Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24519

Disentangled Representation Learning and its Application in Clinical Diagnosis

Yuhao Cui

Dundee International Institute of Central of University, Changsha, China
2617367@dundee.ac.uk

Abstract. Over the past few decades, with the skyrocketing development of GPU by NVIDA, the speed of computer calculations of different kinds of deep learning models with millions or even billions of parameters based on high-speed calculation like Convolutional Neural Network (CNN) and Transformer have come out and been widely applied in different fields like natural language processing (NLP) and facial and voice recognition. However, despite these models being capable of processing amounts of data and producing the result, their black box working process still obstructs them from having a better ability of generalization. Also, due to the lack of proper interpretability and explainability, it is hard for people to trust these results when it comes to certain situations like legislation and medical contexts where the interpretability of data is of great significance. To solve this problem, Disentangled Representation Learning (DRL) comes out with explainable representations. This review mainly investigates different optimization models of disentangled representation learning, its application in different specific clinical diagnosis and future directions and prospects.

Keywords: Disentangled representation learning, Explainability, Clinical diagnosis

1. Introduction

Currently, traditional deep learning models like artificial neural network (ANN) and convolutional neural network (CNN) have already been applied in medical contexts, including clinical diagnosis by image processing or data analyses [1,2]. For example, breast cancer tumors classification by either images or data. However, although these models show great efficiency in diagnosis, intrinsically they just find a distribution or discrimination criteria to fit the data best, and their working process remains a black box that nobody could understand because of the high-dimensional data and complex correlations [3]. In light of this, many researchers are dedicated to designing better structures based on original models with more complex parameters in hopes to increase the model explainability and interpretability, but still limited [3]. Some other researchers consider designing a model to extract features from the input with semantic meanings, which brings about disentangled representation learning (DRL)[4].

Disentangled representation learning (DRL) is a deep learning model identifying and separating underlying factors from the image with semantic meaning and then encoding them. Based on the basic assumption that each extracted factor is independent from each other, the model just learns

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24519

how the change of a certain factor in latent space can in turn affect the certain property of the real object and just study such relations. So DRL imitates the way humans think, first analyze the intrinsic properties disentangled from an object, and then do judgement through a series of analysis of the properties [4]. By doing this, it is easier for researchers to ensure the model's robustness and generalization since researchers could understand the detailed process and make adjustment on the parameters in turn. Moreover, in medical contexts, it provides doctors and patients with specific explanations of diagnosis, which may greatly improve the trust of people and broader utilization of the use of AI [5].

This review explores several methods of DRL, including different optimization models and application in clinical diagnosis. By analyzing these approaches, the review attempts to not only find the current advances of DRL methods on the structure of representations, but also limitations including scalability, evaluation metrics, and adaptability to medical contexts. Also, this review proposes potential future directions for enhancing the clinical utility of DRL, aiming to bridge the gap between state-of-the-art machine learning research and practical medical applications.

2. Literature review

Recent advances in AI have significantly transformed clinical diagnosis across medical specialties. Deep learning models now demonstrate expert-level performance in medical image analysis, including detection and classification of various pathologies [2]. Beyond imaging, AI systems increasingly integrate multimodal clinical data to support diagnostic decision-making and risk prediction. While these developments show great prospect, persistent challenges in model interpretability and clinical integration still remain, which have driven growing interest in disentangled representation learning (DRL) approaches. Disentangled representation learning has evolved through various approaches that can be categorized along several dimensions: the types of models used for training, the structure of the learned representations, the level of supervision required, and whether they rely on strict independence assumptions between factors [3] . Recent advances have particularly focused on addressing challenges specific to medical applications, such as handling incomplete or inconsistent data modalities and providing clinically relevant interpretations [6]. Below, the author examines three representative approaches that demonstrate the potential of DRL in clinical settings.

2.1. Factorized Diffusion Autoencoder (FADE)

The Factorized Diffusion Autoencoder (FDAE) represents a significant advancement in unsupervised disentangled representation learning for medical image analysis [7]. A structure proposed by Wu and Zheng introduces a content-mask factorization network (CMFNet) that decomposes input images into two complementary components: content codes that capture semantic information in latent space, and masks that represent spatial information [8]. This factorization is particularly valuable in medical imaging, where both the "what" (pathological findings) and the "where" (anatomical location) are crucial for accurate diagnosis [8].

The FDAE architecture consists of three main components: (1) an image encoder that extracts latent features, (2) a mask decoder that generates spatial attention maps, and (3) a conditional diffusion model that reconstructs the input image from the disentangled representations. The model employs two novel loss functions to enhance disentanglement: a content decorrelation loss that encourages independence between different content factors in latent space, and a mask entropy loss that promotes spatial separation of different concepts in the image plane [8].

