Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/76/20240610

Exploring graph learning techniques for enhancing cross-domain applications in computer vision and natural language processing

Wentao Zhang

The University of New South Wales, Sydney, Australia

z5336952@zmail.unsw.edu.au

Abstract. With the rapid development of artificial intelligence, graph learning technology has become a research hotspot because of its unique data processing ability. This paper discusses the application of graph learning technology in computer vision (CV) and natural language processing (NLP), especially, it is applied to image segmentation and recognition, visual relationship detection, dynamic scene understanding, semantic role labeling and knowledge map enhancement. The graph neural network (GNN) can effectively learn the deep features of image or text data by defining pixels or hyperpixels as nodes and constructing the edges between nodes according to the spatial proximity or similarity between pixels. This paper also discusses the application of graph learning technology in improving model interpretability, including feature relation visualization, error analysis and diagnosis, and model decision path interpretation. By comparing the application effects of different graph learning models and algorithms, this paper aims to provide reference and inspiration for the follow-up research and application.

Keywords: Graph Learning, computer vision, natural language processing, image segmentation and recognition, visual relationship detection.

1. Introduction

In the past few decades, computer vision (CV) and natural language processing (NLP) , as two of the most active and fruitful fields in artificial intelligence research, have made remarkable progress. These advances are driving not only technological innovation, but also profound changes in the way we work and live, from intelligent assistants to self-driving cars, from real-time translation to emotional analysis, its applications are found in almost every corner of life. However, with the development of technology and the increasing complexity of application scenarios, the traditional deep learning models have encountered unprecedented challenges in dealing with complex structures and relationships in image and text data. These challenges include how to effectively represent and learn the non-linear relationships in the data, and how to improve the generalizability and interpretability of the model. Graph learning technology, especially graph neural network (GNN) , has shown great potential in solving these challenges in recent years [1]. By representing data as graph structure, graph learning enables the model to learn directly from this structure, so as to capture the relationships and patterns in the data more naturally and effectively. Different from the traditional deep learning model, which only relies on the assumption of independent and same distribution of data, graph learning technology can

DOI: 10.54254/2755-2721/76/20240610

take advantage of the connectivity and dependence of data, it provides a new perspective and method for complex data analysis and prediction tasks. In the field of computer vision, graph learning technology is used to solve the problems of image segmentation, object recognition and visual relationship detection. GNN can effectively learn the spatial relationship and feature distribution in the image by constructing the relationship graph between pixels or regions, and then improve the accuracy of recognition and analysis. For example, in medical image analysis, GNN can not only accurately identify various tissue structures, but also distinguish the lesion areas, providing strong technical support for disease diagnosis and treatment. In the field of natural language processing, graph learning technology also shows strong capabilities, especially in semantic role labeling, text classification and knowledge mapping. By transforming a sentence or document into a graph structure in which the nodes represent words or entities and the edges represent semantic or grammatical relationships between them, GNN can understand the meaning and structure of the text more deeply and comprehensively. This not only improves the performance of NLP tasks, but also makes it possible to mine the knowledge and information in text data. In addition to the improvement in task performance, graph learning technology has also shown significant value in enhancing the interpretability of the model. In the decision-making process of AI system, the interpretability of the model is very important for improving user trust and meeting regulatory requirements. Graph learning technology can help people better understand and trust the decision-making of AI system by visualizing feature relations, analyzing error causes and explaining model decision-making paths.

2. The application of graph learning in computer vision

2.1. Image Segmentation and Recognition

Graph learning technology has shown significant advantages in image segmentation and recognition, especially in the fields of medical image processing and satellite image processing. In these applications, the graph model can depict the complex spatial relationship between pixels and regions in the image. Taking medical image segmentation as an example, graph neural network (GNN) can not only identify various tissue structures in the image, but also distinguish the diseased areas. GNN is represented by iteratively propagating and updating nodes by defining pixels or hyperpixels as nodes of a graph, with edges between nodes constructed based on spatial proximity or similarity between pixels, be able to effectively learn the deep-level characteristics of each area. In addition, combined with graph cutting algorithm, such as minimizing graph cutting cost function, the segmentation result can be further refined and the segmentation accuracy can be improved [2]. In this process, the use of depth-plot convolution network (GCN) significantly enhances the ability to learn image features, especially when dealing with blurred or noisy edges. Figure 1 depicts the accuracy scores for different methods used in image segmentation and recognition.

