Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Optimizing Causal Inference Approach for Exploring Shallow Reading Behavior with Generative Adversarial Networks

Yu BAI[a], Fuzheng ZHAO[b], Wenhao WANG[c] & Chengjiu YIN[d]*
[a]*Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan*
[b]*Education Technology Center, Jilin University, China*
[c]*Graduate School of System Informatics, Kobe University, Japan*
[d]*Research Institute for Information Technology, Kyushu University, Japan*
*yin.chengjiu.247@m.kyushu-u.ac.jp

Abstract: The prevalence of shallow reading in online digital learning is steadily increasing, which has sparked interest in revealing the mechanisms behind shallow reading behavior, especially analyzing the causal relationship between its constituent features and learning performance. However, current causal analysis methods have many limitations in terms of experimental conditions, data independence assumptions, and analysis costs. Drawing on the application experience of Markov chain theory in the field of causality, this study adopts the structure-agnostic model (SAM) algorithm to design the structure, parameter loss, and learning process, and proposes an evaluation method for causal exploration based on generative adversarial neural networks (GANs). The study shows that the proposed maximum mean diversity (MMD) optimization method improves the stability of the model analysis results and clarifies that reading speed is a key factor in the occurrence of shallow reading behavior.

Keywords: Shallow reading, generative adversarial neural network, causal inference

1. Introduction

With the rise of the Internet and digital learning technologies, there has been significant attention on issues like fragmented reading time, shallow reading, and the challenge of deep comprehension in fast-paced reading styles (Romero & Ventura, 2020). The digital learning environment fostered by computer-assisted learning has placed a heightened emphasis on shallow reading (Haleem et al., 2020) This has led shallow reading to become a pivotal reading behavior within the online learning context alongside deep reading (Carr, 2020).

It's noteworthy that commonly employed causal analysis techniques, including randomization, quasi-experimental methods, and machine learning approaches like Bayesian networks, come with strict requirements related to experimental assumptions, data independence, and linearity (Spirtes & Zhang, 2016). These requirements limit their applicability and scope, resulting in an insufficient utilization of causal inference. The rapid advancement of deep learning is driven by its adaptability to diverse data distributions, suitability for training nonparametric models, and the robustness gained from transfer learning capabilities (Livingstone, Manallack, & Tetko, 1997). GANs exhibit substantial potential for shedding light on causal inference, offering several advantages including the flexibility to capture intricate data distributions, simulate causal effects under various interventions, enhance the resilience of causal models, and fit nonparametric models seamlessly.

Drawing on the application experience of Markov chain theory in the field of causal relationships, the SAM algorithm was used to design the structure, parameter loss, and learning process, and a model evaluation method based on GANs for causal relationship exploration was proposed to analyze the causal relationship of shallow reading behavior

2. Literature Review

Shallow reading definition and features. Defining "shallow reading" is challenging (Bai et al., 2024). Often contrasted with "Deep reading", it's linked to fragmentation in digital environments (Gordon, 2023). Goodfellow et al. (2014) oppose it to deep comprehension, critical thinking, and reflection. The digital age, with its interactive features, has reshaped reading habits (Haleem et al., 2022). Carr (2020) highlights the internet's role in fostering shallow reading, reducing attention span, and hindering comprehension. Shallow reading garners attention not only for its behavioral characteristics in the digital age but also as a learning strategy. Its principles and effects in learning scenarios are under scrutiny (Zhao, Hwang, & Yin, 2021). While shallow reading can also limit understanding (Sarris, 2022).

Causality inference methods and challenges. Traditional methods are time-consuming and resource intensive. Machine learning including Bayesian networks faces limitations in handling high-dimensional data and identifying causation (Nauta Bucur & Seifert, 2019). Neural networks excel in capturing non-linear relationships and learning causal structures from raw data. GANs offer advantages in modeling complex data distributions, simulating causal effects, and handling nonparametric models (Goodfellow et al., 2014). They enable counterfactual inference (Grari, Lamprier & Detyniecki, 2023). However, GANs suffer from mode collapse, vanishing gradients, and inadequate evaluation metrics (Xu et al., 2018).

