A neural network with attention mechanisms was constructed to predict students' service-learning outcomes by modeling the contextual information from their various learning experiences. In-depth evaluation experiments on a large-scale dataset collected from more than 10,000 students showed that this proposed model achieved better accuracy on predicting service-learning outcomes. More importantly, it could capture the interdependence between different aspects of student learning experience and the learning outcomes. We believe that this framework can be extended to student modeling for other types of learning activities.

Keywords Computational modeling · Learning experience · Learning outcomes · Service-learning · Neural networks

✉ Grace Ngai
grace.ngai@polyu.edu.hk

Extended author information available on the last page of the article.

Published online: 01 April 2023

1 Introduction

Service-learning is a popular educational pedagogy worldwide, and many studies have documented its positive influence on students' development, particularly associated with their intellectual, social, civic, and personal learning outcomes (Astin et al., 2000; Celio et al., 2011; Yorio & Ye, 2012). Modeling students' learning experiences in service-learning courses empowers instructors to better understand the process of service-learning, and the corresponding factors such as the student, course, and pedagogical characteristics, including the student's motivation for taking the course, interest, and experience in the course and project. By modeling them, we can uncover impactful factors, and interpret how each individual factor impacts students' attainment of particular learning outcomes. However, service-learning usually involves a wide range of highly inter-correlated, interactive and non-linearly related variables, which poses a challenge for statistically-driven methods. Most of the existing analyses are restricted to small-scale datasets with a limited diversity of student samples, which limits their generalizability.

This paper presents an investigation that uses machine learning to analyze the impact of students' learning experiences on different service-learning outcomes. Machine learning approaches have been demonstrated to have the potential to encode representational information from noisy data, in essence directly solving for the matrix of variables rather than trying to isolate individual variables. Their effectiveness in real applications is evident from previous work across a variety of areas. In addition, they can be augmented with other techniques. For example, neural networks can be augmented with attention mechanisms, which assist in discriminating the significance of different inputs and helping a model to attend to the most important ones (Bahdanau et al., 2014). It is therefore expected that machine learning models will be similarly effective at investigating the impact of different learning experiences on different service-learning outcomes.

In our study, a deep neural network model with attention mechanisms was constructed to tackle the problem of predicting service-learning outcomes by modeling the contextual information directly from various student, course and learning characteristics. In previous work, embedding layers in deep neural network models have been shown to work well in capturing indicative representation of inputs (e.g., words) into fixed-size and low-dimension feature vectors. We postulated that for our problem, the embedding layers would similarly be able to encode students' learning experiences into representative features. These representative features can then be further analyzed by attention modules that can identify the interdependence between different learning experiences, and determine the importance of each learning experience in predicting different service-learning outcomes.

The proposed model thus, in the process of predicting students' learning outcomes from their learning experiences, conducted an analysis of the effects of different aspects of the student learning experiences on the service-learning outcomes. To our best knowledge, this is the first study that applies machine learning techniques towards this problem.

Our model was evaluated with a large dataset from more than 10000 students, collected from actual service-learning programs across a broad range of projects and

disciplinary topics. The experimental results illustrated that our model could effectively predict students' service-learning outcomes and capture the relative impact of different aspects of the student learning experience.

Specifically, this study explored the following research questions:

- RQ1: Can machine learning approaches, in particular, neural networks, predict students' service-learning outcomes from the various student, course and learning experience factors?
- RQ2: What are the effective models to that end?

The contributions of this paper include: 1) we designed a deep neural network with embedding layers and attention modules to model students' learning experiences related to various factors, from which to predict outcomes associated with service-learning; 2) we evaluated the proposed model with extensive experiments on a large dataset to understand its performance; 3) we evaluated the different components of our resulting model to ascertain their relative contributions.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 delineates the context and Section 4 presents the proposed methodology and model. Section 5 presents the experiments and results. This is followed by our findings and discussions in Section 6. Finally, this paper is concluded in Section 7.

2 Literature review

The background and related work of this study involves service-learning, computational analysis for education, and attention mechanisms. We expound on them as follows.

2.1 Service-learning

Service-learning is an experiential pedagogy that combines service to the community with academic learning. In academic service-learning programs, students learn about a societal issue and acquire the skills and knowledge that may be applied to address this issue. They then are arranged to carry out a service project that addresses this issue, and link their service experiences with the academic topic through critical reflection. A number of studies have affirmed the benefits of service-learning to students' development. These include positive effect on students' cognitive and academic outcomes (Lemons et al., 2011; Novak et al., 2007; Prentice & Robinson, 2010); communication and leadership skills (Simons & Cleary, 2006; Wurr & Hamilton, 2012); civic responsibility and engagement (Greenwood, 2015; Weber & Weber, 2010); and personal understanding and growth (Simons & Cleary, 2006; Weiler et al., 2013). These can be roughly categorized into intellectual, social, civic, and personal developments respectively. Fullerton et al. (2015) further suggested that the impact of service-learning persists even after graduation.

Even though *outcomes* of service-learning have been well studied and documented, the *process* of service-learning, in particular, the impact of the different

course, student and learning experience factors on the student learning outcomes, is less well-understood. There are commonly accepted guidelines of "good practices" believed to be impactful in effecting students' learning outcomes, such as the 10 curricular and pedagogical factors from the Wingspread Report (Honnett & Poulsen, 1989). Most of these guidelines, however, are backed up only by anecdotal evidence or long-time practice, or, at best, by small-scale studies.

Mabry (1998) demonstrated that service-learning is more effective when students undertake at least 15-20 hours of service, keep frequent contact with their service beneficiaries, introspect with weekly in-class reflection, maintain ongoing and summative written reflection, and carry out discussions of their service experiences with both instructors and site supervisors. Astin et al. (2000) found that the most important factor associated with a positive service-learning experience is the student's degree of interest in the course matter, followed by class discussion, connecting the service experience to the course matter, and the amount of training received prior to service. Lambright and Lu (2009) identified three key factors: the extent to which the project is integrated with class materials, whether students work in groups, and whether the participating students are full-time. Celio et al. (2011) conducted a meta-analysis of 62 studies on the impact of service-learning on students, and found four key practices: linking to curriculum, youth voice, community involvement, and reflection. Moely and Ilustre (2014) found that the two outcomes that are most closely related to service-learning – learning about the community and academic learning – were strongly predicted by students' perceived value of the service, the opportunities for reflection, and students' social change orientation.

Some other approaches used statistical methods to analyze students' learning from service-learning. Chan et al. (2019) analyzed four dimensions of student learning in mandatory service-learning and their correlation with factors of student background, interest, and learning experience. It was found that among all the factors investigated, the quality of the learning experience was the only consistently significant factor that has an impact upon the student learning outcomes. Among different course, student and learning experience factors, Ngai et al. (2018) found highly statistically significant correlations between factors such as the perceived benefits of the service to the community, challenging and meaningful tasks, students' interest and regular and structured reflection, etc. to the various student learning outcomes, including intellectual, personal, social and civic.

Traditional statistical analysis methods can establish some understanding of the various factors and outcomes, but they face challenges when it comes to dealing with real-world data. To begin with, statistical methods make assumptions about data distribution and/or linearity in variable relationships, but real-world data is often highly interrelated and non-linearly correlated. Another limitation is that the number of variables included in a model cannot be too large, but service-learning (and other experiential learning) often involves large numbers of factors in the student learning experience. In contrast, machine learning algorithms make minimal assumptions about the data and are generally more adept at handling noisy data with potentially complicated interactions between the factors. To bridge the gaps, we proposed a deep neural network model to predict students' service-learning outcomes and examine the impact of different learning experiences in this study.