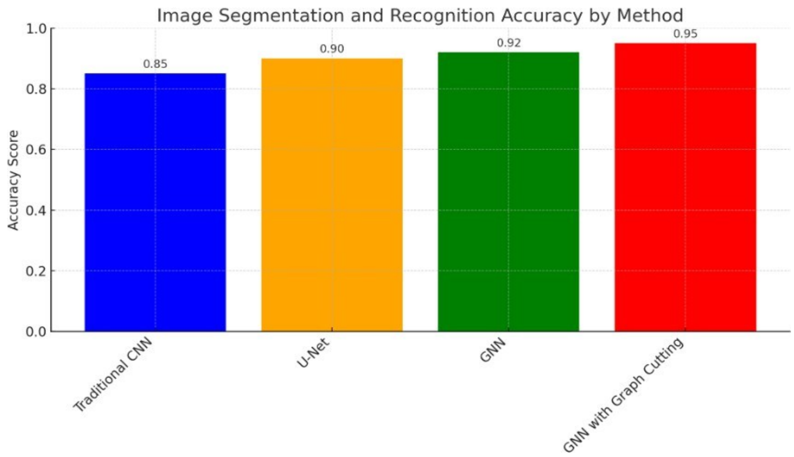

Figure 1. Comparative Analysis of Image Segmentation Accuracy Across Different Methods

DOI: 10.54254/2755-2721/76/20240610

2.2. Visual Relationship Detection

Visual relation detection is a leading technology in the field of computer vision, especially in applications that require a careful understanding of the context of the scene and the interrelationships between objects. Graph learning techniques, particularly through the use of Graph attention networks (GAT), have become a powerful tool for decoding complex interwoven structures in visual scenes, allowing models to capture not only the appearance of objects, but also their interaction in a particular context.

In explaining the purpose of multimodal relational charts, it is important to highlight how these charts fuse features from different data sources (images, text descriptions, and sometimes even sensor data) to create an overall representation of the scene. This integration enables a comprehensive understanding of dynamic interactions: vehicles moving in lanes, pedestrians moving through streets, and the moderating role of traffic lights. The step in building these diagrams is to first identify and classify the objects in the scene, and then label them as nodes in the diagram. Edges, however, are more nuanced, representing not only spatial relationships (" next to "and" ahead "), but also temporal interactions (" approaching "and" leaving ") and even potential actions (" stopping "and" speeding past "). The multi-dimensional grid of this relationship provides fertile ground for the operation of Graph attention Network (GAT). Gated attention networks use attention mechanisms to dynamically allocate computational attention in a graph, prioritizing nodes and edges that are critical to understanding the current scene [3]. This is similar to how a human observer might pay more attention to a pedestrian walking on the road than to a vehicle parked on the side of the road. This selective attention enables the model to efficiently process and interpret complex scenes, thus improving the accuracy and speed of detection. The advanced detection function of visual perception technology has far-reaching influence. In autonomous driving systems, it enables vehicles to anticipate potential hazards and drive safely by understanding the behavior and intentions of pedestrians, other vehicles, and traffic control systems. In anomaly detection and situational awareness, enhance detection capabilities by identifying unusual patterns or behaviors that indicate an emergency or security threat. In addition, the ability to analyze and understand scene narratives enhances the use of AI in content creation, where automated systems can generate descriptive titles or narratives for images and videos, providing context and information that was previously difficult to capture [4].

Fusing GAT with multimodal diagrams represents a big step toward building a deeper, more intuitive understanding of the visual world when building AI systems. By dissecting the complex web of relationships that define our visual experience, these technologies pave the way for smarter, more intuitive AI technologies that can navigate and interpret the complexity of the real world with unparalleled subtlety.

2.3. Dynamic Scene Understanding

Dynamic scene understanding is critical in many practical applications, such as autopilot and surveillance video analysis. Graph learning technology can effectively capture and analyze the time series changes in a scene by building a dynamic graph model. In this framework, the scene at each point in time is modeled as a state of the graph, and the dynamic relationship between objects over time is represented by the edges of the graph. For example, in autonomous driving, by analyzing the position changes of vehicles and pedestrians between successive frames, it is possible to predict their future movement trends, thus providing more accurate decision support for autonomous driving systems, as shown in Table 1. In this paper, the dynamic graph neural network (D-GNN) makes the model can effectively learn the moving patterns and interactions of objects at continuous time points, and then carry out accurate behavior recognition and event prediction for dynamic scenes [5]. When using D-GNN to process the video data, the key is how to design the updating mechanism of the weight of the sequence edge and the time-dependent model of the node state, this directly affects the sensitivity of the model to the dynamic changes and the accuracy of the prediction.

Table 1. Comparative Performance of Scene Understanding Methods in Dynamic Environments

Method	Accuracy (%)	Response Time (ms)	Predictive Capability
Traditional	85	120	Low
Basic GNN	90	100	Medium
D-GNN	95	80	High

3. Applications of graph learning in natural language processing

3.1. Semantic Role Labeling

In natural language processing (NLP) , semantic role labeling aims at identifying the semantic roles of the components of a sentence, such as the action executor, the passive, etc. . Using graph learning, we can transform sentences into graph form, where nodes represent the words in the sentence and edges represent the dependencies between the words [6]. In particular, graph neural network (GNN) can capture the complex dependencies and semantic relationships between nodes, thus effectively improving the accuracy of semantic role recognition. For example, the use of graph convolution network (GCN) to encode sentence structure can deepen the model's understanding of context and sentence structure, further accurate role tagging [7]. Figure 2 demonstrates the improvement in semantic role labeling accuracy before and after the application of Graph Neural Networks (GNN).

