Neural Networks in the Educational Sector: Challenges and Opportunities

Ugo FIORE
Department of Management and Quantitative Studies Parthenope University, Napoli, Italy
ugo.fiore@uniparthenope.it

ABSTRACT

Given their increasing diffusion, deep learning networks have long been considered an important subject on which teaching efforts should be concentrated, to support a fast and effective training. In addition to that role, the availability of rich data coming from several sources underlines the potential of neural networks used as an analysis tool to identify critical aspects, plan upgrades and adjustments, and ultimately improve learning experience. Analysis and forecasting methods have been widely used in this context, allowing policy makers, managers and educators to make informed decisions. The capabilities of recurring neural networks—in particular Long Short-Term Memory networks—in the analysis of natural language have led to their use in measuring the similarity of educational materials. Massive Online Open Courses provide a rich variety of data about the learning behaviors of online learners. The analysis of learning paths provides insights related to the optimization of learning processes, as well as the prediction of outcomes and performance. Another active area of research concerns the recommendation of suitable personalized, adaptive, learning paths, based on varying sources, including even the tracing of eye-path movements. In this way, the transition from passive learning to active learning can be achieved. Challenges and opportunities in the application of neural networks in the educational sector are presented.

Keywords: neural networks; recurring networks; learning paths.

INTRODUCTION

Learning useful representations from raw data means extracting relevant information in a compact form and removing redundant information as well as noise. In other words, constructing a simplified model that explains observed data. Analysis of the obtained representation can highlight latent factors, disclose previously unseen relationships among variables, and ultimately help gaining useful insight into the phenomenon being observed. Finding a good representation is crucial in multiple research fields, where data come from several sources and are characterized by high complexity. Neural networks are a widely used and successful representation learning technique. Neural networks, as their name suggests, are inspired by the structure of the cortex in the human brain. They consist of a number of units arranged in a directed graph (undirected for the Boltzmann machines) by means of connections. A unit takes as input a weighted sum of the outputs of the units connected to it and produces its output by applying to that sum a nonlinear activation function—typical such functions are the hyperbolic tangent and the logistic sigmoid. The neural computation model has some nice theoretical properties and neural networks can be shown to be universal approximators (Goodfellow et al., 2016).
Neural networks learn from a collection of training samples. Training a neural network is usually done by means of Stochastic Gradient Descent, with the calculation of the gradient of the loss function (quantifying the prediction error) with respect to the network parameters being obtained through the backpropagation algorithm. To keep the architecture simple, restrictions are applied to the topological structure of networks: Units are arranged in layers, with connections only between units in adjacent layers (Fig. 1). Intermediate layers are called hidden layers. Neural networks with at least two (three for some authors) hidden layers are called deep learning networks. It is this hierarchical structure that provides deep network with the ability to build powerful representations. Subsequent layers work on intermediate representation constructed by previous layers, so that internal representations are at an increased level of “abstraction”.

RECURRING NEURAL NETWORKS

Like other models, neural networks work on the assumption that the examples are independently and identically distributed according to an (unknown) distribution. Thus, the order in which examples appear is unimportant. Sequential data raises unique challenges for neural networks, because order-based dependences among data need to be captured. Even though networks can be designed to cope with fixed—length sequences, dependences may extend over variable—length intervals, with possibly long gaps. An architectural change is therefore required. While a conventional neural networks has connections only between units in adjacent layers, a Recurrent Neural Networks (RNN) may have cycles in its graph structure. In this way, a state can be constructed and maintained that contains information—Goodfellow et al. (2016) effectively called it a “lossy summary”—about the whole sequence observed so far. Upon observing new sequence elements, RNNs update their current state vector to reflect changes. The problem becomes how to isolate important changes and discard irrelevant ones.

LSTM networks

In theory, RNNs are able to treasure on dependences of any length. In practice, however, very long chains of gradient propagation when the network is unrolled in time will lead to vanishing gradients (Bengio et al., 1994). A mechanism to control the accumulation and propagation of state variations is needed. To cope with this problem, gated RNNs were introduced, including for the Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) and for the Gated Recurrent Unit (GRU) networks (Cho et al., 2014).
These networks have the ability of controlling the amount of information about past inputs that is be preserved at each stage. The self-loop is regulated by additional units—gates—that introduce the ability to forget old state information.

