Introduction to Neural Networks and Uses in EDM

Agathe Merceron
Berlin University of Applied Sciences
merceron@bht–berlin.de

Ange Tato
École de Technologie Supérieure
ange-adrienne.nyamen-tato@etsmtl.ca

ABSTRACT

In this **half-day** tutorial, participants first explore the fundamentals of feed-forward neural networks, such as the backpropagation mechanism; the subsequent introduction to the more complex Long Short Term Memory neural networks builds on this knowledge. The tutorial also covers the basics of the attention mechanism, the Transformer neural networks, and their application in education with Deep Knowledge Tracing. There will be some hands-on applications on open educational datasets. The participants should leave the tutorial with the ability to use neural networks in their research. A laptop capable of installing and running Python and the Keras library is required for full participation in this half-day tutorial.

Keywords

Neurons, Neural networks, LSTM, Attention mechanism, Transformers

1. INTRODUCTION

Neural networks (NN) are as old as the relatively young history of computer science: McCullogh and Pitts already proposed nets of abstract neurons in 1943 as Haigh and Priestley report in [7]. However, their successful use, especially in the form of convolutional neural networks (CNN), Long Short Term Memory (LSTM), or Transformer neural networks, in areas such as image recognition, language translation, or chatbot in the last years has made them widely known, also in the Educational Data Mining (EDM) community. This is reflected in the contributions that are published each year in the proceedings of the conference.

In [11], we counted the percentage of the contributions in the EDM proceedings of the Educational Data Mining (EDM) conference from the beginning of the conference in 2008 till 2019 (long and short papers, posters and demos, young research track, doctoral consortium, and papers of the industry track) that have used some kind of neural networks in

their research. While the percentage stayed below 10% till 2015, it started to increase in 2016 to reach 28% in 2019. This trend has continued since then with 14 long papers from 26 mentioning some kind of neural networks in their research in the EDM proceedings of 2022.

Recognizing the growing importance of neural networks in the EDM community, this tutorial aims to provide 1) an introduction to neural networks in general and to LSTM neural networks with a focus on the attention mechanism and the Transformer neural networks and 2) a discussion venue on these exciting techniques. Compared with our precedent tutorial [11], the main difference is the introduction to Transformer neural networks. This tutorial targets 1) participants who have no or very little prior knowledge about neural networks and would like to use them in their future work or would like to better understand the work of others, and 2) participants interested in exchanging and discussing their experience with the use of neural networks. A simple kind of neural network is a feedforward neural network also often called a multilayer perceptron. It propagates the calculation of each neuron from its inputs through all layers in a directed way forward to its outputs. In education, such a NN has been used, for example, to predict the performance of students. The work of Romero et al. [18] presented at the first EDM conference in 2008 uses it to predict the final mark of students in a course taught with the support of the learning platform Moodle. while the work of Wagner et al. [24] uses it to predict whether students will drop out of a study program.

While their primary use was in Natural Language Processing (NLP) Tasks, LSTM neural networks have been extensively used in education and have achieved remarkable results [22, 20, 6]. Unlike feedforward neural networks that cannot remember the past, LSTM have cycles and are recurrent neural networks. The LSTM [9] architecture can learn long-term dependencies using a memory cell that can preserve states over long periods. It is suitable for contexts where sequential information and temporal prediction is important such as in education, where we are interested in predicting students' outcome based on past behavior. Deep Knowledge Tracing [14] is probably the best example of using LSTM to track a student's state of knowledge while interacting with a tutoring system. Numerous variants of LSTM have been proposed, such as the Gated Recurrent Unit (GRU) [4] or the LSTM combined with the attention mechanism, especially the Transformer neural networks [23].

Attention [3] in machine learning refers to a model's ability to focus on specific elements in data. It helps the LSTM to learn where to look in the data. It was initially designed in Neural Machine Translation using sequence-to-sequence (Seq2Seq or encoder-decoder) [19] models. However, since the attention mechanism can improve the prediction results of NN models, it is now widely used in text mining in general. Especially in the education domain, it has been used for question-answering tasks, sequential modeling for student performance prediction, or to predict essay or short answer scoring [25, 17]. Transformer neural networks aim to solve sequence-to-sequence tasks while handling long-range dependencies. It uses the attention mechanism and GPU (Graphics Processing Unit) computing. The input sequence of the Transformer neural network can be passed parallelly, which speeds up the training. It can also overcome the vanishing gradient issue thanks to its multi-headed attention layer. The use of transformers in education is only in its infancy. However, given its notable results (e.g., Generative Pre-trained Transformer (GPT)[2], Bidirectional Encoder Representations from Transformers (BERT)[10]), we think that we will see an increasing number of research papers using this architecture in EDM.