3. Constructing causal inference mechanism based on Markov chains

Causal inference optimization using GANs begins with applying Markov chains for causal relationship analysis. This study references Janzing and Schölkopf (2010) for Markov chain integration. Optimization of data log-likelihood, inspired by Brown et al. (2012), is to address Learning Markov kernels. For structural loss in the causal structure graph, Kalinathan et al. (2022) provide a regularization term method. Parameter loss evaluation adheres to Neyshabur et al. (2017), controlling causal mechanism complexity using the Frobenius norm.

Markov chains-based causal mechanism. Janzing and Schölkopf (2010) integrate Markov kernels and causality mechanisms, learning d Markov kernels. Each j-th Markov kernel q_j, represents the conditional density of x_j, with candidate parents denoted as $\hat{g}$. The learning process involves optimizing data likelihood based on the conditional distribution of $q_j\left(x_j \middle| x_{\mathrm{pa}(j;\hat{g})}\right)$, where $\mathrm{pa}(j;\hat{g})$ is an estimated cause of x_j, represented by a binary vector set. Additionally, $\hat{g}$ is specified to enforce sparsity and acyclicity. They designate each Markov kernel as a causal mechanism $\widehat{f_j}$defined as the below equation.

$$\widehat{\boldsymbol{X}}_{\boldsymbol{j}} = \widehat{f_j}\left(\left[a_j \odot X, E_j\right], \theta_j\right)$$

Where X represents a distinct sample comprising d variables, i.e., $X = (x_1, \cdots, x_d)$. The j-th Markov kernel, represented as $a_j = (a_1, j, \cdots, a_d, j)$ with a binary vector. If the coefficient $a_i$in a_i, j is 1, it signifies that variable x_i in sample X can generate x_j, and edges $x_i \rightarrow x_j$ belong to the graph $\hat{g}$. θ_j denotes the parameters used in calculating $\widehat{f_j}$, such as the weights in neural networks. E_j encompasses all noise variables that lack observed causal relationships with $x_j$and are independent. Derived from the causal mechanism $\widehat{f_j}$, the estimated cause based on x_j is defined as $q_j\left(x_j \middle| x_{-\mathrm{j}}, a_j, \theta_j\right)$, and it can be simplified to $q_j\left(x_j \middle| x_{\mathrm{pa}(j;\hat{g})}, \theta_j\right)$.

Learning Markov kernels. According to Brown et al. (2012), learning the causal mechanism $\widehat{f_j}$ involves both learning the mechanism itself and its minimal set of parent nodes. By minimizing the conditional log-likelihood of the data, we evaluate and optimize the model to make its predicted distribution closely match the true data distribution. Specifically, it compares the predicted distribution of the model (i.e., q_j, the distribution of x_j predicted based on model parameters θ_j and parent nodes $x_{\mathrm{pa}(j;\hat{g})}$ with the true data distribution. Additionally, to more accurately identify the true causal graph structure, researchers introduce an extra term $\hat{I}^n$, which is used to recognize the Markov equivalence class of the true causal graph g. During optimization, constant terms (cst) aid in stabilizing computations without altering the gradient. In summary, learning Markov kernels aims to accurately predict and explain causal relationships in data through optimizing a scoring function.

Assessing the structural loss of the generated causal structure graph. When calculating structural loss of generating causal structure graphs, identifying the minimum subset of $\mathrm{pa}(j; \hat{g})$ corresponds to a feature selection problem. The challenge lies in the Markov Blanket of x_j converging to the true causal graph under the large sample limit. Kalainathan et al. (2022) tackle this by optimizing a log-likelihood function with a regularization term λ_S, which penalizes the number of parent variables for each node. This hyperparameter λ_S, acts as a tuning knob to control the sparsity of the selected Markov Blanket, effectively balancing between model simplicity and explanatory power.

Evaluating the parameter loss to conform to the Markov equivalence class. To achieve model identification within the Markov equivalence class of the true directed acyclic graphs, Neyshabur et al. (2017) and Kalainathan et al. (2022) introduce an additional layer of complexity control. They leverage the Frobenius norm of the model parameters λ_F as a regularization mechanism. This approach restricts the complexity of the causal mechanisms, preventing overfitting and promoting generalizability. By combining data-fitting terms with this regularization, they define a parametric loss function that guides the model towards identifying a parsimonious yet accurate representation of the true causal relationships.