![Figure 1: An LSTM cell (Olah 2015)](image)

Figure 1 (Olah 2015) shows the structure of the repeating module in an LSTM network, highlighting the cell state (upper part), the output (lower part). From left to right, the forget gate block, the input gate block, the candidate state gate block, and the output gate are shown. The symbol $\sigma$ stands for the logistic sigmoid, which squashes its input in the interval $(0,1)$, where the multiplication sign indicates the Hadamard product. Differently from an LSTM networks, in a GRU forgetting and the updating of the cell state is delegated to a single gate. GRUs, being simpler, have shown improvements over LSTMs in computational performance. The two models are otherwise competitive on a wide range of problems.

**APPLICATION TO THE EDUCATIONAL SECTOR**

It is worth mentioning that psychological studies on human and animal learning have been conspicuous sources of inspiration in developing machine learning paradigms. Machine learning, in its general meaning of automatically deriving knowledge from experience—crystallized in data—is particularly attractive in the educational sector. There are two reasons for this. Firstly, the educational environment is so complex that little assumptions can be made about the data distribution. Secondly, vast amounts of data are available for exploration.

Useful applications of machine learning in education include a variety of objectives (Coelho & Silveira, 2017). Accurate monitoring student’s states during learning can support personalized, flexible, and adaptive learning, with direct benefit for students and an increased retention rate for providers. Student modeling can be based on several data sources, including for interaction logs, facial features, and eye movements.

The application of deep learning models to educational data gained momentum in 2015 (Guo et al., 2015), when a prediction system for student performance was introduced. An interesting benefit of such a system is its capability of providing early warnings so that students at risk could be identified where there is still time for corrective actions. While applying deep learning and RNN models to an educational context is obviously desirable, the scenario creates some unique challenges that need to be addressed. In particular, inhomogeneity and redundancy often characterize data in educational analysis, especially in detection of student boredom, and they should be handled properly.

Designing handcrafted feature to represent student behavior can be challenging (Bosch & Paquette 2017). Unsupervised autoencoders are trained to find data embeddings, mappings to low-dimensional spaces that (a) improve the performance of classifiers, and (b) have the potential of showing interesting insights in data, highlighting previously unseen connections. Despite being useful as building blocks in modular architectures of complex neural networks, the embedding themselves can be analyzed and studied separately, looking for clues about unexpected associations evidenced by spatial closeness in the simplified representation.
In a personalized and adaptive learning environment the learning path, instead of being fixed, is continuously adapted, based on student’s individual characteristics and knowledge state, to help students achieve their learning objectives in the shortest possible time. Personalized recommendation systems enable the realization of customized learning path for different individuals, treasuring on the experience of others. Recommendation systems should be optimized in terms of diversity, novelty and interaction intensity. In early recommendation systems, content based filtering derived recommendations for a learner on the basis of what was preferred in the past by learners with similar tastes. In order to aggregate learners with similar preferences in Collaborative Learning, it is natural to think to clustering algorithms based on various similarity metrics (Pelánek, 2019).

Sparsity and volume of the data volume call, however, for different solutions that can scale in a better way. Kim et al. (2017) combined Probabilistic Matrix Factorization with a Convolutional Neural Network (CNN) to model contextual information and consider Gaussian noise. Features used to represent learning resources need to keep some fundamental assumptions into account (Zhou et al., 2018). In particular, some knowledge is regarded as essential in a learning plan and it ought to be included in any path relative to that plan. Zhou et al. (2018) used an LSTM predictor for learning paths, in particular because of its ability to handle sequences of different length. In contrast, Kim et al. (2017) preferred a CNN to a LSTM or GRU, because of the faster training times offered by the former. In fact, CNN’s, due to their fixed structure, can use simple backpropagation, whereas recurrent networks have to resort to backpropagation through time in order to keep long-term dependencies.

The relationship between learners, items, and tags can be represented by means of a tripartite graph, which was originally static and based on historical information. Recently, an approach where the interaction tripartite graph—modeling the ternary relation among learners, interaction behaviors, and learning content—is made dynamic has been proposed (Hu et al., 2019). In this way, trendy topics attracting much attention may easily propagate among learners. The weights in the dynamic interaction tripartite graph are initialized and then through an attention-driven CNN.

