MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding

Giảng viên hướng dẫn: TS. Đỗ Thị Thanh Hà

Nguyễn Mạnh Linh

Nguyễn Đức Thịnh

***Tóm tắt nội dung*—Một lượng lớn các đồ thị hay mạng trong thực tế vốn dĩ không đồng nhất, có nhiều loại nút và nhiều loại quan hệ. Embedding đồ thị không đồng nhất là việc embed từ cấu trúc lớn và nhiều thông tin của đồ thị về biểu diễn nút trong không gian thấp chiều. Các mô hình đã tồn tại tường định nghĩa metapaths trong một đồ thị không đầu nhất để ghi lại các quan hệ và định hướng lựa chọn "hàng xóm". Tuy nhiên các mô hình này bỏ qua đặc trưng của từng nút mà tìm hiểu ngay lập tức các nút trên metapath hoặc chỉ xem xét một metapath. Để khắc phục ba giới hạn này, tác giả đề xuất một mô hình mới là *Metapath Aggregated Graph Neural Network* (MAGNN) để tăng tốc hiệu năng cuối cùng. Đặc biệt, MAGNN sử dụng ba thành phần chính, biến đổi nội dung của nút thành các thuộc tính đóng gói của nút đầu vào, tổng hợp intra-metapath để kết hợp các nút ngữ nghĩa trung gian và tổng hợp inter-metapath để kết hợp thông tin từ nhiều metapaths. Các thí nghiệm được thực hiện trên ba bộ dữ liệu đồ thị không đồng nhất trong thực tế để phân loại nút, phân cụm nút và dự đoán liên kết chỉ ra rằng MAGNN đạt được kết quả dự đoán chính xác hơn so với các mô hình state-of-the-art hiện tại .**

I. Introduction

Nhiều bộ dữ liệu thực tế được biểu diễn với cấu trúc dữ liệu đồ thị, trong đó các đối tượng và quan hệ giữa chúng được biểu diễn bằng các nút và cạnh. Các ví dụ bao gồm mạng xã hội [14, 29], hệ thống vật lý [2, 10], mạng giao thông [18, 34], mạng trích dẫn [1, 14, 16], hệ thống gợi ý [26, 35], đồ thị tri thức [3, 24], ... Bản chất non-Euclidean của đồ thị khiến chúng khó được mô hình hóa bằng các mô hình học máy truyền thống. Với tập hàng xóm của mỗi nút, không hề có thứ tự hoặc giới hạn về kích thước, Tuy nhiên, hầu hết các mô hình thống kê giả định rằng một đầu vào có thứ tự và kích thước cố định trong không gian Euclid. Do đó, sẽ thuận tiện nếu các nút có thể được biểu diễn bằng các vector thấp chiều trong không gian Euclid và từ đó có thể lấy làm đầu vào của mô hình học máy khác.

Các kĩ thuật embed đồ thị khác nhau được đề xuất cho cấu trúc dữ liệu đồ thị. LINE [25] sinh node embedding dựa vào các nút gần nhất và gần thứ 2. Các phương pháp dựa trên bước ngẫu nhiên (Random-walk) bao gồm DeepWalk [21], node2vec [13] và TADW [32] sinh dãy nút được sinh ra bởi các bước ngẫu nhiên đến một mô hình skip-gram [19] để học node embeddings. Với sự phát triển nhanh chóng của deep learning, mạng neuron đồ thị (Graph neural networks - GNNs) được đề xuất, mô hình học các biểu diễn đồ thị bằng việc sử dụng các lớp neuron được thiết kế đặc biệt. Spectral-based GNNs bao gồm ChebNet [8] và GCN [16] biểu diễn các toán tử tích chập đồ thị trong miền Fourier của một đồ thị đầy đủ. Các mô hình dựa trên spatial, bao gồm GraphSAGE [14], GAT [28] và nhiều biến thể khác [17, 34, 45], giải quyết các vấn đề xung quanh khả năng mở rộng và khả năng tổng quát hóa của các mô hình dựa trên phổ bằng cách biểu diễn các phép toán tích chập đồ thị trực tiếp trên miền đồ thị.