4. Optimizing causal inference model evaluation using MMD algorithm

Assessing the stability of causal inference results involves considering factors like model collapse, training instability, vanishing gradients, and data quality, impacting the accuracy of generated outcomes (Radford, Metz & Chintala, 2015). GANs evaluation metrics, typically used for stability assessment (Spirtes & Zhang, 2016), include MMD (Gretton et al., 2012). MMD quantifies the distinction between distributions of two given samples, making it effective across various models compared to other GANs evaluation metrics like Inception score, Wasserstein distance, Frechet inception distance, and 1-Nearest Neighbor (Xu et al., 2018).

$$MMD\lfloor p_x, p_y \rfloor = \left\| \frac{1}{n^2}\sum_{i}^{n}\sum_{i'}^{n} k(x_i, x_{i'}) + \frac{1}{m^2}\sum_{j}^{m}\sum_{j'}^{m} k(y_j, y_{j'}) - \frac{2}{nm}\sum_{i}^{n}\sum_{i'}^{m} k(x_i, y_{j,}) \right\|_H$$

p_x and $p_y$represent the probability distributions of samples x and y, respectively, while $k(x_i, x_{i'})$ and $k(y_j, y_{j'})$ are kernel functions. These kernel functions compute the average similarity between all pairs of data points in the sets x and y. The paired kernel evaluations are then summed and averaged to quantify the self-similarity within these sets. Additionally, $k(x_i, y_{j,})$ calculates the cross-similarity between data points in sets x and y, while $\| \quad \|_H$ signifies the norm in the Hilbert space. This norm measurement is instrumental in gauging the distance between distributions within the high-dimensional feature space. Generally, $0 \leq \mathrm{MMD}\lfloor p_x, p_y \rfloor < \infty$. When MMD is close to zero, it indicates a high similarity between p_x and p_y. Conversely, larger values signify greater disparities between the distributions.

5. Experiment

5.1 Identifying model for shallow reading behavior

The shallow reading model is based on three primary behavioral characteristics established in previous research (Bai et al., 2024): reading speed, rapid page jumping, and Nonsentence reading. Firstly, a benchmark of over 400 words per minute identifies fast reading behavior associated with shallow reading. Secondly, rapid page jumping, indicating quick page turns without prioritizing effective memory or deep understanding, is another notable characteristic, assessed using an 8-second benchmark to detect ineffective retention of information. Lastly, Nonsentence reading, representing fragmented reading, involves a preference for short word counts. Shallow readers often concentrate on phrases and individual words rather than complete sentences. Moreover, fragmented reading places relatively low importance on content, quantified using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to assess material significance.

Identification begins with an 8-second threshold to detect page-jumping behavior. If this criterion isn't met, the next assessment considers sentence-level reading behavior. For Nonsentence reading, the number of words in the shortest meaningful sentence is calculated, excluding auxiliary details. This determines if reading behavior prioritizes sentences, words, or letters, indicating shallow reading. Fast reading within shallow reading is identified if speeds meet or exceed 400 words per minute, categorized as fragmentary reading otherwise. Within sentence-level reading behavior, speeds exceeding 400 words per minute are considered fast reading in the context of shallow reading, while lower speeds are non-shallow reading.

5.2 Data Collection Tool and Strategy for Shallow Reading Behavior

Our study utilizes an ebook system developed by our laboratory as the data collection tool. This online reading platform provides various functionalities such as prev, next, underline, highlight, and memo, logging these actions (Zhao et al., 2023). It quantitatively assesses shallow reading by extracting reading content and measuring duration. To collect data on shallow reading behavior, a reading strategy was devised. Students are instructed to use the underline or highlight buttons when reading, initiating the action at the start and ending it upon finishing, facilitating data collection for identifying shallow reading behavior.