DOI: 10.54254/2755-2721/2025.PO24519

In medical applications, this approach offers several advantages. The explicit separation of content and spatial information allows radiologists to understand whether the model's decisions are based on the correct anatomical locations. For instance, in lung nodule detection (de Margerie-Mellon, C. and Chassagnon, G., 2023), the model might identify malignant features (content) and correctly localize them in the lung periphery rather than the mediastinum [9]. Moreover, diffusion-based reconstruction provides a natural mechanism for handling noisy or incomplete medical images, a common challenge in clinical practice.

2.2. Disentangled Representation Learning Framework for Size Generalization (DISGEN)

The DISGEN framework addresses a critical challenge in medical image analysis: the variation in anatomical sizes across patients [10]. Traditional deep learning models often struggle with size variations, requiring extensive data augmentation or resizing that can distort important diagnostic features. DISGEN creates representations that stay the same regardless of size or task, clearly distinguishing size-related changes from other important clinical features. The key innovation of DISGEN lies in its dual-branch architecture that processes input images through both size-specific and size-invariant pathways. The size-specific branch captures scale information, while the size-invariant branch extracts features relevant to the diagnostic task regardless of anatomical dimensions. These representations are then combined through an adaptive fusion module that maintains the relative importance of size-dependent and size-independent features [10].

This approach has shown promise in pediatric imaging, where normal anatomical sizes vary dramatically with age. For example, in congenital heart disease assessment, DISGEN can separately represent the absolute size of cardiac structures (which varies with patient age) from the relative proportions and morphological features that indicate pathology [11]. An adaptive fusion module then combines these representations, preserving the relative importance of size-dependent and size-independent features [10].

2.3. DrFuse architecture

Clinical decision-making often relies on multiple data modalities including electronic health records (EHR), medical images and lab results [12]. However, in practice, some modalities may be missing or inconsistent. DrFuse addresses these challenges through a novel approach to multi-modal fusion that combines disentangled representation learning with dynamic modality weighting [6].

The DrFuse architecture consists of three key components: (1) modality-specific encoders that extract both shared and distinct representations from each data type, (2) a shared representation alignment module that identifies common information across modalities, and (3) a disease-aware attention fusion mechanism that dynamically weights modalities based on their relevance to specific diagnostic tasks [6].

A particularly innovative aspect of DrFuse is its handling of missing modalities. Rather than inputting missing data or training separate models for different modality combinations, DrFuse leverages the disentangled shared representations—when one modality is missing, the information it typically contributes can be partially inferred from the shared representations learned from other modalities [6]. The disease-aware attention mechanism further enhances clinical relevance by adjusting modality importance based on both the patient's condition and the specific diagnostic question. For instance, in pneumonia diagnosis, chest X-rays might receive higher weight for typical cases, while EHR data might be more heavily considered for immunocompromised patients with atypical presentations [13].

DOI: 10.54254/2755-2721/2025.PO24519

3. Future directions

While disentangled representation learning shows tremendous promise for clinical applications, several important challenges and opportunities for future research emerge from current work. These directions reflect both technical limitations that need addressing and promising opportunities to expand the clinical utility of DRL approaches.

3.1. Multimodal integration enhancement

Current multimodal approaches like DrFuse represent important first steps, but more sophisticated methods are needed to handle the complex relationships between diverse clinical data types [6]. Future work should explore hierarchical disentanglement that captures relationships between modalities at different levels of abstraction [14]. For example, genomic data might influence the interpretation of imaging findings differently than vital signs do [15]. Developing frameworks that can learn these relationships automatically while maintaining interpretability will be crucial for comprehensive clinical decision support systems.

3.2. Dynamic disentanglement for temporal data

Many medical conditions evolve over time, and their representations may need to adapt accordingly. Current DRL methods typically treat each time point independently, missing opportunities to disentangle persistent patient factors from transient states [16]. Future methods could incorporate temporal dynamics explicitly, perhaps through recurrent or attention-based architectures that maintain disentanglement across time series. This would be particularly valuable for monitoring chronic conditions or tracking treatment responses.

3.3. Causal disentanglement

Moving beyond statistical independence to discover causal relationships between factors could significantly enhance the clinical relevance of DRL [17]. By identifying which factors causally influence others, models could provide more actionable insights and better support intervention planning. Integrating causal discovery methods with disentangled representation learning remains an open challenge with high potential impact.

3.4. Standardized clinical evaluation

As noted by Janiesch et al., the lack of standardized benchmarks for interpretability methods makes comparative evaluation difficult [18]. The field would benefit from developing clinical-specific evaluation protocols that assess both the quality of disentanglement and the clinical utility of the representations. This might include metrics for physician trust, decision-making accuracy with and without explanations, and time efficiency in clinical workflows [3].

Addressing these challenges will require close collaboration between machine learning researchers and clinical experts to ensure that technical advances translate into genuine clinical value. As disentangled representation learning matures, its ability to provide interpretable, robust, and clinically meaningful insights positions it as a key technology for the next generation of medical AI systems.