2.2 Computational analysis for education

Much previous work has demonstrated that machine learning and data mining techniques are effective in modeling complex and noisy data in a wide range of domains, including education. Educational data mining applies machine learning and data mining models to predict students' performance via a variety of features, to understand how students learn and what affects their learning. This understanding facilitates evidence-based improvements in the design and delivering of the learning experiences, such as the design of course elements, teaching strategies, assignments, projects and learning activities, so as to enhance students' learning experiences and improve their learning outcomes. For instance, Minaei-Bidgoli et al. (2003) built classifiers to predict students' final exam performance using features extracted from their action logs in an online education platform. The results revealed that students' homework performance and engagement are the most useful indicators of students' final grade. Tomasevic et al. (2020) further confirmed the effectiveness of the engagement-based and past performance-based features with a large public dataset (Kuzilek et al., 2017). They also suggested that neural networks perform better than other machine learning approaches in these contexts. Romero et al. (2013) showed that certain features of students' participation in online discussion forums, such as the quantity and quality of their posts, and their online social relationship with other students, are also good predictors of their exam performance. Asif et al. (2017) studied students' four-year graduation rates based on their pre-university and first and second year courses grades. Azcona et al. (2019) successfully identified students at risk in computer programming classes by modeling students' demographic information, prior academic performance and behavior logs in online education platform. Bosch (2021) used decision trees to learn student-level variables and mindset interventions that were indicative to students' GPA improvements when they transit from middle school to high school. Other work has also demonstrated that machine learning models are able to understand and identify the course, instructor and learner factors that affect students' selection and completion of online courses (Kardan et al., 2013; Hew et al., 2020).

Most studies in educational data mining are in the context of online education, such as Massive Open Online Courses (MOOCs), as it is relatively easy to obtain large amounts of student data from these contexts. Online systems can also provide large amounts of interaction data such as click streams that can be analyzed to learn how students interact with different learning contents (Geigle & Zhai, 2017). These insights enabled the development of affective-aware adaptive learning environments to create a better learning experience for students based on their interactions and detected affective state (Grawemeyer et al., 2017). In addition, it is feasible to detect students' learning performance by modeling their interaction behaviors with machine learning and data mining approaches, such as the interaction with different course contents during the learning process (Brinton & Chiang, 2015; Chen et al., 2018). Chen et al. (2014) also probed into students' interaction behaviors in assessments. They found that students' gaze behaviors during a test, especially the average duration of eye fixations, could efficiently predict their performance in the

said assessment. Students' gaze interaction can also be analyzed for detecting mind wandering in classrooms (Hutt et al., 2019).

Despite these recent initiatives, there has not been any previous work applying data mining methods to investigate service-learning-related outcomes. To our knowledge, the closest work is that presented in Lo et al. (2019), that used association rules to identify learning experience factors that are impactful to the learning outcomes of students from different disciplines.

2.3 Attention mechanisms

Real applications often involve a rich set of input variables, solving for some target output. When the number of input variables gets too big, challenges for models arise as the search space becomes too wide. This also limits the applicability of complex machine learning techniques (such as neural networks) in learner modeling. On one hand, there is no guidance as to which of the various learning aspects should be included in the models. On the other hand, it is also difficult to interpret the results – i.e., to explain the reasons behind a well-performing model and to identify the contributing factors (Pelánek, 2017). Attention mechanisms have been introduced to allow models to selectively pick out specific elements in the input variables to focus on Bahdanau et al. (2014). In brief, an attention module computes a vector of learned weights to indicate which parts of the data are more important. The computed weights are applied to the input data to differentiate and "highlight" the more important parts for the next layer to "pay attention". In essence, attention mechanisms borrow from the concept of human attention. It is similar to the case when we recognize a person in a photograph. Our the brain will "tell" us to focus on processing the information from the facial region, which correlates to the pixels in the facial region getting higher weights and being highlighted.

Since their introduction, attention mechanisms have been proven to be critical for deep learning in a diversity of problems across many disciplines. One common application of the technique is to find the word(s) in a sentence with the highest match to a particular output word in machine translation (Bahdanau et al., 2014). This can be applied to other human language comprehending studies, such as finding the words that are most indicative to the semantic intent of the whole query sentence for machine dialogue systems (Goo et al., 2018; Liu & Lane, 2016; Liu et al., 2020), and recognizing entities and their potential relations within a sentence (Nayak & Ng, 2019).

Attention mechanisms also perform well in personalized recommendation and diagnosis, since models can tell how likely a factor of an item is attractive to a specific user by modeling the user's demographic information with a factor-based attention layer (Cheng et al., 2018). For example, in news recommendation, an attention module can visit individual words in a news article based on the user's features to estimate whether a user would be interested in that article (Wu et al., 2019). They have also been used for disease prediction (Gao et al., 2019), to measure the relevancy between patients and diseases based on their historical health record. In the area of computer vision, attention modules are also commonly employed in tasks such as recognizing multiple objects (Ba et al., 2014), detecting the saliency map in an image (Kruthiventi

et al., 2017), and automatic image captioning, which connects the areas of computer vision and natural language processing (You et al., 2016).

In addition to learning the relevance of specific input items to the target output, attention mechanisms can also be applied to learn the inner relationships, that is, the relationships between different items in the input data, in a process also known as self-attention. For instance, Cui et al. (2017) utilized a self-attention module to compute the pair-wise matching score between a document word and a query word in cloze-style reading comprehension tasks. Huang et al. (2018) used a feature-wise self-attention module for contextual user behavior modeling.

Inspired by these studies, we experimented with two attention modules in our model. The first one adopted a self-attention mechanism to identify the inner relevance between different course, student and learning experience factors. The second attention module focused on understanding the impact of the different factors on the various learning outcomes. We demonstrated that the attention and the embedded modules could successfully encode latent information in the dataset and contribute to predicting student learning outcomes.

3 Background and data

This study proposed the *Context-Reinforced Experience Attention Modeling* (CREAM) network for predicting learning outcomes from student-acquired or respondent-based data in service-learning courses. Given a set of respondent data covering the student-perceived learning experience and their learning outcomes, our model was designed to carry out the following tasks: (1) identifying and comparing the impact of different learning experience on different service-learning outcomes; and (2) encoding contextual information from students' learning experience that was effective for predicting service-learning outcomes.

3.1 Context

The context of this study is a large, comprehensive, public university in Hong Kong that has incorporated academic service-learning into their undergraduate core curriculum since 2012. All undergraduate students are required to pass a three-credit service-learning course before graduation. There are almost 70 service-learning courses distributed over 33 disciplines, and each includes the following three components:

- An experiential component, in which students participate in a substantive service project to address some societal issue;
- An academic component, in which students are taught the academic concepts pertaining to the aforementioned societal issue, and the skills needed to execute the said service project; and
- A reflective component, that facilitates students to link theory with practice and to develop ethical and social responsibility.

These components are standard and commonly encountered in almost all academic service-learning contexts.

Students' experience and learning outcomes in the service-learning courses are measured via a Student Post-Experience Questionnaire (SPEQ), developed by the university with reference to the literature review and the specific contexts in which the service-learning courses and projects are implemented. It includes the following questions:

- Learning Outcomes (Table 1) : Questions asking students to rate, on a 7-point scale (1 = very little; 4 = a fair amount; 7 = very much), their attainment of the intended learning outcomes relating to their intellectual (four items), social (two items), civic (five items), and personal (one item) learning outcomes.
- Learning Experience (Table 2): Questions inviting students to indicate their experience, on a 7-point scale (1 = strongly disagree, 4 = neutral; 7 = strongly agree), their level of agreement with items on the following:
  - Whether they took the course primarily to fulfill the graduation requirement (1 item).
  - Whether they took the course because they were interested in the service project (1 item).
  - Various aspects of their learning experience of the service learning course/project they had completed (16 items).

The instrument was reviewed by a panel of experienced service-learning teachers and researchers to ensure content and face validity. The construct validity of the multiple-item scales was also examined using exploratory and confirmatory factor analyses. The results show that the instrument is reasonably valid, with all of the fit indices meeting the criteria for goodness of fit (CFI = 0.973, TLI = 0.956, NFI = 0.971, RMSEA = 0.073).

The questionnaire is administered to the students at the end of the course as part of the university's quality assurance process. Students are well-informed of the purpose of the study and assured that their responses will not be made known to the teachers nor affect their grade in the course.