2. PROPOSED FORMAT

Table 1: Timeline and activities

Time	Item
45 minutes	**Presentation**: introduction - Feedforward neural networks and backpropagation
45 minutes	Application - Discussion - Hands-on
30 minutes	Break
60 minutes	**Presentation**: LSTM, Attention Mechanism, and Transformer
60 minutes	Application - Implementation of a LSTM for student performance prediction - Discussion

3. DESCRIPTION OF THE TUTORIAL

3.1 Introduction to feed-forward neural networks

This part begins with artificial neurons and their structure - inputs, weight, output, and activation function - and the calculations that are feasible and not feasible with one neuron only. It continues with feedforward neural networks or multi-layer perceptrons (MLP). A hands-on example taken from [8] illustrates how a feedforward neural network calculates its output. Further, this part introduces loss functions and the backpropagation algorithms and makes clear what a feedforward neural network learns. Backpropagation is demonstrated with the hands-on example introduced before.

3.2 Application of feedforward NN

This part discusses the use of feedforward neural networks in EDM research. These networks are often used to predict students' performance and students at risk of dropping out, see for example [5, 1, 24]. It must be noted that feedforward neural networks do not necessarily give better results than other algorithms for this kind of task. Other uses emerge. For example, Ren et al. use them to model the influence on the grade of a course taken by a student on all other courses that the student has co-taken [16]. As another example, Or and Russel [13] uses intentionally a feedforward "neural network model to both automatically assess the design of a program and provide personalized feedback to guide students on how to make corrections".

It must be noted that neural networks are considered not interpretable, see [12]. When explanations are crucial, it might be worthwhile to evaluate whether interpretable algorithms might be used instead; another way is to generate explanations with other algorithms, see [20] for challenges in doing so.

The main activity of this part is for participants to solve a classification task on an educational dataset; participants will create, inspect and evaluate a feedforward neural network with Python and relevant libraries.

3.3 LSTM

In this part of the tutorial, basic concepts of LSTM are covered. We will focus on how the architecture of different elements (cell, state, etc.) works. Participants will learn how to use an LSTM for the prediction of learners' outcomes in an educational system. Concepts such as Deep Knowledge Tracing (DKT) will also be covered.

3.4 Attention Mechanism

In this part, the attention mechanism is introduced. Participants will learn how this mechanism works and how to use it in different cases. We will explore concepts such as global and local attention in neural networks.

3.5 Transformer neural networks

This part introduces the Transformer neural network architecture. Concepts such as multi-headed attention layer and parallel inputs with the use of GPU will be covered.

3.6 Application

This hands-on part will explore existing real-life applications of LSTM (especially Deep Knowledge Tracing and Knowledge tracing with transformer) in education. We will also explore the combination of LSTM with Expert Knowledge (using the attention mechanism) for Predicting Socio-Moral Reasoning skills [21, 22]. Participants will implement an LSTM, especially a Transformer, with an attention mechanism for the prediction of students' performance in a tutoring system [15]. We will use Keras (Python) library for coding and also use open educational datasets (e.g., Assistments benchmark dataset).

3.7 Objectives and outcomes

The objectives of this tutorial are twofold: 1) introduce the fundamental concepts and algorithms of neural networks to newcomers and then build on these fundamentals to give them some understanding of LSTM and the attention mechanism, especially the Transformer neural networks; 2) provide a place to discuss and exchange about experiences while using neural networks with educational data. Newcomers should leave the tutorial with a good understanding of neural networks and the ability to use them in their own research or to appreciate better research works that use neural

networks. Participants already knowledgeable about neural networks get a chance to discuss and share about this topic and connect with others. A website will be created to display important information to participants: schedule, slides, data, and software to download and install.

4. SHORT BIOGRAPHIES

Agathe Merceron is a Professor of Computer Science at Berlin University of Applied Sciences teaching courses such as machine learning. She was head of the online study program "Computer Science and Media" (Bachelor and Master) till March 31, 2022. Her research interest is in Technology Enhanced Learning with a focus on Educational Data Mining and Learning Analytics. She has served as a program chair for national and international conferences and workshops, in particular for the international conferences Educational Data Mining and Learning Analytics and Knowledge. She is Editor of the Journal of Educational Data Mining and member of the board of the Journal "Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation" (STICEF).

