Deep Learning for Automated Scoring*

Aoife Cahill

1 Introduction

Recently, deep-learning, neural network algorithms have successfully been applied to a range of tasks including automated speech recognition, image recognition and machine translation. The main advantage of neural networks is that they can automatically learn useful features and patterns from data, removing the need for manual feature design and engineering. In particular, recurrent neural networks can process arbitrary sequences of inputs, ideal for processing natural language. In the last few years, researchers have applied neural networks to the field of automated scoring and are beginning to see results comparable to existing state-of-the-art systems. We will present an overview of the recent work in applying deep-learning algorithms to automated scoring tasks.

2 Neural Networks

Deep learning refers to a technology based on neural networks. It is thought that they received their name because of the perceived link between the structure of the models and the neural networks found in the brain. Each network is made up of a collection of connected nodes (or neurons). Information is passed from node to node, and the connections between nodes often have weights associated with them. Typically, nodes are arranged in layers, where each layer can perform different transformations on its inputs before passing on its outputs. Data is processed by being passed into the first layer, undergoing several transformations before finally arriving at the final layer.

A convolutional neural network (CNN) is a type of neural network that consists of an input and an output layer, as well as multiple hidden layers. The main differentiating characteristic of these neural networks is the use of the mathematical concept of convolution. CNNs can be more efficient than neural networks because nodes are typically only connected to a subset of all nodes. This reduces the number of parameters that need to be learned in the model.

Recurrent Neural Networks (RNNs) are a type of neural network that have the ability to memorize and can therefore make use of sequential information.

*Paper presented at the annual conference of the National Council on Measurement in Education, New York, April, 2018.

This is particularly important in NLP tasks where the sequence of words is vital to the correct interpretation of the language. The most common types of RNNs are Long-Short Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and the bi-directional variant (Bi-LSTM) (Graves, 2012) has also shown considerable promise in NLP tasks.

3 Automated Scoring with Neural Networks

Research on automated scoring using neural networks has only appeared since around 2016, and most of that research is focused on automated essay scoring. Most of the work has been conducted on English data (e.g. on the ASAP dataset[1], or the FCE dataset (Yannakoudakis et al., 2011) or the TOEFL11 corpus (Blanchard et al., 2013)). However, there has also been some initial work on other languages including German (Horbach et al., 2017).

Most of the neural network approaches to automated scoring are based on recurrent neural network architectures (Taghipour and Ng, 2016; Alikaniotis et al., 2016; Dong et al., 2017; Tay et al., 2017; Cummins and Rei, 2018). Zhao et al. (2017) investigate the combination of the recurrent and convolutional neural network architectures and Östling and Grigonyte (2017) build a quality assessment model for feedback based on deep convolutional networks with residual connections.

There has not been as much attention on automated scoring of other modalities using neural networks. Riordan et al. (2017) described a neural architecture based on that of Taghipour and Ng (2016) for automated scoring of short answer content items. Yu et al. (2015) presented a neural architecture based on recurrent neural network for automated scoring of spoken response. Malinin et al. (2017b,a) present a neural network for off-topic detection in spoken response automated scoring.

4 Discussion

Deep neural networks appear to be a promising area of research in the area of automated scoring for essays, content and speech. Performance of the systems proposed is roughly in line with state of the art using current machine learning methods. One of the main advantages of using neural network approaches to automated scoring is that the need for careful manual feature engineering is removed and the bulk of the effort in model development is in designing the model architecture and tuning the model parameters. However, this also means that the traditional method of measuring the construct coverage of the models — by aligning features to aspects of the scoring rubrics — is impossible. Knowing that an automated scoring model is measuring the construct correctly, and not simply measuring spurious noise in the signal, is important for test fairness and validity. Without this step, models are susceptible to gaming strategies that

[1] ttps://www.kaggle.com/c/asap-aes

can take advantage of the spurious noise in the data (e.g. the high correlation between essay length and human scores).

The interpretability of neural network models is a very active area of research. Most of the research has focused on the interpretation of models used in the field of computer vision (Simonyan et al., 2013; Yosinski et al., 2015), however, there have also been some recent developments in the field of NLP (Li et al., 2016) and Speech (Tan et al., 2015). If we are to consider the use of deep neural network models for automated scoring, interpretability will play an important role in determining how well the models are measuring the relevant construct correctly. Alikaniotis et al. (2016) generate interpretable visualizations of their automated essay scoring network. While the output has some drawbacks (the context of word usage is not taken into account), it is an important step forward in making sure that we pay close attention to what these models are measuring.

While research into deep learning methods and their interpretability continues, one possibility for including deep learning in automated scoring is by using deep learning for lower-level construct-aligned feature development. For example, Eger et al. (2017) proposed a neural architecture for automatically identifying argumentation elements. These kinds of features could be used in automated essay scoring in a more traditional simpler linear model that is easier to interpret and link to the construct.

The field of deep learning and automated scoring does not seem to be quite at the point where we could deploy such models in a production scenario. We do not yet have sufficient evidence that the models meet our ethical standards, are measuring the construct correctly and are not susceptible to gaming techniques.[2] However, there does seem to be promising evidence that deep learning could be used to improve components of automated scoring systems.

References

Alikaniotis, D., Yannakoudakis, H., and Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2).

Cummins, R. and Rei, M. (2018). Neural Multi-task Learning in Automated Assessment. *ArXiv e-prints*.

[2]Known gaming strategies could of course be detected by external filtering components, however this does not guarantee that these models are not susceptible to new, currently unknown, gaming strategies.

Dong, F., Zhang, Y., and Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Graves, A. (2012). Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711.*

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

Horbach, A., Scholten-Akoun, D., Ding, Y., and Zesch, T. (2017). Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366, Copenhagen, Denmark. Association for Computational Linguistics.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2016). Visualizing and understanding neural models in nlp. In *Proceedings of NAACL-HLT*, pages 681–691.

Malinin, A., Knill, K., and Gales, M. J. (2017a). A hierarchical attention based model for off-topic spontaneous spoken response detection. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 397–403. IEEE.

Malinin, A., Knill, K., Ragni, A., Wang, Y., and Gales, M. J. (2017b). An attention based model for off-topic spontaneous spoken response detection: An initial study. In *at ISCA Workshop on Speech and Language Technology for Education (SLaTE).*

Östling, R. and Grigonyte, G. (2017). Transparent text quality assessment with convolutional neural networks. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 282–286, Copenhagen, Denmark. Association for Computational Linguistics.

Riordan, B., Horbach, A., Cahill, A., Zesch, T., and Lee, C. M. (2017). 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, Copenhagen, Denmark. Association for Computational Linguistics.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034.*

Taghipour, K. and Ng, H. T. (2016). A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

Tan, S., Sim, K. C., and Gales, M. (2015). Improving the interpretability of deep neural networks with stimulated learning. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 617–623.

Tay, Y., Phan, M. C., Tuan, L. A., and Hui, S. C. (2017). Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. *arXiv preprint arXiv:1711.04981*.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., Tao, J., Ivanou, A., and Qian, Y. (2015). Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 338–345. IEEE.

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., and Heffernan, N. (2017). A memory-augmented neural model for automated grading. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 189–192. ACM.