The data that support the findings of this study are available from the institution (Hong Kong Polytechnic University) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. For this study, the authors sought and received approval from the university ethics committee to access this data for analysis. Data are however available from the authors upon reasonable request and with permission of the Hong Kong Polytechnic University.

3.2 Participants

The dataset in this study was collected from undergraduate students enrolled in one of 74 service-learning courses offered between the 2014/15 and 2018/19 academic years at the Hong Kong Polytechnic Unversity. These courses originated from a wide spectrum of disciplines including engineering, health sciences, hard sciences, design,

Table 1 Items from SPEQ measuring students' self-perceived learning outcomes on their intellectual, social, civic, personal and overall development from the service-learning course

lo	Learning Outcomes	Survey Items
lo_1	Intellectual	Deeper understanding of the linkage between service-learning and the academic content of the course. Applying/integrating knowledge to deal with complex issues. Solving challenging real-life problems. Thinking critically.
lo_2	Social	Working effectively in teams. Communicating effectively with peers, collaborators, and service recipients.
lo_3	Civic	Better understanding of the problems facing underprivileged members of the community. Increased interest/commitment to serve people in need. Becoming a more responsible member of your community. Increased understanding and respect for people from different backgrounds. Becoming a more responsible global citizen.
lo_4	Personal	Better understanding of my own strengths and weaknesses.
lo_5	Overall	Overall learning gain.

business, humanities, and social sciences, etc. The service projects also exhibited much diversity across different geographical locations (e.g. China, Cambodia, Myanmar, Rwanda, etc.), for different beneficiaries (e.g. elderly, and children, etc.), and with different types of activities (e.g. teaching/tutoring, construction, design, etc.).

Students were asked to fill out the SPEQ upon course completion. In total, we received 18,175 responses. Responses with one or more missing items were removed from our dataset. After the filtering, our dataset contained 11,185 instances from 11,100 students (5,494 females). 85 students took more than one service-learning course, contributing to more than one data instance for the dataset.

4 Methodology

This study investigated the potential of using a deep learning model to predict students' self-perceived learning outcomes from their self-reported course, student and other learning factors.

We constructed our dataset based on student responses. A student response R^{ks} completed by a student s who enrolled in a specific service-learning course k, is a vector of integers. Specifically, we have $R^{ks} = \langle RO^{ks}, RE^{ks} \rangle$, where $RO^{ks} = \langle ro_{11}, \ldots, ro_{14}, ro_{21}, \ldots, ro_{22}, ro_{31}, \ldots, ro_{35}, ro_{41}, ro_{51} \rangle$, and $RE^{ks} = \langle re_1, re_2, \ldots, re_{18} \rangle$. The ro_{ij} and re_i are integers $\in \{1, 2, \ldots, 7\}$. Each of the ro_{ij} corresponds to one item from Table 1 and each re_i to an item from Table 2. Since we

Table 2 Items from the SPEQ measuring the student learning experience with respect to different course and pedagogical elements of the service-learning course

le	Questionnaire Item
le_1	The main reason for me to take this service-learning course is to fulfill the university's Service-Learning requirement for graduation.
le_2	I took this course because I was very interested in the service-learning project of the course.
le_3	The service I performed was closely related to my chosen major/discipline of study.
le_4	I put a lot of effort into planning, preparing and delivering the service.
le_5	I believe that the service I performed in the service-learning project has benefited the people I served.
le_6	I felt that my service was appreciated by the collaborating agency/service recipients.
le_7	My instructors and TAs prepared me appropriately for performing the service.
le_8	I could feel the enthusiasm and passion of my instructors and TAs in delivering the course and the service.
le_9	Help and support was usually available from the instructors/TAs/collaborative agency when I needed it.
le_{10}	I benefited a lot from the interaction I had with the instructors, TAs and other students in class.
le_{11}	My team-mates in the service-learning project were generally motivated and supportive.
le_{12}	There were a lot of opportunities for me to meet and interact with the people I served.
le_{13}	I developed a good personal relationship with my teammates.
le_{14}	The service-learning project provided challenging and meaningful tasks for me to accomplish.
le_{15}	The service-learning project challenged me to try things that I had never done before.
le_{16}	In my service-learning project, I carried out tasks that were mainly designed by me/my team rather than following instructions.
le_{17}	I was required to engage regularly in reflective activities (e.g. writing reflective journals or project logs, debriefing sessions, project reports) during and after the service-learning project.
le_{18}	The reflective activities of the course were well structured with clear instructions and guidelines.

All items are rated on a scale of 1-7

measure five learning outcomes $\langle lo_1, lo_2, lo_3, lo_4, lo_5 \rangle$, and each lo_i corresponds to one or more items from Table 1, the overall student-perceived learning outcome ro_i for lo_i is the mean of all the ro_{ij} assigned to the items corresponding to lo_i.

We modeled each student response as a *learning profile*, composed of a pair $LP^{ks} = (OP^{ks}, EP^{ks})$:

- $OP^{ks} = \langle (lo_1, ro_1), (lo_2, ro_2), \ldots, (lo_M, ro_M) \rangle$, $ro_i \in \mathbb{R}$ and $M = 5$, are the (learning outcome, rating) pairs corresponding to the student's self-assessed learning gains, where $\mathbb{R}$ denotes the set of real numbers. As illustrated in the above example, each (lo_i, ro_i) is the average rating for all the items related to learning outcome lo_i in Table 1.
- $EP^{ks} = \langle (le_1, re_1), (le_2, re_2), \ldots, (le_N, re_N) \rangle$, $re_i \in \{1, 2, \ldots, 7\}$ and $N = 18$, are the (learning experience item, rating) pairs corresponding to the student's responses to the items in Table 2.

Our dataset could therefore be viewed as a set of *learning profiles*: $LP = \{(OP^{11}, EP^{11}), (OP^{12}, EP^{12}), \ldots, (OP^{ks}, EP^{ks}), \ldots\}$. We illustrate with an example. Student 15 takes service-learning course 17 and completed the SPEQ with the following ratings:

- $RO^{17,15} = \langle 5, 5, 6, 5, 4, 5, 4, 4, 5, 4, 6, 5, 4 \rangle$ for the learning outcome items in Table 1
- $RE^{17,15} = \langle 3, 5, 2, 5, 6, 6, 7, 6, 6, 4, 6, 4, 6, 5, 5, 6, 6 \rangle$ for the learning experience items in Table 2

This means that the student's attainment of the five learning outcomes would be calculated as:

$$\langle (5+5+6+5)/4, (4+5)/2, (4+4+5+4+6)/5, 5, 4 \rangle = \langle 5.25, 4.5, 4.6, 5, 4 \rangle$$

and the student's learning profile would be:

$$\left(OP^{17,15}, EP^{17,15}\right) = (\langle (lo_1, 5.25), (lo_2, 4.5), (lo_3, 4.6), (lo_4, 5), (lo_5, 4) \rangle, \langle (le_1, 3), (le_2, 5), (le_3, 2), \ldots, (le_{18}, 6) \rangle)$$

4.1 Learning outcome prediction from respondent ratings

Given a learning experience profile EP^{ks}, we trained a neural network model to predict the value for a particular learning outcome ro_i. The most straightforward way was to learn directly from the raw numeric values of students' respondent ratings in our dataset. For each learning experience profile EP^{ks}, we extracted the RE^{ks} given to the model as the feature vector. The model then applied dense layers to estimate the final rating of one learning outcome:

$$u_{re} = \sigma\left(W_f \times RE^{ks} + b_f\right) \tag{1}$$

$$\hat{ro} = W_u \times u_{re} + b_u \tag{2}$$

where σ is the activation function, and the weights in the W_f, b_f, W_u, and b_u were trained during the learning process. This process was repeated five times, each time predicting for one of the learning outcomes.

We adopted a 10-fold cross-validation approach for model evaluation. The dataset was segmented into 10 groups, or "folds". The model was then trained and validated on 8 and 1 of the folds respectively, and evaluated on the remaining fold. This process was repeated 10 times, until each of the folds had been used once for evaluation. The performance of the model was taken as the mean (averaged) performance that was

achieved on each of the evaluation folds. Since our dataset included some students who have contributed multiple instances, and there would likely be correlations in data that is contributed by the same individual, we needed to ensure that we would not end up having data points from the same student in the training and the test set. Therefore, we constructed our folds by segmenting our data on the *student* level.