A 3-week learning experience was incorporated into a university course for 51 graduate students. Prior to the course, students received guidance on activities, reading techniques, and ebook usage. Week 1 featured a 30-minute ebook system practice. In Weeks 2 and 3, participants read a game-based research paper on the ebook, applying advised reading strategies. Afterward, they completed a 30-minute test with multiple-choice and short-answer questions. This endeavor yielded 10,694 data pieces including user-specific data like ID and Page, shallow reading behavior characteristics, corresponding labels such as reading speed (RS) and Page jumping (PJ), data types including Word number of Non-sentence (WNNS), TF-IDF, Importance of Non-sentence (INS), and learning performance scores.

6. Construction, evaluation, and analysis of GANs causality models

6.1 GANs generator and discriminator

PyTorch served as the cornerstone framework for GANs implementation, bolstered by GPU resources from Google Colaboratory for enhanced training efficiency. The generator network comprises a Linear 3D layer fusing input and noise data for linear combinations, a ChannelBatchNorm 1d for normalizing data to avoid “elimination singularities” and promote stability, a Tanh activation for introducing nonlinearity, and another Linear 3D layer for final data transformation. Structural loss is evaluated by calculating causal mechanism complexity. The discriminator, designed for mixed data (real and generated), follows a sequence of Linear layers, 1D batch normalization for stability and faster convergence, and LeakyReLU activation to mitigate gradient issues. This structure assesses each sample's authenticity, outputting a probability array per variable, effectively differentiating between genuine and fake data.

6.2 Analysis of the GANs model performance

The primary concern addressed here is the instability observed in the results of the generated adversarial neural network. The GANs model exhibits a consistent trend in structural loss, indicating that the generated causal relationships align with the DAG and accurately represent the influence of causal links between elements. However, the generator and discriminator loss exhibit significant fluctuations. This variability points to instability during the training process, which, in turn, affects the stability of the causal results. This outcome can be attributed to the delicate balance inherent in GAN models between the Generator and Discriminator components. When one dominates the other, causing slow convergence, it can result in instability (Goodfellow et al., 2014). Additionally, the causal relationships derived from the GANs-based causal exploration method are probabilistic in nature, further contributing to result instability as the model undergoes learning iteration.

6.3 Analysis of the causal inference results

MMD results of the 16 causal outcomes show values ranging from 0.01 to 0.34, indicating overall small dissimilarities. Except for Results 1 and 7, with MMD values exceeding 0.1, the remaining results fall within 0 to 0.1, indicating low dissimilarity. This suggests a high degree of similarity among the 16 causal relationship results. Despite slight deviations in the two results, the overall MMD value remains small, indicating model-generated outcomes' instability. Multiple calculations are averaged to ensure acceptable stability. Following Kalainathan et al.'s methodology (2022), who ran the model 8 times and averaged the results, this study learned the model 15 times, computing the average of all outcomes (Gawa, 2020)

Table 1. *Probability of causal relationship between variables in shallow reading behavior*

	RS	PJ	WNNS	INS	Score
RS	0	0.67	0.37	0.15	0.07
PJ	0.32	0	0.63	0.61	0.26
WNNS	0.25	0.52	0	0.94	0.16
INS	0.33	0.49	0.27	0	0.68
Score	0.29	0.35	0.13	0.18	0

Table 1 sets a probability threshold of 0.6 to determine causal relationships. Values exceeding this threshold indicate a higher likelihood of causation. By aggregating and averaging causal probability values across 16 runs, the causal matrix relationship values are derived. These conclusive results encompass shallow reading behavior characteristics such as RS, PJ, WNNS, INS, and Score. The result highlights two significant causal relationships: Firstly, a causal link exists between shallow reading behavior and academic performance, with a probability of 0.68, indicating shallow reading behavior's influence. Specifically, “Nonsentence behavior” directly impacts academic performance. Secondly, causal relationships exist among various behavioral aspects of shallow reading. RS is the causal factor for PJ with a probability of 0.67. Additionally, PJ affects “Nonsentence behavior”, including WNNS and INS, with probabilities of 0.63 and 0.61, respectively. Notably, INS within “Nonsentence behavior” is the causal factor for WNNS, with a probability of 0.94.