Ange Tato is a Senior Lecturer in computer science at École de Technologie Supérieure de Montréal. She has worked as a research scientist in machine learning at Bem Me Up Augmented Intelligence Montreal for 4 years. Her research interest is in the fundamentals of machine learning algorithms applied to user modeling in intelligent systems. Some of her notable works focus on improving first-order optimization algorithms (with gradient descent); improving neural network architectures for multimodal data to predict or classify user behaviors (players, learners, etc.) in adaptive intelligent systems; and integrating expert knowledge into deep learning models to improve their predictive power and for better traceability of these models. She has served as Poster and Demo Track Co-Chair for Educational Data Mining 2021, Program Committee Member of international conferences such as ICCE, or AIED.

5. REFERENCES

[1] J. Berens, K. Schneider, S. Gortz, S. Oster, and J. Burghoff. Early detection of students at risk - predicting student dropouts using administrative student data from german universities and machine learning methods. *Journal of Educational Data Mining*, 11(3):1–41, 12 2019.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 577–585, Cambridge, MA, USA, 2015. MIT Press.

[4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*. MIT Press, 2014.

[5] G. Dekker, M. M. Pechenizkiy, and J. Vleeshouwers. Predicting students drop out: A case study. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura, editors, *Proceedings of the second International Conference on Educational Data Mining (EDM 2009)*, pages 41–50. International Educational Data Mining Society, July 2009.

[6] A. Ghosh, N. Heffernan, and A. S. Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2330–2339, 2020.

[7] T. Haigh and M. Priestley. von neumann thought turing's universal machine was' simple and neat.' but that didn't tell him how to design a computer. *Communications of the ACM*, 63(1):26–32, 2019.

[8] J. Han, M. Kamber, and J. Pei. *Data Mining - Concepts and Techniques*. Morgan Kaufmann, 2012.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186, 2019.

[11] A. Merceron and A. Tato. An introduction to neural networks. In A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, editors, *Proceedings of the International Conference on Educational Data Mining (EDM 2020)*, pages 821–823. International Data Mining Society, 2020.

[12] C. Molnar. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/, Nov 2022. Last checked on Dec 07, 2022.

[13] J. W. Orr and N. Russell. Automatic assessment of the design quality of python programs with personalized feedback. In I.-H. S. Hsiao, S. S. Sahebi, F. B. chet, and J.-J. Vie, editors, *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, pages 495–501. International Educational Data Mining Society, July 2021.

[14] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 505–513, Cambridge, MA, USA, 2015. MIT Press.

[15] S. Pu, M. Yudelson, L. Ou, and Y. Huang. Deep knowledge tracing with transformers. In *International Conference on Artificial Intelligence in Education*, pages 252–256. Springer, 2020.

[16] Z. Ren, X. Ning, A. Lan, and H. Rangwala. Grade prediction based on cumulative knowledge and co-taken courses. In M. Desmarais, C. F. Lynch, A. Merceron, and R. Nkambou, editors, *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pages 158–167. International Educational Data Mining Society, July 2019.

[17] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, pages 159–168, 2017.

[18] C. Romero, S. Ventura, P. Espejo, and C. Hervás.

Data mining algorithms to classify students. In R. S. J. de Baker, T. Barnes, and J. E. Beck, editors, *Proceedings of the first International Conference on Educational Data Mining (EDM 2008)*, pages 8–17. International Data Mining Society, 2008.

[19] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[20] V. Swamy, B. Radmehr, N. Krco, M. Marras, and T. Käser. Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In N. Bosch and A. Mitrovic, editors, *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, pages 98–109, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[21] A. Tato and R. Nkambou. Infusing expert knowledge into a deep neural network using attention mechanism for personalized learning environments. *Frontiers in Artificial Intelligence*, 5:921476, 2022.

[22] A. A. N. Tato, R. Nkambou, and A. Dufresne. Hybrid deep neural networks to predict socio-moral reasoning skills. In M. Desmarais, C. F. Lynch, A. Merceron, and R. Nkambou, editors, *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pages 623–626. International Educational Data Mining Society, 2019.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[24] K. Wagner, A. Merceron, and P. Sauer. Accuracy of a cross-program model for dropout prediction in higher education. In *Companion Proceedings of the 10th International Learning Analytics & Knowledge Conference (LAK 2020)*, pages 744–749, 2020.

[25] X. Xiong, S. Zhao, E. G. Van Inwegen, and J. E. Beck. Going deeper with deep knowledge tracing. In T. Barnes, M. Chi, and M. Feng, editors, *Proceedings of the International Conference on Educational Data Mining (EDM 2016)*, pages 545–550. International Data Mining Society, 2016.