The root mean squared error (RMSE) was used as the loss function to train our model. Given a set of (course, student) pairs $\mathcal{K} = \{(1, 1), \ldots, (k, s), \ldots\}$, the RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{(k,s)\in\mathcal{K}} \left(ro^{ks} - \hat{ro^{ks}}\right)^2}{|\mathcal{K}|}} \tag{3}$$

where ro^{ks} is the student-assigned rating for one learning outcome, and $\hat{ro^{ks}}$ is the predicted rating from the neural network. During the training phase, the trainable values of the model were updated and optimized by minimizing the loss function with backpropagation algorithm (Rumelhart et al., 1986).

The Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 0.001 was adopted for training the model. In order to prevent model overfitting, an early stop strategy was applied in the training process, which terminated the training if the model did not improve on the validation fold over 5 epochs. To minimize the likelihood of overfitting, we also used dropout (Srivastava et al., 2014) with a rate of 0.2.

The averaged RMSE over the 10 folds of the data is shown in Table 3, with the standard deviation of the calculated outcome included for comparison.

The RMSE was lower than the standard deviation, suggesting that the performance of the model was reasonable, but we hypothesized that the current mode of constructing the feature vector u could make assumptions about the data that may be naive. Simply concatenating the respondents' ratings together into one feature vector assumed that the numeric ratings are linear, or, at least, follow a distribution that is easily modeled mathematically. However, previous work did not support this. For example, Nadler et al. (2015) and Velez and Ashworth (2007) found the midpoint in a rating scale often did not truly express a neutral opinion. Instead, survey respondents often chose the midpoint in a rating scale to avoid making a choice, or to save cognitive energy. It stands to reason that the relationship between the successive rating values is most likely more complex than is apparent from the raw numeric values.

In contrast to many other commonly-encountered problems (e.g. face recognition, speech processing, etc) tackled by machine learning, our data was very inconsistent and subjective. It was not difficult to imagine respondents carelessly or mindlessly filling out the questionnaire. The value of the ratings also differed across respondents

Table 3 Predicting Service-Learning outcomes: baseline performance

Learning Outcome	Intellectual	Social	Civic	Personal	Overall
RMSE	0.624	0.671	0.616	0.774	0.702
Standard deviation	0.884	0.930	0.888	1.002	0.985

– "somewhat agree" and "agree" could mean very different things to different people. Even for the same respondent, there could be shifts in his/her perception of a particular rating across different types of learning experience. For example, rating an "agree" for le_1 (motivation for taking the course) could be different from rating an "agree" for le_4 (put a lot of effort).

An added complication arose from the fact that some of the items in the questionnaire were linked. For example, the items le_5 (the service was beneficial to the community) and le_6 (the service was appreciated) were obviously correlated. Given a different context, the same numeric rating, given to the same item, could have a very different meaning. For example, supposing student s_1 gave a rating of 4 to item le_i, and ratings of 3 for other items that are related to le_i. Another student, s_2, gave the same rating 4 to item le_i but 5 to all the other related items. Even though s_1 and s_2 had both given item i the same rating, it is likely that s_1 had a more negative perception about that learning experience than s_2, given his/her rating on other highly correlated items. These observations suggested that the model should be trained to learn these latent relationships in the data.

4.2 Context-reinforced experience attention modeling network (CREAM)

From our observations, it appeared that at least three types of latent information should be encoded: (1) the relationships between the different components of the dataset – i.e. the items on the questionnaire and the respondent ratings; (2) the correspondence between different learning experience items within the same learning experience profile – i.e. between the different le_i; (3) the linkage between the different learning experience items and the target learning outcome – i.e. between the various le_i and the target lo_i.

We thus designed Context-Reinforced Experience Attention Modeling Network (CREAM) for learning outcome prediction with a pipeline that incorporates three modules. Each module focused on extracting one type of latent information from the data. The different components in the process are described as follows.

4.2.1 Stage 1. EMBED: Mapping questionnaire items and responses to feature space

The first module in CREAM implemented embedding layers (EMBED) to map the data from the questionnaire into feature space. Like some other problems such as automatic captioning of images (Frome et al., 2013), our dataset contained heterogeneous data. Each learning experience pair (le_i, re_i) was composed of le_i, the experience, dimension or characteristic measured by the survey item, and re_i, a number from a set of numeric values denoting "a certain degree". le_i and re_i thus constitute different natural and physical meanings.

Our preliminary result suggested that the respondent ratings in our dataset, though represented by numbers, could be more similar in behavior to categorical types rather than numeric values. Indeed, even though we *encode* the ratings with integers, they were *presented* to the respondents in the survey as descriptions, e.g. "disagree", "somewhat disagree", etc.

Prior studies (Wang et al., 2021) have demonstrated the potential of using embedding techniques to learn latent features of ratings, and that rating embedding could effectively encode useful information from ratings, especially when they interacted with heterogeneous data. We thus applied two embedding layers to learn the latent features from the ratings and the questions respectively and mapped them to feature space. As depicted in Fig. 1, the rating embedding layer converted each numeric rating value re_i to a dense vector $h_i^r \in \mathbb{R}^{d_r}$, where d_r is the dimension of the rating latent features; and the question embedding layer converted each learning experience question item le_i to a dense vector with a dimension of d_q: $h_i^q \in \mathbb{R}^{d_q}$. This is similar to work done in recommender systems (Covington et al., 2016; Huang et al., 2018; Wang et al., 2018; Yan et al., 2019), which often incorporated the IDs of the goods or the videos along with the review ratings and other demographic information.

4.2.2 Stage 2. CLEMM: Modeling the learning experience context

The second stage of our model is targeted to understand and quantify the interdependence between the different items that measured student experience on the post-experience survey.

Previous work commonly used multiple regression to analyze questionnaire data, including those relating to learning experiences and learning outcomes. These studies usually worked with the raw respondent ratings, or, in the case of multi-item scales, the mean of the rating values. This can be problematic as the ratings are simplified to put them on a ratio scale, or, at least, a simple mathematical relationship between the successive rating values is assumed. Neural networks offer an alternative approach. Given a particular data item i, it is possible to learn the interdependence between i

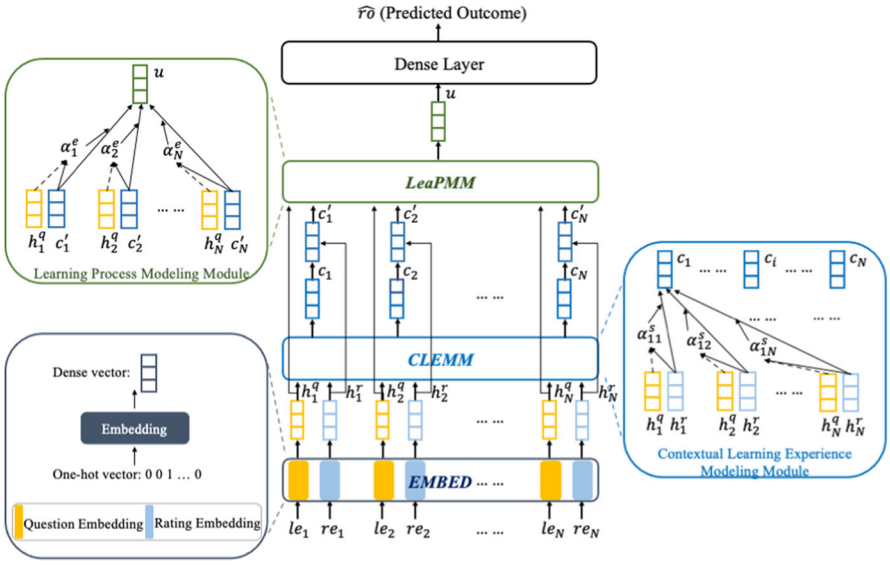

Fig. 1 Framework of the proposed CREAM model for service learning outcome prediction

and other data items directly from the data. In essence, it is possible to pick out the more relevant or salient parts of the input data, relevant to item i, and assign them higher weights.