7. Discussion and conclusion

GANs offer a novel approach to deep learning in causal inference, presenting challenges in model design and evaluation. Integrating Markov chains into neural networks establishes causality probabilistically, contributing to inference result instability. Furthermore, GANs-based causal inference relies on probability-based inferences, rendering traditional evaluation metrics inadequate. Notably, existing evaluation metrics, rooted in image processing, predominantly focus on image realism and pixel richness. Shallow reading behavior, often categorized negatively, especially in digital learning, is neutrally viewed as a strategy for quick information retrieval. Historical mediums like newspapers and television were also considered shallow reading. The correlation between reading speed and shallow reading has long intrigued researchers, with fast reading being a prominent characteristic.

GANs-based causal analysis faces output instability due to issues like poor data quality, model collapse, and vanishing gradients. Our study employs ChannelBatchNorm 1d and Tanh for feature extraction and introduces global penalized min-max optimization in SAM training to stabilize models. Assessing causal inference outcomes, absent true causal samples, is tricky.

We mitigate this by averaging 16 results and utilizing MMD to measure causal structure differences, ensuring minimal result distribution variation for more reliable estimates.

Causal relationships between shallow reading and learning performance are confirmed, with reading speed identified as the fundamental cause. Reading speed causally influences other characteristics like page jumping and scanning. Clarifying these relationships sheds light on the formation mechanism of shallow reading behavior and its impact on learning performance, providing insights for educational interventions. Reading speed alone does not directly affect learning performance. And content engagement plays a crucial role in the learning process.

References

Bai, Y., Zhao, F., Kumamoto, E., & Yin, C. (2024). Precisely identifying and mitigating shallow reading behaviors through online reading data: exploring quantitative reading behavior methods and precise behavioral intervention. *Interactive Learning Environments*, 1-15.

Brown, G., Pocock, A., Zhao, M. J., & Luján, M. (2012). Conditional likelihood maximization: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, *13*, 27-66.

Carr, N. (2020). The shallows: What the Internet is doing to our Brains. WW Norton & Company.

Gawa, Y. O. (2020). Learn while creating! Advanced deep learning with PyTorch. Mynavi.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. (2014). Generative adversarial nets. In Z. Ghahramani and M. Welling and C. Cortes and N. Lawrence and K.Q (Eds.), *Advances in Neural Information Processing Systems*. Weinberger.

Gordon, C. (2023). Reading literature in/against the digital age: Shallow assumptions, deep problems, expectant pedagogies. *Convergence*, *29*(1), 28-46.

Grari, V., Lamprier, S., & Detyniecki, M. (2023). Adversarial learning for counterfactual fairness. *Machine Learning*, *112*(3), 741-763.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, *13*(1), 723-773.

Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). Understanding the role of digital technologies in education: A review. Sustainable Operations and Computers, *3*, 275-285.

Janzing, D., & Schölkopf, B. (2010). Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, *56*(10), 5168-5194.

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., & Sebag, M. (2022). Structural agnostic modeling: Adversarial learning of causal graphs. *The Journal of Machine Learning Research*, *23*(1), 9831-9892.

Livingstone, D. J., Manallack, D. T., & Tetko, I. V. (1997). Data modelling with neural networks: Advantages and limitations. *Computer-aided molecular design*, *11*, 135-142.

Nauta, M., Bucur, D., & Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, *1*(1), 312-340.

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. *Neural Information Processing Systems*, *30*, 5947–5956.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*. 1-16

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10(3), e1355.

Sarris, M. (2022). Learning to read in a shallow orthography: the effect of letter knowledge acquisition. *International Journal of Early Years Education*, *30*(4), 661-678.

Spirtes, P. & Zhang, K. (2016) Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics, 3*(3), 1-28.

Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., F. Wu, & Weinberger, K., (2018). An empirical study on evaluation metrics of generative adversarial networks, *arXiv:1806.07755*, 1-14

Zhao, F., Hwang, G. J., & Yin, C. (2021). A result confirmation-based learning behavior analysis framework for exploring the hidden reasons behind patterns and strategies. *Educational Technology & Society*, *24*(1), 138-151.

Zhao, F., Liu, G. Z., Zhou, J., & Yin, C. (2023). A Learning Analytics Framework Based on Human-Centered Artificial Intelligence for Identifying the Optimal Learning Strategy to Intervene in Learning Behavior. *Educational Technology & Society*, *26*(1), 132-146