Visual-Explainable AI: The Use Case of Language Models

Tanja Munz-Körner, Sebastian Künzel, Daniel Weiskopf

University of Stuttgart, Visualization Research Center, Germany

We address the research problem of explainable AI (Artificial Intelligence) in the context of language models using visualization and visual analytics. Our goal is to open the black box of language models to make them more transparent and interpretable for humans, find problems (debugging), and improve accuracy.

Different strategies can be chosen to analyze deep learning models [1]. One approach is to explore a model's internal states while performing a prediction. This helps see how they change and contribute to the final prediction. Another method is to explore the actual prediction results that can provide insight into the prediction and provide a quality assessment about the performance of the model. Other approaches, such as exploring the training process, are also possible but are not considered in this work.

These strategies can be applied to a variety of data types. Here, we chose the use case of language models. In the area of natural language processing (NLP), AI techniques can create impressive results, but results may still be incorrect. Explainable AI can help understand why certain predictions were made. We investigate NLP as part of data-integrated simulation science: in the form of supporting cognitive aspects in the creation of a digital human model, which is one of the visionary examples of the Cluster of Excellence "SimTech".[1]

We developed two approaches in this context: (1) One[2] [2] shows how internal structures (hidden states of long short-term memories) and expected predictions change when processing text sequences and performing a classification task (see Figure 1). It shows how the prediction would be if we stopped at an earlier step and how the final prediction was created. For both correct and incorrect predictions, our approach can help analyze what factors influenced such a prediction. (2) Our neural machine translation system[3] [3, 4] (Figure 2) allows the translation of a document, provides information about the quality, and, for individual sentences, shows how the prediction for the translation was made (using a beam search visualization). The recommended prediction can be adapted by choosing an alternative translation or adding user input. Furthermore, the attention visualization of internal states shows the relationship between the words of a source sentence and their translation. Finally, the quality of the model can also be improved by fine-tuning the model using the corrections of the user.

In the future, we plan to apply our methods in an adapted form to other simulation-related data and develop new methods to explain AI models in other language-related areas and beyond. Currently, a popular research area is Visual Question Answering (VQA) [5], where we also intend to make internal states visible to users such that they better understand prediction results.

[1] https://www.simtech.uni-stuttgart.de/

[2] https://github.com/MunzT/hiddenStatesVis

[3] https://github.com/MunzT/NMTVis

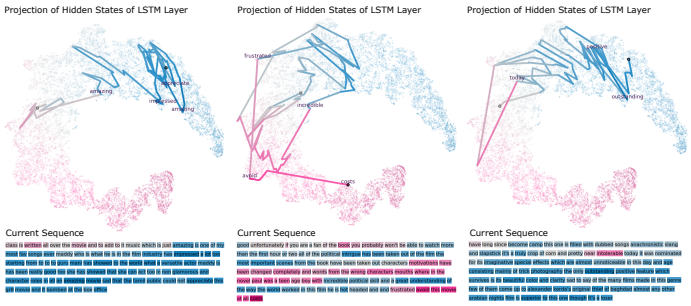

Figure 1: Examples of correctly and incorrectly classified sequences: (1) Words like "amazing" contribute to a positively classified movie. (2) The prediction result changes multiple times until the final negative decision. (3) A wrong prediction. Positive words dominate it this sequence. Data: IMDB [6].

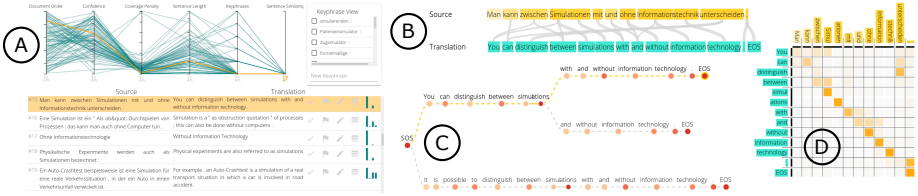

Figure 2: Our approach consists of multiple visualizations: (A) The main view shows all sentences of a document and a visualization of multiple quality metrics. (B) and (D) show the attention graph-based and in form of a matrix, and (C) is an interactive beam search representation. Data: German Wikipedia article for *simulation* [7].

Funding. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2075 – 390740016.

References

[1] Garcia, R., Telea, A. C., da Silva, B. C., Tørresen, J., Dihl Comba, J. L. A task-and-technique centered survey on visual analytics for deep learning model engineering. *Computers & Graphics*, 77:30–49, 2018.

[2] Garcia, R., Munz, T., and Weiskopf, D. Visual analytics tool for the interpretation of hidden states in recurrent neural networks. *Visual Computing for Industry, Biomedicine, and Art*, 4(1):1–13, 2021.

[3] Munz, T., Väth, D., Kuznecov, P., Vu, N. T., and Weiskopf, D. Visual-interactive neural machine translation. In: *Proceedings of Graphics Interface 2021*, pages 265–274. 2021.

[4] Munz, T., Väth, D., Kuznecov, P., Vu, N. T., and Weiskopf, D. Visualization-based improvement of neural machine translation. *Computers & Graphics*, 103:45–60, 2022.

[5] Antol, S., et al. VQA: Visual question answering. In: *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425-2433, 2015.

[6] Chollet, F.: Keras. GitHub. `https://github.com/fchollet/keras` (2015)

[7] Wikipedia, Simulation – Wikipedia, die freie Enzyklopädie, 2021. URL: `https://de.wikipedia.org/wiki/Simulation`, accessed February 28, 2023.