Attention mechanisms (Huang et al., 2018) have been successfully used to consolidate and encode latent contextual information from input data. In the same spirit, we implemented the *Contextual Learning Experience Modeling Module (CLEMM)* in our model based on a *self-attention mechanism* to calculate the interdependence between the learning experience items – that is, to encode the *inner relationship* between the input elements. This inner relationship is the latent information that can be extracted based on the context of "the student rated re_i for le_i when she rated $\{re_j\}$ for other related items le_j".

Figure 1 illustrates the process. The dense vectors $h_i = \left[h_i^q, h_i^r\right]$, for all the items i in SPEQ that were constructed by EMBED are passed as input to CLEMM. CLEMM then quantified the relationship for each pair of experience items (i, j) based on the interaction of their latent features (h'_i, h'_j) as follows:

$$h'_i = \sigma\left(W_h \times h_i + b_h\right) \tag{4}$$

$$v_{ij}^s = {h'_i}^T W_s h'_j \tag{5}$$

$$\alpha_{ij}^s = \frac{exp(v_{ij}^s)}{\sum_{k=1}^{N} exp(v_{ik}^s)} \tag{6}$$

where σ is the activation function, and $W_h \in \mathbb{R}^{d_h \times (d_q + d_r)}$, $b_h \in \mathbb{R}^{d_h}$ and $W_s \in \mathbb{R}^{d_h \times d_h}$ are the trainable model parameters. Based on this, we then compute the contextual rating feature c_i for experience item i as follows:

$$c_i = \sum_{j=1}^{N} \alpha_{ij}^s h_j^r \tag{7}$$

The output of CLEMM was the contextual feature vector $c'_i = \left[h_i^r, c_i\right]$, $c'_i \in \mathbb{R}^{2 \times d_r}$, the concatenation of c_i and h_i^r. This process was repeated for all the experience items to extract $\left\{c'_1, c'_2, \ldots, c'_N\right\}$.

4.2.3 Stage 3. LeaPMM: Modeling the learning process

We focused on the interdependence between the different parts of the learning process and the learning outcomes in the next stage. Service-learning involves multiple types of pedagogical modes and learning experiences, including classroom learning, project execution and reflecting on the experience. It is reasonable to expect that different types of learning experiences will impact the learning outcomes differently. For example, a good personal relationship with teammates positively correlates with learning outcomes rated to social development (Ngai et al., 2018).

In Stage 3 of our model, we implemented the *Learning Process Modeling Module (LeaPMM)*, an attention module, to focus on the interdependence between each learning experience item and the learning outcome to be predicted (Fig. 1). The inputs to LeaPMM were the contextual feature vectors $\{c'_1, c'_2, \ldots, c'_N\}$ from CLEMM and the

dense vector embedding for each learning experience item $h_i^q, i \in N$ from EMBED. For the i-th experience item, we calculate its *impact weight* on target service-learning outcome as follows:

$$h_i^c = \sigma\left(W_c \times \left[h_i^q, c_i'\right] + b_c\right) \tag{8}$$

$$v_i^e = W_e h_i^c \tag{9}$$

$$\alpha_i^e = \frac{exp\left(v_i^e\right)}{\sum_{j=1}^{N} exp\left(v_j^e\right)} \tag{10}$$

where $W_c \in \mathbb{R}^{d_c \times (d_q + 2 \times d_r)}$, $b_c \in \mathbb{R}^{d_c}$ and $W_e \in \mathbb{R}^{1 \times d_c}$ were the trainable model parameters, and $[\cdot]$ denoted the concatenation of different feature vectors. This was repeated for all N experience items. The final feature vector (u) is then the summation of all the contextual rating features of all the learning experiences weighted by their impact weights. It thus encoded the contextual information of the entire learning experience:

$$u = \sum_{i=1}^{N} \alpha_i^e c_i' \tag{11}$$

4.2.4 Making the final prediction

Given the final feature vector u, a dense layer is used to estimate the final rating of one learning outcome:

$$\hat{ro} = W_u \times u + b_u \tag{12}$$

As depicted in Section 4.1, W_u and b_u were trainable model parameters in the final dense layer.

We adopted RMSE as the loss function, and adopted the same training setting as presented in Section 4.1 to train our models.

5 Experiments and results

We tested our designed model from two aspects: (1) the efficacy of the model in predicting students' self-perceived learning outcomes given their learning experience; and (2) the construction of the model framework, i.e. the contributions of the individual components. We evaluated the models with 10-fold cross-validation and RMSE as the metric.

Table 4 summarizes the results (averaged RMSE, lowest is best) of our ablation study. We evaluated the different models constructed by leaving out various components of our framework The first row shows the model previously illustrated in Section 4.1, leaving out all 3 modules and working directly on the raw numeric values. These models were all evaluated on the intellectual, social, civic, personal and overall service-learning outcomes respectively.

Our first observation was that the models using EMBED consistently outperformed the models without that across the board. One possible reason was that it was easier to handle heterogeneous data in the form of latent features learned

Table 4 Performance of CREAM on predicting service-learning outcomes (RMSE). Best performance for each outcome shown in bold

Modules Incorporated			RMSE for Outcome				
EMBED	CLEMM	LeaPMM	Intellectual	Social	Civic	Personal	Overall
–	–	–	0.624	0.671	0.616	0.774	0.702
✓	–	–	0.607	0.656	0.599	0.764	0.692
–	✓	–	0.602	0.653	0.594	0.759	0.694
–	–	✓	0.587	0.647	0.608	0.761	0.696
✓	✓	–	0.579	0.633	**0.573**	0.745	0.671
✓	–	✓	**0.577**	0.632	**0.573**	0.745	0.670
–	✓	✓	0.583	0.636	0.575	0.746	0.672
✓	✓	✓	0.578	**0.631**	**0.573**	**0.743**	**0.667**

from embedding layers. Another possible reason could be that the rating latent features could encode more useful information and expose more complex relations of different ratings than the raw numeric rating values.

Our second observation is that the two attention modules, working on their own, can also improve the prediction performance. LeaPMM, in particular, appeared to contribute more to performance improvement. This corroborated previous work in demonstrating that different experiences impact learning outcomes differently, and that the learned experience-impact attention weight was effective at helping the model to focus on the experience items that have greater influence on the learning outcomes. The models incorporating LeaPMM thus are less likely to be distracted by experience items that have low impact on the learning outcomes.

Thirdly, when CREAM incorporated all the modules, it consistently outperformed the alternative models over almost all the learning outcomes. In this composition, CREAM achieved RMSE of 0.578, 0.631, 0.573, 0.743 and 0.667 for predicting intellectual, social, civic, personal and overall learning outcomes respectively. In particular, our model performed much better in predicting intellectual and civic learning outcomes than the others. This can be explained buy the fact that there are 4 and 5 items related to intellectual and civic learning outcomes respectively, meaning that more comprehensive information was available for these two learning outcomes, hence making them easier to predict than the others. Nevertheless, even for the other learning outcomes, our model also attained promising performance.

Finally, there is the question of whether the model achieves a "good enough" performance. One metric, proposed by model evaluation literature to determine if the RMSE is low enough, is the RMSE-observations SD Ratio (RSR), calculated as the ratio of RMSE to the standard deviation of the measured data: $RSR = \frac{RMSE}{SD}$ (Moriasi et al., 2007; Singh et al., 2005). Generally, $RSR \leq 0.7$ indicates a *satisfactory* performance of the model. To meet this condition, our model should yield RMSE of no more than 0.7, given that the standard deviation of the student ratings is about 1 in our dataset. It is encouraging to note that with all three modules incorporated into CREAM, this requirement was achieved for most of the learning outcomes.

Besides RMSE, the satisfactory rate (SR) achieved by the models could also help to further understand their performance. We define a prediction as *satisfactory* if the square of the error between the predicted and actual value is below 0.5 (which meets the *satisfactory* guideline of $RMSE \leq 0.7$ in our task). The SR, therefore, measures the proportion of the *satisfactory* predictions made by a model.