DOI: 10.54254/2755-2721/2025.PO24519

4. Conclusion

Disentangled representation learning represents a significant step forward in developing interpreted AI systems for clinical diagnosis. By explicitly separating the underlying factors that contribute to medical observations, DRL provides a pathway to overcome the "black box" limitations of conventional deep learning approaches. The methods discussed in this review—FDAE's content-mask factorization, DISGEN's size-invariant representations, and DrFuse's dynamic multimodal fusion—demonstrate how tailored disentanglement approaches can address specific challenges in medical applications. The clinical benefits of these approaches are manifold. They provide interpretable explanations that build clinician trust, enable more robust performance across diverse patient populations. By continuing to focus on both technical innovation and clinical relevance, researchers can develop systems that not only achieve high performance but also enhance transparency, trust, and ultimately patient care. As the field progresses, disentangled representation learning is poised to play a pivotal role in bridging the gap between powerful AI capabilities and the nuanced requirements of medical diagnosis. But still, as outlined in the future directions, significant work remains to fully realize the potential of DRL in medicine. Current DRL faces challenges including the optimization of disentanglement strategies, lack of reliable evaluation metrics, and instability in learning consistent representations. While theoretically appealing, learned disentangled features often fail to enhance performance in practical applications, especially with complex real-world data where factors are naturally correlated. These issues highlight the gap between DRL's promise and its current practical utility.

References

[1] Amato, F., López, A., Peña-Méndez, E.M., Vaňhara, P., Hampl, A. and Havel, J. (2013). Artificial neural networks in medical diagnosis. Journal of applied biomedicine, 11(2), pp.47-58.

[2] Arena, P., Basile, A., Bucolo, M. and Fortuna, L. (2003). Image processing for medical diagnosis using CNN. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 497(1), pp.174-178.

[3] Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T. and Zhang, S. (2020). CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE transactions on medical imaging, 40(2), pp.699-711.

[4] Xin Wang., Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu, l. (2024). 'Disentangled Representation Learning', IEEE transactions on pattern analysis and machine intelligence, 46(12), pp. 9677–9696. Available at: https: //doi.org/10.1109/TPAMI.2024.3420937.

[5] Zai-Dao, W., Jia-Rui, W., Xiao-Xu, W. and Quan, P. (2022). A review of disentangled representation learning. Acta Automatica Sinica, 48(2), pp.351-374.

[6] Yao, W. et al. (2024) 'DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency', in Proceedings of the ... AAAI Conference on Artificial Intelligence, pp. 16416–16424. Available at: https: //doi.org/10.1609/aaai.v38i15.29578.

[7] Yuan, X. and Maire, M. (2024). Factorized diffusion architectures for unsupervised image generation and segmentation. Advances in Neural Information Processing Systems, 37, pp.68686-68712.

[8] Wu, A. and Zheng, W.-S. (2024) 'Factorized Diffusion Autoencoder for Unsupervised Disentangled Representation Learning', in Proceedings of the ... AAAI Conference on Artificial Intelligence, pp. 5930–5939. Available at: https: //doi.org/10.1609/aaai.v38i6.28407.

[9] de Margerie-Mellon, C. and Chassagnon, G. (2023). Artificial intelligence: A critical review of applications for lung nodule and lung cancer. Diagnostic and interventional imaging, 104(1), pp.11-17.

[10] Huang, Z., Yang, Q., Zhou, D. and Yan, Y. (2024). Enhancing Size Generalization in Graph Neural Networks through Disentangled Representation Learning. [online] arXiv.org. Available at: https: //arxiv.org/abs/2406.04601 (Accessed 13 Apr. 2025).

[11] Holt, D.B., El-Bokl, A., Stromberg, D. and Taylor, M.D. (2025). Role of Artificial Intelligence in Congenital Heart Disease and Interventions. Journal of the Society for Cardiovascular Angiography & Interventions, 4(3), p.102567.

DOI: 10.54254/2755-2721/2025.PO24519

[12] Gaw, N., Schwedt, T.J., Chong, C.D., Wu, T. and Li, J. (2018). A clinical decision support system using multi-modality imaging data for disease diagnosis. IISE Transactions on Healthcare Systems Engineering, 8(1), pp.36-46.
[13] Ayan, E. and Ünver, H.M. (2019). Diagnosis of pneumonia from chest X-ray images using deep learning. In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-5). Ieee.
[14] Qiao, S., Wang, R., Shan, S. and Chen, X. (2023). Hierarchical disentangling network for object representation learning. Pattern Recognition, 140, p.109539.
[15] Baek, K.W., Park, J.J., Gim, J.A., Baek, K.W., Park, J.J. and Gim, J.A. (2021). Machine Learning Application in Genomic, Exercise, and Vital Datasets. Exercise Science, 30(2), pp.131-146.
[16] Yao, W., Chen, G. and Zhang, K., (2022). Temporally disentangled representation learning. arXiv preprint arXiv: 2210.13647.
[17] Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z. and Zhang, T. (2020). Disentangled generative causal representation learning.
[18] Janiesch, C., Zschech, P. and Heinrich, K. (2021). 'Machine learning and deep learning', Electronic markets, 31(3), pp. 685–695. Available at: https: //doi.org/10.1007/s12525-021-00475-2.