In online platforms, a large number of exercises are prepared and loaded to assess the degree to which a learner has mastered a topic. The ability to find similar exercises, i.e., exercises sharing the same purpose, can substantially improve the richness of learning. Automatically grouping exercises on the basis of similarity is not at all trivial, because exercises usually contain heterogeneous data such as text and images, and similarity at word level—and even at notion level—can easily lead to erroneous grouping. For this task, a CNN and an Attention-based LSTM have been combined (Liu et al., 2018). The CNN processes images, an embedding layer creates representations for notions, while the Attention-based LSTM produces the final, semantic, representation. Such combination of components is telling of a research trend that is in progress. In future developments, subnetworks will either continue to be juxtaposed in a modular way, each component dedicated to the portions of input it handles best, or we might witness the development of new, hybridized architecture designed specifically so that it will be natively able to process all the data.

CONCLUSIONS

Discovering hidden structure and patterns in data originated from online learning systems is valuable in education, as it permits to gain a deeper understanding and to devise a highly flexible, adaptive, and personalized offering. Deep learning networks and their capability to untangle previously unanticipated connections are very promising tools in this endeavor. Choosing the most appropriate deep network architecture for a given task is still a problem requiring skill and expertise. The main architectures offer advantages and disadvantages, in terms of capabilities and performance, and careful weighing is needed when making a selection. Once the choice has been made, the next step is to determine suitable architectural hyperparameters, which also requires extensive experiments to find the level of inductive bias that improves generalization capability.

The availability of public datasets to experiment new ideas and evaluate their performance is a
critical factor for the research in this field. Currently available datasets for education, for instance the Edx or the WorldUC datasets, are a starting point but cannot completely cover the requirements for some experiments (Hu et al., 2019). Extensions of publicly available data would therefore be welcome.

Perspectives for future research are wide and auspicious. Regarding improvements to the RNN architecture, several attempts have been made, among which the most exciting appears to be attentional interfaces (Vaswani et al., 2017), where an RNN can focus, depending on the context, on salient parts of its input that are relevant for the prediction of the next target; a specific module regulates the decision. For example, an RNN can control the output of another RNN. All of the proposed improvements seem to be conducive to a relaxing of the topological constraints in network layout, an idea which has started to yield interesting results with skip connections in residual networks (He et al., 2016), and hypernetworks (Ha et al., 2017).

ACKNOWLEDGEMENT

The author gratefully acknowledges the support of Lucian Blaga University, Sibiu, where the colleagues create a uniquely productive collaboration environment.

REFERENCES

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157–166.

Bosch, N., & Paquette, L. (2017). Unsupervised deep autoencoders for feature extraction with educational data. Paper presented at the Deep Learning with Educational Data Workshop at the 10th International Conference on Educational Data Mining, Urbana, IL, USA.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1724–1734), Association for Computational Linguistics.

Coelho, O. B., & Silveira, I. (2017). Deep Learning applied to Learning Analytics and Educational Data Mining: A Systematic Literature Review. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE) (Vol. 28, No. 1, p. 143–152).

Goodfellow, I, Bengio, Y, & Courville, A. (2016). Deep learning: MIT Press.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In 2015 International Symposium on Educational Technology (ISET) (pp. 125–128), IEEE.

Ha, D., Dai, A., & Le, Q. V. (2016). Hypernetworks. arXiv preprint arXiv:1609.09106.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition – CVPR 2016 – (pp. 770–778), IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hu, Q., Han, Z., Lin, X., Huang, Q., & Zhang, X. (2019). Learning peer recommendation using attention-driven CNN with interaction tripartite graph. Information Sciences, 479, 231–249.
Kim, D., Park, C., Oh, J., & Yu, H. (2017). Deep hybrid recommender systems via exploiting document context and statistics of items. *Information Sciences, 417*, 72–87.

Liu, Q., Huang, Z., Huang, Z., Liu, C., Chen, E., Su, Y., & Hu, G. (2018). Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1821–1830), ACM.

Olah, C. (2015). Understanding LSTM networks. Retrieved August 15, 2019, from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Pelánek, R. (2019) Measuring Similarity of Educational Items: An Overview. *IEEE Transactions on Learning Technologies.* (Early Access: DOI:10.1109/TLT.2019.2896086).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems 30 – NIPS 2017* – (pp. 5998–6008), Curran Associates, Inc.

Zhou, Y., Huang, C., Hu, Q., Zhu, J., & Tang, Y. (2018). Personalized learning full-path recommendation model based on LSTM neural networks. *Information Sciences, 444*, 135–152.