Table 5 presents the model performance in SR (highest is best). In general, all three modules contributed to performance improvement in terms of SR. With all three modules incorporated, CREAM again achieves the best performance across the board. This promising result affirmed that the design of our framework is sound, as our model is able to precisely predict learning outcomes for most of the data, especially for the intellectual and civic learning outcomes, in which more than 82% of the data was satisfactorily predicted.

6 Discussion

In this paper, we investigated the possibility of modelling and predicting students' learning outcomes given their learning experience in service-learning. Based on our understanding of the different learning processes, we designed and implemented a model, CREAM, which yields better prediction accuracy than its counterparts. This has thus addressed our research questions, and raised a follow-up to RQ2: *what is the most effective module that was incorporated into the model?* In this section, we discuss the contributions of different components of our model and our findings.

Table 5 Performance of CREAM on predicting service-learning outcomes (SR). Best performance for each learning outcome shown in bold

Modules Incorporated			SR for Outcome				
EMBED	CLEMM	LeaPMM	Intellectual	Social	Civic	Personal	Overall
–	–	–	76.88%	70.50%	77.67%	60.07%	63.60%
✓	–	–	78.24%	75.96%	80.43%	64.61%	69.98%
–	✓	–	80.32%	76.67%	81.22%	69.09%	73.71%
–	–	✓	81.38%	76.79%	80.32%	69.24%	73.62%
✓	✓	–	81.99%	**77.62%**	82.45%	69.67%	74.84%
✓	–	✓	82.07%	77.49%	82.61%	69.44%	74.31%
–	✓	✓	81.62%	77.33%	82.43%	69.56%	**75.05%**
✓	✓	✓	**82.15%**	77.49%	**82.77%**	**69.82%**	74.33%

Table 6 Experiment setup to investigate contributions of different modules

Target	Base Model	Modules Incorporated	
		Base Model	Target Model
CLEMM	leave-all-out	–	CLEMM
	leave-two-out	LeaPMM	LeaPMM, CLEMM
		EMBED	EMBED, CLEMM
	leave-one-out	EMBED, LeaPMM	EMBED, LeaPMM, CLEMM

6.1 Contributions of different modules

In order to fully understand the contributions of the 3 modules in CREAM (shown in Fig. 1), we further analyzed their contributions to performance improvement. The following definitions are used for clarity:

- Target: the module that is under investigation.
- Base model: the model before the *target module* is added. It can be:
  - Leave-all-out: the model that leaves out all 3 modules.
  - Leave-two-out: the model that leaves out the *target* and one other module.
  - Leave-one-out: the model that leaves out the *target* only.
- Target model: the model that adds the *target* to the *base model*.

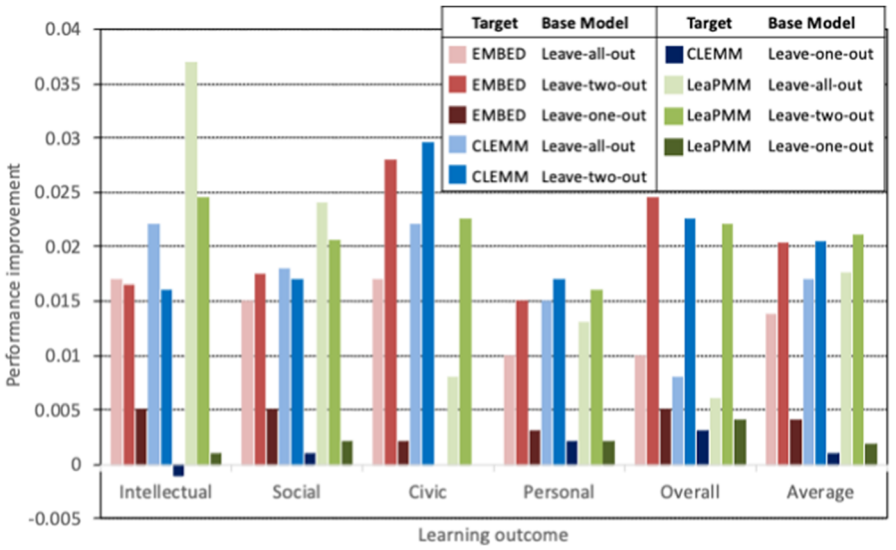

Fig. 2 Performance improvement (in RMSE) contributed by different target modules over a base model

- Performance improvement: the change in the RMSE/SR achieved by the *target model* over the *base model*.

For a systematic investigation, we constructed and tested all possible (*target*, *base model*) pairs and compared the *performance improvement* for each pair. Thus, each *target* was compared against one *leave-all-out* model, two possible *leave-two-out* models and one *leave-two-out* model. Table 6 illustrates an example:

Figures 2 and 3 show the resulting *performance improvement* (the higher the better) in RMSE and SR respectively. Since there were two *leave-two-out* models, their performance improvements were averaged to obtain an overall *leave-two-out* picture. In general, all the modules contributed to a positive *performance improvement*, but the improvement was especially significant when the *base model* was less complex (i.e. we see bigger performance improvements over *leave-all-out* or *leave-two-out* models). In particular, CLEMM and LeaPMM made larger contributions to the less complex models across the board, lowering the RMSE with a reasonable increase in the number of satisfactory predictions. The two modules modeled the interdependence between different learning experiences, and the impact of each learning experience on the learning outcomes. The results suggested that these dependencies are key to student learning experience modeling and learning outcome prediction, and that our proposed attention modules were able to capture them to some extent.

The improvement of adding LeaPMM to a model that has already included two of the previous modules (i.e. a *leave-one-out* model) is marginal, according to the figures. This is particularly true when the *target* is CLEMM or LeaPMM. We hypothesize that, since both of these modules model the learning experience, even when one of them was omitted, the remaining one could serve as a surrogate for it. For example, when LeaPMM captured the impact of different learning experiences on

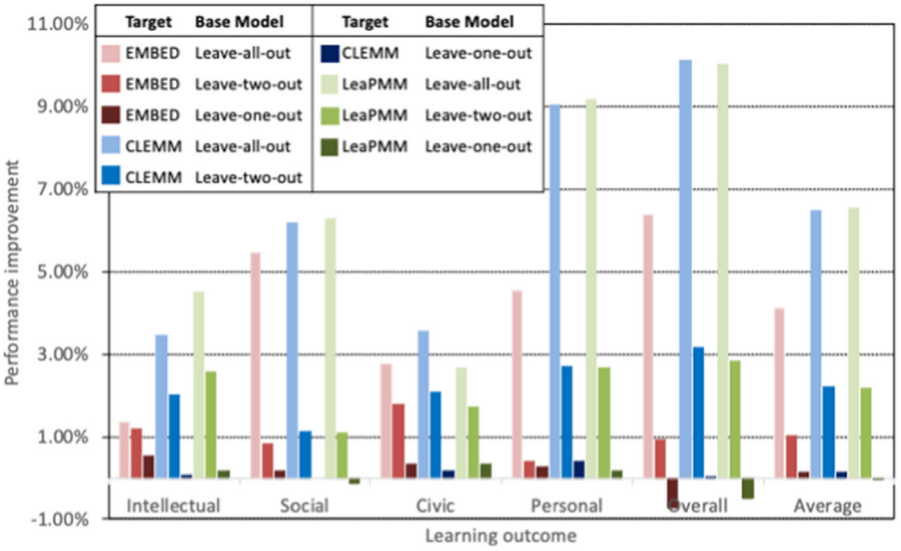

Fig. 3 Performance improvement (in SR) contributed by different target modules over a base model

learning outcomes, it would also indirectly capture information on the inner relations between different learning experiences, since experiences that were strongly related might have similar impact on a learning outcome.

In contrast, in the *leave-one-out* cases, EMBED generally contributes more than the other two modules. This indicates that EMBED was also an important component, even though Tables 4 and 5 suggested that its individual contribution was not as high as the other two modules. This module learned the latent features of ratings, i.e., it modeled the students' interpretation of the questions and their responses. It helped the other two modules to better model the student responses, and thereby to capture the students' learning experience.