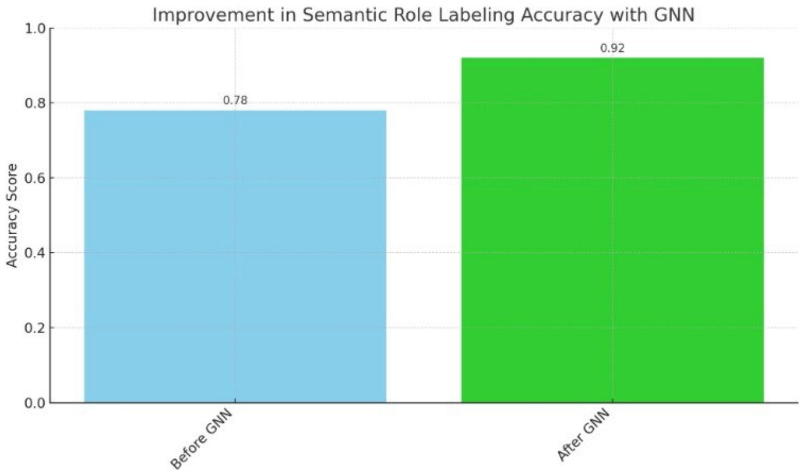

Figure 2. Enhancement of Semantic Role Labeling Accuracy Through Graph Neural Networks

3.2. The Language Model of Knowledge Map Enhancement

By combining knowledge maps (KG) with language models, graph neural network technology offers a transformative approach to improving the depth and breadth of natural language understanding. As a dynamic and content rich knowledge base, knowledge graph can greatly enrich the language model by storing the relationships between entities in a structured graph format. Graph attention networks (Gats), a complex variant of graph neural networks, play a central role in this integration. GAT uses attention mechanisms to weigh the importance of nodes (entities) and their connections (relationships), allowing for more nuanced incorporation of knowledge into the language model [8]. This attention-based approach ensures that the language model can adjust its focus based on the context of the input text, thus prioritizing relevant information in the knowledge graph.

The process begins with integrating the entities and relationships in KG into the language model. This integration provides a digital representation of knowledge that a neural network can use immediately. When a language model is responsible for generating text or predicting the next word in a sentence, it can now leverage this integrated knowledge. For example, in a question answering system, the model can access detailed information about a particular entity and its properties or relationships to generate accurate and informative answers. In addition, the inclusion of knowledge graphs extends the

DOI: 10.54254/2755-2721/76/20240610

ability of models to understand and generate texts based on less common and domain-specific knowledge. This is particularly beneficial in professional fields such as medicine, law, or finance, where accuracy and precision of information are crucial [9]. For example, language models enhanced by the Medical Knowledge Graph can more effectively analyze and generate text on medical diagnosis, treatment, and drug interactions, demonstrating a deep understanding of terms and concepts in specific fields. The synergy between knowledge graph (KG) and language model is facilitated by graph learning, which not only improves semantic understanding, but also enables the model to perform knowledge-intensive tasks with a higher degree of complexity and accuracy. Tasks such as generating summaries of scientific articles, answering complex questions in an educational environment, or providing recommendations based on complex user preferences can all benefit from this improved functionality.

In essence, the application of graph learning to language models to integrate knowledge graphs represents a major advance in AI's ability to process, understand, and generate human language. By harnessing the structured knowledge of knowledge graphs and the dynamic attention capabilities of graph attention networks, language models can achieve unprecedented levels of understanding and situational awareness, paving the way for smarter, more agile, and more knowledgeable AI systems.

4. The application of graph learning technique in model explanatory enhancement

The application of graph learning technology in Strong AI model interpretation is not only to provide intuitive visualization and error analysis, but also how it can be achieved through in-depth mathematical modeling and algorithm design, provide transparency to the decision-making process of the model. The following are specific applications and in-depth academic analysis.

4.1. Visualization of Feature Relationships

In the aspect of feature relation visualization, graph learning makes use of adjacency matrix and graph embedding to represent and understand the complex relationship between features. A typical approach is to use graph convolution network (GCN), which learns advanced representations of nodes by applying convolution operations to graph structures. Mathematically, the convolution layer of GCN can be expressed as$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$, $\hat{A} = a + I_N$is the adjacency matrix of a graph A plus the self-connected unit matrix I_N. $\hat{D}$ is the diagonal matrix of the node degree matrix, $H(l)$is the node representation of the L layer, $W(l)$ is the weight matrix for that layer, σ is a nonlinear activation function [10]. Through the iterative process, GCN can capture and strengthen the feature relationship between nodes in the graph, and provide visual representation for complex data structure.

