Transforming Process Mining: A Transformer-Based Approach to Semantic Clustering in Event Log Analysis

Zineddine Bettouche[1]

[1]Deggendorf Institute of Technology (DIT), Dieter-Goerlitz-Platz 1; 94469 Deggendorf

Research Question

Can the employment of transformer models and clustering algorithms facilitate event log analysis in process mining by forming more cohesive case clusters that capture meaningful semantic coherence, allowing for the representation of specific behavioral patterns within the logs?

Motivation

The field of process mining plays a pivotal role in optimizing business processes by examining digital footprints. Central to process mining is event log analysis, where event sequences are scrutinized to unveil patterns, irregularities, and potential bottlenecks. While traditional deep learning networks, such as recurrent neural networks (RNNs), have excelled in predicting process behaviors due to their precision and adaptability, they face challenges in handling long-range dependencies, gradient issues, and computational efficiency for lengthy sequences. Addressing these challenges led to the introduction of the attention mechanism, embodied by transformer models.

Transformer models excel in tasks like machine translation and natural language processing. Encoder-only transformers have consistently achieved state-of-the-art results, particularly when paired with clustering techniques. This extends beyond the domain of natural language processing to structured data, where encoding textual information and subsequently clustering the encoded representations can lead to semantically coherent clusters, providing valuable insights and revealing hidden patterns. These patterns can be critical in areas such as understanding customer behavior in e-commerce, optimizing manufacturing processes in industry, or tracking disease progression in healthcare. Discovering these patterns informs decision-making and process improvements across various domains. However, despite their demonstrated success in various applications, the application of transformer models in process mining remains a largely unexplored frontier.

ICPM Doctoral Consortium and Demo Track 2023

zineddine.bettouche@th-deg.de (Z. Bettouche)

Methodology

This research adopts a two-step approach. Initially, a transformer model will be trained as a masked language model, requiring it to predict masked portions of input data based on surrounding context. This training process ensures that the encoder learns to represent input data as vectors in a latent space, capturing both contextual and semantic relationships.

The second step involves developing a methodology that leverages the encoder model to transform event logs into semantically meaningful vectors within the latent space. These encoded vectors will then be subjected to a suitable clustering algorithm, aiming to create meaningful clusters that reveal behavioral patterns in the event logs. However, assessing the quality of the constructed semantic clusters requires the introduction of novel evaluation measures.

When selecting a tokenizer for transformer models, careful consideration must be given to model compatibility, granularity of the text, and domain specificity. The efficiency of the chosen tokenizer impacts computational costs and sequence lengths, which must align with practical constraints such as available computational resources and tool support. Empirical testing on the specific task and dataset is essential to strike a balance between performance and computational requirements.

Configuring the transformer model, including decisions on the number of encoder layers and attention heads, should be driven by dataset size, task complexity, and available computational resources. While initial configurations can be informed by established architectures like BERT [1] or RoBERTa [2], fine-tuning these settings based on validation performance is critical. A systematic hyperparameter search, combined with insights from recent literature, will guide this decision-making process to optimize model performance and computational efficiency.

To address the challenges of clustering high-dimensional vectors generated by transformers for structured data like event logs, dimensionality reduction techniques, such as t-SNE [3] or PCA [4], will be applied before clustering. Alternatively, algorithms like DBSCAN [5] or HDBSCAN [6], which do not require a priori specification of the number of clusters and can handle varying densities, will be explored for clustering.

Beyond the pre-training phase, the research will evaluate the versatility of the model through fine-tuning tasks. An exemplary task is anomaly detection, evaluated using metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) [7] and the Area Under the Precision-Recall Curve (AUC-PR) [8]. These metrics, especially AUC-ROC, effectively capture the model's ability to distinguish between normal and anomalous sequences, serving as a testament to its post-fine-tuning effectiveness.

Additionally, the research will incorporate novel metrics, as proposed by Sommers et al. [9], to assess the model's capacity to generalize beyond its training data. These metrics will address fundamental questions: Can the neural network effectively learn its task? How do the rediscovered models (semantic clusters) compare to their original counterparts both structurally and behaviorally? How does the model perform in real-world scenarios without a definitive ground truth? Incorporating these evaluation metrics will provide a robust framework for analyzing the efficacy and versatility of the research findings.

State-of-the-Art

Deep neural networks, since their introduction in the field of process mining, have improved business process monitoring and discovery. Evermann et al. [10] introduced the use of recurrent neural networks (RNNs) to predict real-time process behaviors. Building on this, a range of prediction models grounded in RNNs and their derivatives like the long-short term memory (LSTM) networks have emerged. They involved several tasks such as predicting the next activity [11], suffix generation [12], and more.

However, there are notable limitations with these networks, especially when handling lengthy sequences. For instance, the efficiency of LSTM networks tends to decrease in relation to the increasing length of event sequences [13]. This is undesired since the intricacies of event logs often involve control flows linking activities, creating the need for recognizing both short and long-range dependencies. Also, due to their sequential nature, LSTMs and RNNs don't support parallel processing, leading to significant inefficiencies during learning and inference phases.