Similar to many previous studies (Kardan et al., 2013; Paechter et al., 2010), we modeled students' learning experiences based on self-reporting, i.e. via questionnaire responses. Our findings suggested that in these contexts, an effective model should explicitly include a module for modeling respondent ratings and at least one more attention-based module focusing on modeling the learning experience. We believe that the findings would benefit future studies about student modeling in different learning activities.

6.2 Studying the student responses

Our analyses showed that EMBED, the question and rating embedding module, which mapped the student responses to points in vector space, had a large impact upon the performance of the system. Since all of our input data (indeed, all data in respondent surveys) were captured by this approach, it behooves us to study the rating embeddings more closely. Fundamentally, we aimed to understand: *even though the respondent ratings are commonly represented by numbers, do they truly behave in a numeric manner?* In other words, do the students interpret the ratings in a numeric manner?

The dense feature vectors h_i^r constructed by EMBED enabled us to analyze the ratings as points in multi-dimensional space. We constructed a *ratings distance matrix* based on these learned rating feature vectors. Given a pair of ratings (re_i, re_j), we computed their distance as:

$$dist_{re_i, re_j} = Distance\left(h_i^r, h_j^r\right) \tag{13}$$

where (h_i^r, h_j^r) are the rating latent features extracted from a trained CREAM model, and $Distance(\cdot)$ denotes the distance between two feature vectors. We computed the distance for all rating pairs in the set of $\{1, 2, ..., 7\}$ for one model. As we had different models for predicting different service-learning outcomes, we summarized the average distance for each pair of ratings across all the models.

Figure 4 presents the distance matrix between all rating pairs. Given a pair of ratings, the distance between them, as learned by the CREAM model, is represented as the intersection of the column and row corresponding to the ratings in question.

An inspection of the distance matrix showed that the relationship between the ratings, as incorporated into the rating latent features, was somewhat consistent with

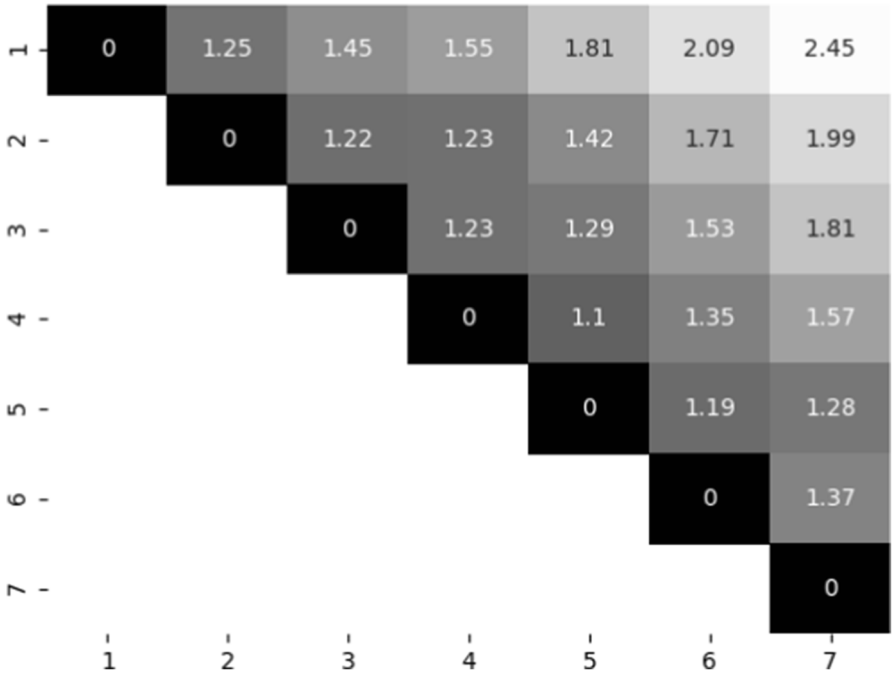

Fig. 4 The distance matrix of different ratings based on the learned latent rating features

what we would expect from the raw numeric values. The ratings increment monotonically – rating "1" was closer to rating "2" than to the other ratings. The distances were also symmetrical – distance(2,3) equaled distance(3,2). However, the rating latent features also suggested that the ratings did not behave like numbers. From the distance matrix, it can be seen that rating "2" was closer to "3" than to "1". Rating "6" also had a larger distance to "7" than to "5". Summation also did not work in the way that we would expect from numbers: distance(2,3) + distance(3,4) did not equal distance(2,4).

On reflection, this is not altogether surprising when one considers how respondents usually behave when filling in such questionnaires. The extreme end-points ("1" and "7" on a 7-point scale, i.e., *strongly dis/agree*) are often selected only when respondents have strong feelings, and not used for incremental differences in opinion. It is also common experience that incremental differences in opinion are difficult to quantify – for example, many people may not make much distinction between a "2" and a "3" on a 7-point scale.

Finally, the results also revealed that the mid-point rating "4" (*neutral*) actually may not necessarily represent a truly neutral response, but tilted slightly towards the positive rating "5" (*somewhat agree*). The finding was quite consistent with those exposed in previous studies, which found that survey participants often choose the midpoint in a rating scale to avoid making choice and to save cognitive energy, etc. instead of truly expressing a neutral opinion (Nadler et al., 2015; Velez & Ashworth, 2007).

7 Conclusions, limitations and future work

This paper presents the novel *Context-Reinforced Experience Attention Modeling* (CREAM) model for service-learning outcome prediction. The model used a neural network with attention mechanisms to model students' experience with respect to various student, course and learning factors, and eventually predict students' service-learning outcomes that are related to their intellectual, social, civic, personal and overall development. Our model achieved promising performance, especially in predicting the intellectual and civic learning outcomes. Our evaluation results also demonstrated the effectiveness of the proposed embedding and attention modules. This suggests that the learned weights in the attention modules are able to capture information about the interdependence between different learning experiences, and the interactions between the different aspects of the student learning experience and the learning outcomes.

Given the prediction results, and the analysis of the effectiveness of the different models, a natural follow-up question would investigate the particular student, course and learning factors which more significantly impact students' learning outcomes. This investigation would contribute to a better understanding of students' learning process in service-learning and offer insights for teaching and learning practices, and this will be part of our focus in the future. We also plan to conduct a thorough analysis of the trained attention weights to better understand the process of learning in service-learning, and also extrapolate our findings to student modeling in other kinds of learning activities.

The limitations of this work are mainly due to data constraints. It is difficult to collect a large volume of data with enormous diversity in practical life. In our study, we tried to make our dataset as diverse as possible: we collected data from a large number of students from diverse departments and across different academic years; we included courses originating from various disciplines; and we involved a diversity of service projects. However, our dataset, similar to many other datasets in such contexts, will eventually be constrained by the institution diversity. All the students were enrolled in, and all the courses are offered by the same university. Hence, the ethnic backgrounds (university strategies promoting internationalisation notwithstanding) and the age of the students would be very similar. In the future, we will investigate possibilities of enlarging our dataset by involving more institutions from different countries and regions.

Declarations

Conflict of Interests We have no known conflict of interest to disclose.

References

Asif, R., Merceron, A., Ali, S. A., & Haider, N.G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177–194.

Astin, A. W., Vogelgesang, L. J., Ikeda, E. K., & Yee, J.A. (2000). How service learning affects students. *Higher Education*, 144.

Azcona, D., Hsiao, I. H., & Smeaton, A.F. (2019). Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. *User Modeling and User-Adapted Interaction*, *29*(4), 759–788.

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. arXiv:14127755.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv:14090473.

Bosch, N. (2021). Identifying supportive student factors for mindset interventions: A two-model machine learning approach. *Computers & Education*, 104190.

Brinton, C. G., & Chiang, M. (2015). MOOC performance prediction via clickstream data and social learning networks. In *2015 IEEE conference on computer communications (INFOCOM)* (pp. 2299–2307). IEEE.

Celio, C. I., Durlak, J., & Dymnicki, A. (2011). A meta-analysis of the impact of service-learning on students. *Journal of Experiential Education*, *34*(2), 164–181.