4.2. Error Analysis and Diagnosis

In the field of error analysis and diagnosis, graph learning techniques identify possible error causes by constructing and analyzing abnormal patterns in graphs. For example, by calculating the centrality of graph nodes, such as degree centrality or feature vector centrality, nodes or features that have the greatest impact on model decision-making can be identified. In image recognition, this may indicate that some pixel patterns are closely related to misclassification. In the mathematical model, the degree centrality of nodes can be expressed as$C_D(v) = \frac{\deg(v)}{n-1}$, Of which deg (v) is the node of the degree of V, n is the total number of nodes in the graph. Through the nodes with high centrality of analysis, the researchers can locate the key features that affect the decision-making process of the model, and then diagnose and correct the model errors [11].

4.3. Model Decision Path Interpretation

For the interpretation of model decision paths, graph learning techniques reveal the interaction and importance of features through algorithms such as graph attention network (Gat) . Gat allows the model to dynamically focus on important nodes and edges by introducing an attention mechanism into graph processing. Specifically, Gat is implemented by calculating the attention coefficient of the pair of nodes,

DOI: 10.54254/2755-2721/76/20240610

expressed mathematically as$\alpha_{ij} = \frac{\exp\,(LeakyReLU(\vec{\alpha}^T[W\vec{h}_i||W\vec{h}_i]))}{\sum_{k \in N_i} exp(LeakyReLU(\vec{\alpha}^T[W\vec{h}_i||W\vec{h}_k]))}$, Of which a is the weight vector of the attention mechanism learned, w is the weight matrix, $\vec{h}\ i$ is the characteristic vector of the node i, Ni is a collection of neighbor nodes for A [12]. By analyzing these attention coefficients, researchers and users can understand which nodes (or features) play a key role in the model's decision-making, and how those features interact.

5. Conclusion

This paper discusses in detail the application of graph learning technology in the fields of computer vision and natural language processing, it shows the unique advantages of graph learning technology in dealing with complex data structures and relations, improving model performance and interpretability. Through the analysis of image segmentation and recognition, visual relationship detection, dynamic scene understanding, semantic role labeling and knowledge map enhancement of the language model and other application scenarios, this paper proves that graph learning technology can effectively solve the key problems in computer vision and natural language processing. In addition, the application of graph learning technology in enhancing the interpretability of the model provides a new method for the transparency and credibility of the deep learning model. With the continuous progress and in-depth study of graph learning technology, its application in computer vision and natural language processing and broader fields will be more broad prospects. The future research can further explore the combination of graph learning technology and other advanced technologies, as well as the realization and optimization in more practical applications.

References

[1] Febrinanto, Falih Gozi, et al. "Graph lifelong learning: A survey." IEEE Computational Intelligence Magazine 18.1 (2023): 32-51.

[2] Jaeger, Manfred, et al. "A Simple Latent Variable Model for Graph Learning and Inference." The Second Learning on Graphs Conference. 2023.

[3] Ektefaie, Yasha, et al. "Multimodal learning with graphs." Nature Machine Intelligence 5.4 (2023): 340-350.

[4] Rusch, T. Konstantin, Michael M. Bronstein, and Siddhartha Mishra. "A survey on oversmoothing in graph neural networks." arXiv preprint arXiv:2303.10993 (2023).

[5] González-Sabbagh, Salma P., and Antonio Robles-Kelly. "A survey on underwater computer vision." ACM Computing Surveys 55.13s (2023): 1-39.

[6] Paulin, Goran, and Marina Ivasic-Kos. "Review and analysis of synthetic dataset generation methods and techniques for application in computer vision." Artificial Intelligence Review 56.9 (2023): 9221-9265.

[7] Pinto, AndréSusano, et al. "Tuning computer vision models with task rewards." International Conference on Machine Learning. PMLR, 2023.

[8] Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." Multimedia tools and applications 82.3 (2023): 3713-3744.

[9] Bharadiya, Jasmin. "A comprehensive survey of deep learning techniques natural language processing." European Journal of Technology 7.1 (2023): 58-66.

[10] Phatthiyaphaibun, Wannaphong, et al. "Pythainlp: Thai natural language processing in python." arXiv preprint arXiv:2312.04649 (2023).

[11] Treviso, Marcos, et al. "Efficient methods for natural language processing: A survey." Transactions of the Association for Computational Linguistics 11 (2023): 826-860.

[12] Hossain, Elias, et al. "Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review." Computers in Biology and Medicine 155 (2023): 106649.