Mặc dù GNNs đã đạt được những kết quả tốt nhất trong nhiều bài toán, hầu hết các mô hình dựa trên GNN giả định rằng đầu vào là đồ thị đồng nhất với chỉ một loại nút và một loại cạnh. Hầu hết đồ thị trong thực tế bao gồm nhiều loại nút và cạnh tương ứng với các thuộc tính trong các không gian thuộc tính khác nhau. Ví dụ, một mạng đồng tác giả chứa ít nhất hai loại nút, cụ thể là các tác giả và bài báo. Các thuộc tính của tác giả có thể bao gồm nơi làm việc, trích dẫn và lĩnh vực nghiên cứu. Thuộc tính của bài báo bao gồm từ khóa, địa điểm, năm phát hành ... Tác giả gọi loại đồ thị này là *mạng thông tin không đồng nhất* (HINs) hoặc *đồ thị không đồng nhất.* Sự không đồng nhất trong cả cấu trúc biểu diễn và nội dung của nút khiến GNN gặp khó khăn trong việc mã hóa thông tin vào không gian vector thấp chiều.

Hầu hết các phương pháp nhúng đồ thị không đồng nhất hiện có dựa trên ý tưởng về metapaths. Một *metapath* là một trình tự có thứ tự của các loại nút và loại cạnh được xác định trên lược đồ mạng, mô tả mối quan hệ tổng hợp giữa các loại nút liên quan. Ví dụ một mạng với tác giả, bài báo và địa điểm, *Tác giả-Bài báo-Tác giả* (APA) và *Tác giả-Bài báo-Địa điểm-Bài báo-Tác giả* (APVPA) là các metapaths mô tả mối quan hệ khác nhau giữa các tác giả. Metapath APA tương ứng với hai đồng tác giả, trong khi APVPA tương ứng với hai tác giả xuất bản bài báo cùng địa điểm. Do đó, ta có thể xem metapath là khoảng cách gần bậc cao giữa 2 nút. Do GNNs truyền thống xử lý tất cả các nút như nhau, chúng không thể mô hình hóa cấu trúc phức tạp và thông tin có nghĩa trong đồ thị không đồng nhất.

Mặc dù các phương pháp nhúng dựa trên metapath này hoạt động tốt hơn phương pháp nhúng mạng truyền thống trên các bài toán khác nhau, chẳng hạn như phân loại nút và dự đoán liên kết, chúng vẫn có ít nhất một trong những hạn chế sau. (1) Mô hình không sử dụng đặc trưng về nội dung nút nên nó hiếm khi hoạt động tốt với các đồ thị không đồng nhất với nút có nội dung nhiều đặc trưng (ví dụ metapath2vec [9], ESim [22], HIN2vec [11], HERec [23]). (2) Mô hình loại bỏ tất cả

Identify applicable funding agency here. If none, delete this.

các nút trên metapath, chỉ xem xét 2 nút cuối dẫn đến mất thông tin (ví dụ HERec [23] and HAN [31]). (3) Mô hình dựa trên một metapath duy nhất để nhúng đồ thị không đồng nhất. Do đó, mô hình yêu cầu một quá trình chọn metapath thủ công và mất đi các đặc điểm của thông tin từ các metapaths khác dẫn đến hiệu năng không tối ưu (ví dụ metapath2vec [9]).

Để giải quyết các hạn chế này, tác giả đề xuất một mạng neuron tổng hợp metapath (*Metapath Aggregated Graph Neural Network - MAGNN*) mới cho đồ thị không đồng nhất.

II. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m^2” or “webers per square meter”, not “webers/m^2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm^3”, not “cc”.)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

D. LaTeX-Specific Advice

Please use “soft” (e.g., `\eqref{Eq}`) cross references instead of “hard” references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don’t use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you’ve discovered a new method of counting.

BibTeX does not work by magic. It doesn’t get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can’t read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it’s supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won’t be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word “data” is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited,

such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)

- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

Bảng I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Hình 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

TÀI LIỆU

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.