Chan, S. C., Ngai, G., & Kp, Kwan (2019). Mandatory service learning at university: Do less-inclined students learn from it? *Active Learning in Higher Education*, *20*(3), 189–202.

Chen, S. C., She, H. C., Chuang, M. H., Wu, J. Y., Tsai, J. L., & Jung, T.P. (2014). Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities. *Computers & Education*, *74*, 61–72.

Chen, W., Brinton, C. G., Cao, D., Mason-Singh, A., Lu, C., & Chiang, M. (2018). Early detection prediction of learning outcomes in online short-courses via learning behaviors. *IEEE Transactions on Learning Technologies*, *12*(1), 44–58.

Cheng, Z., Ding, Y., He, X., Zhu, L., Song, X., & Kankanhalli, M.S. (2018). Aˆ 3ncf: An adaptive aspect attention model for rating prediction. In *IJCAI* (pp. 3748–3754).

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 191–198).

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2017). Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Vol. 1: Long Papers)* (pp. 593–602).

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26.

Fullerton, A., Reitenauer, V. L., & Kerrigan, S.M. (2015). A grateful recollecting: A qualitative study of the long-term impact of service-learning on graduates. *Journal of Higher Education Outreach and Engagement*, *19*(2), 65–92.

Gao, J., Wang, X., Wang, Y., Yang, Z., Gao, J., Wang, J., Tang, W., & Xie, X. (2019). Camp: Co-attention memory networks for diagnosis prediction in healthcare. In *2019 IEEE international conference on data mining (ICDM)* (pp. 1036–1041). IEEE.

Geigle, C., & Zhai, C. (2017). Modeling MOOC student behavior with two-layer hidden Markov models. In *Proceedings of the 4th (2017) ACM conference on learning@ scale* (pp. 205–208).

Goo, C. W., Gao, G., Hsu, Y. K., Huo, C. L., Chen, T. C., Hsu, K. W., & Chen, Y.N. (2018). Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 2 (Short Papers), pp 753–757).

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., & Rummel, N. (2017). Affective learning: Improving engagement and enhancing learning with affect-aware feedback. *User Modeling and User-Adapted Interaction*, *27*(1), 119–158.

Greenwood, D. A. (2015). Outcomes of an academic service-learning project on four urban community colleges. *Journal of Education and Training Studies*, *3*(3), 61–71.
Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*, 103724.
Honnett, E. P., & Poulsen, S. J. (1989). Principals of good practice for combining service and learning. Guides 27.
Huang, X., Qian, S., Fang, Q., Sang, J., & Xu, C. (2018). CSAN: Contextual self-attention network for user sequential recommendation. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 447–455).
Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., & D'Mello, S.K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, *29*(4), 821–867.
Kardan, A. A., Sadeghi, H., Ghidary, S. S., & Sani, M.R.F. (2013). Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, *65*, 1–11.
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:14126980.
Kruthiventi, S. S., Ayush, K., & Babu, R.V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, *26*(9), 4446–4456.
Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific data*, *4*, 170171.
Lambright, K. T., & Lu, Y. (2009). What impacts the learning in service learning? an examination of project structure and student characteristics. *Journal of Public Affairs Education*, *15*(4), 425–444.
Lemons, G., Carberry, A., Swan, C., & Jarvin, L. (2011). The effects of service-based learning on metacognitive strategies during an engineering design task. *International Journal for Service Learning in Engineering Humanitarian Engineering and Social Entrepreneurship*, *6*(2), 1–18.
Liu, B., & Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. arXiv:160901454.
Liu, Z., Winata, G. I., Lin, Z., Xu, P., & Fung, P. (2020). Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI conference on artificial intelligence*, (Vol. 34 pp. 8433–8440).
Lo, K. W. K., Ngai, G., Chan, S. C. F., & Kp, Kwan (2019). A computational approach to analyzing associations between students' learning gains and learning experience in service-learning. In *International association for research on service-learning and community engagement (IARSLCE)*.
Mabry, J. B. (1998). Pedagogical variations in service-learning and student outcomes: How time, contact, and reflection matter. *Michigan Journal of Community Service Learning*, *5*(1), 32–47.
Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W.F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education, 2003. FIE 2003*, (Vol. 1 pp. T2A–13). IEEE.
Moely, B. E., & Ilustre, V. (2014). The impact of service-learning course characteristics on university students' learning outcomes. *Michigan Journal of Community Service Learning*, *21*(1), 5–16.
Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T.L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900.
Nadler, J. T., Weston, R., & Voyles, E.C. (2015). Stuck in the middle: the use and interpretation of midpoints in items on questionnaires. *The Journal of General Psychology*, *142*(2), 71–89.
Nayak, T., & Ng, H.T. (2019). Effective attention modeling for neural relation extraction. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)* (pp. 603–612).
Ngai, G., Chan, S. C., & Kwan, K.P. (2018). Challenge, meaning and preparation: Critical success factors influencing student learning outcomes from service-learning. *Journal of Higher Education Outreach and Engagement*, *22*(4), 55–80.
Novak, J. M., Markey, V., & Allen, M. (2007). Evaluating cognitive outcomes of service learning in higher education: A meta-analysis. *Communication Research Reports*, *24*(2), 149–157.
Paechter, M., Maier, B., & Macher, D. (2010). Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction. *Computers & Education*, *54*(1), 222–229.
Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3), 313–350.

Prentice, M., & Robinson, G. (2010). Improving student learning outcomes with service learning. *Community College Journal*, *10*(2), 1–16.
Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458–472.
Rumelhart, D. E., Hinton, G. E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.
Simons, L., & Cleary, B. (2006). The influence of service learning on students' personal and social development. *College Teaching*, *54*(4), 307–319.
Singh, J., Knapp, H. V., Arnold, J., & Demissie, M. (2005). Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. *JAWRA Journal of the American Water Resources Association*, *41*(2), 343–360.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.
Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, *143*, 103676.
Velez, P., & Ashworth, S. D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. Survey Research Methods, (Vol. 1 pp. 69–74).
Wang, X., He, X., Feng, F., Nie, L., & Chua, TS (2018). TEM: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 World Wide Web conference* (pp. 1543–1552).
Wang, Z., Xia, H., Chen, S., & Chun, G. (2021). Joint representation learning with ratings and reviews for recommendation. *Neurocomputing*, *425*, 181–190.
Weber, J. E., & Weber, P. S. (2010). Service-learning: An empirical analysis of the impact of service-learning on civic mindedness. *Journal of Business, Society and Government, Spring*, pp 79–94.
Weiler, L., Haddock, S., Zimmerman, T. S., Krafchick, J., Henry, K., & Rudisill, S. (2013). Benefits derived by college students from mentoring at-risk youth in a service-learning course. *American Journal of Community Psychology*, *52*(3-4), 236–248.
Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X (2019). NPA: Neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2576–2584).
Wurr, A. J., & Hamilton, C. H. (2012). Leadership development in service-learning: An exploratory investigation. *Journal of Higher Education Outreach and Engagement*, 213–240.
Yan, W., Wang, D., Cao, M., & Liu, J. (2019). Deep auto encoder model with convolutional text networks for video recommendation. *IEEE Access*, *7*, 40333–40346.
Yorio, P. L., & Ye, F. (2012). A meta-analysis on the effects of service-learning on the social, personal, and cognitive outcomes of learning. *Academy of Management Learning & Education*, *11*(1), 9–27.
You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4651–4659).

Affiliations

Eugene Yujun Fu[1,2] · Grace Ngai[1,3] · Hong Va Leong[1] · Stephen C.F. Chan[3] · Daniel T.L. Shek[4]

1 Department of Computing, Hong Kong Polytechnic University, Hong Kong, China

2 Department of Rehabilitation Sciences, Hong Kong Polytechnic University, Hong Kong, China

3 Service-Learning and Leadership Office, Hong Kong Polytechnic University, Kowloon, Hong Kong, China

4 Department of Applied Social Science, Hong Kong Polytechnic University, Hung Hom, Hong Kong, China