To address these limitations, the attention mechanism was introduced [14]. It has proven effective in handling long-range dependencies in sequences. Transformer neural network architectures, using the self-attention mechanism, have become prominent in neural machine translation and natural language processing. Their contributions span a wide range of tasks such as text generation [15], sentiment analysis [16], and more. Transformers have also been applied to non-textual data, such as images [17] and protein sequences [18]. Their applications include information retrieval tasks with structured data [19].

In the context of event log analysis, trace clustering [20] is a fundamental technique for understanding and enhancing process behaviors. Deep learning techniques have gained traction in predictive analytics for process mining, particularly in tasks like predicting the next event or remaining time [21]. The significance of trace clustering also extends to customer journey analysis, which helps us grasp and improve customer behaviors in omnichannel environments [22]. Utilizing domain-informed similarity metrics [23] enhances the quality of customer journey clustering, contributing to process improvement efforts. Furthermore, the field of predictive process analytics emphasizes the need for interpretability in deep learning-based models. The work by Wickramanayake et al. [24] has shown that attention-based models can provide comprehensive explanations for process predictions, addressing the challenge of "black-box" models.

The research aims to extend the application of transformer models and attention-based mechanisms to event log analysis, leveraging their strengths and addressing identified challenges, including enhancing cluster quality and interpretability in event log analysis. Encoder-only transformers, like BERT, excel in translating diverse inputs into a unified latent space, facilitating downstream tasks like clustering and anomaly detection. Although the next-activity prediction was addressed by training an encoder-decoder transformer [25], their application in transforming input data into semantically meaningful vectors in the latent space is underutilized in process mining.

Acknowledgement

This paper has received funding from the State of Bavaria in the context of the project KIGA, funding no. DIK0313.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Bidirectional encoder representations from transformers, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[3] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605.

[4] I. T. Jolliffe, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 8 (2016) 216–222.

[5] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (1996) 226–231.

[6] R. J. Campello, D. Moulavi, J. Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection, in: Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, 2015, pp. 839–847.

[7] T. Fawcett, An introduction to roc analysis, Pattern recognition letters 27 (2006) 861–874.

[8] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, Proceedings of the 23rd international conference on Machine learning (2006) 233–240.

[9] D. Sommers, V. Menkovski, D. Fahland, Process discovery using graph neural networks, CoRR abs/2109.05835 (2021). URL: https://arxiv.org/abs/2109.05835.

[10] J. Evermann, J.-R. Rehse, P. Fettke, A deep learning approach for predicting process behaviour at runtime, in: M. Dumas, M. Fantinato (Eds.), Business Process Management Workshops, Springer, 2017, pp. 327–338.

[11] S. Pandey, S. Nepal, S. Chen, A test-bed for the evaluation of business process prediction techniques (2011) 382–391.

[12] M. Camargo, M. Dumas, O. Gonz´alez-Rojas, Learning accurate lstm models of business processes (2019) 286–302.

[13] D. Paperno, G. Kruszewski, A. Lazaridou, Q. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, R. Fern´andez, The lambada dataset: Word prediction requiring a broad discourse context (2016) 1525–1534.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762.

[15] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[16] J. Delbrouck, N. Tits, M. Brousmiche, S. Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, CoRR abs/2006.15955 (2020). URL: https://arxiv.org/abs/2006.15955.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[18] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. USA 118 (2021) e2016239118. URL: https://doi.org/10.1073/pnas.2016239118.

[19] Y. Guo, Z. Ma, J. Mao, H. Qian, X. Zhang, H. Jiang, Z. Cao, Z. Dou, Webformer: Pre-training with web pages for information retrieval, in: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1502–1512. URL: https://doi.org/10.1145/3477495.3532086.

[20] S. Sakr, A. Y. Zomaya (Eds.), Encyclopedia of Big Data Technologies, Springer, 2019. URL: https://doi.org/10.1007/978-3-319-63962-8. doi:10.1007/978-3-319-63962-8.

[21] I. Ketykó, F. Mannhardt, M. Hassani, B. F. van Dongen, What averages do not tell: Predicting real life processes with sequential deep learning, in: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (SAC), 2022, pp. 1128–1131. doi:10.1145/3450283.3450309.

[22] M. Hassani, S. Habets, Predicting next touch point in a customer journey: A use case in telecommunication, in: European Conference on Modelling and Simulation (ECMS), 2021, pp. 48–54.

[23] S. van den Berg, M. Hassani, On inferring a meaningful similarity metric for customer behaviour, in: European Conference on Modelling and Simulation (ECMS), 2021, pp. 234–250.

[24] B. Wickramanayake, Z. He, C. Ouyang, C. Moreira, Y. Xu, R. Sindhgatta, Building interpretable models for business process prediction using shared and specialised attention mechanisms, Knowledge-Based Systems 248 (2022) 108773. doi:10.1016/j.knosys.2022.108773.

[25] Z. A. Bukhsh, A. Saeed, R. M. Dijkman, Processtransformer: Predictive business process monitoring with transformer network, CoRR abs/2104.00721 (2021). URL: https://arxiv.org/abs/2104.00721.