Vision Transformers aos LLM's Multimodais

Maria Eduarda Silva Borba
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
maria.borba@discente.ufg.br

Pedro Ribeiro Fernandes
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
pedro.fernandes@discente.ufg.br

Hugo Rodrigues Pessoni
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
hugorodriguespessoni@discente.ufg.br

André Martins Dantas
Universidade Federal de Goiás
Instituto de Informática (INF)
Goiânia, Brasil
andre.dantas@discente.ufg.br

***Resumo* — Este artigo explora a evolução dos modelos de linguagem multimodais, com foco principal na transição dos Recurrent Neural Networks (RNNs) para os atuais Large Language Models (LLMs) multimodais, destacando as limitações temporais das RNNs. Uma breve discussão sobre a revolução proporcionada pelos Transformers na manipulação eficiente de sequências extensas de dados precede uma análise mais detalhada dos Vision Transformers, ressaltando como essas arquiteturas transformaram a visão computacional. O artigo também aborda a integração de diversas modalidades em LLMs multimodais, destacando a importância dos Vision Transformers nesse contexto. Um estudo de caso prático é apresentado, delineando um roadmap essencial para compreender a evolução temporal e a crescente relevância desses modelos na inteligência artificial.**

Palavras chaves — Modelos de Linguagem Multimodais, Recurrent Neural Networks (RNNs), Large Language Models (LLMs), Vision Transformers, Visão Computacional, Inteligência Artificial.

I. INTRODUÇÃO(*HEADING 1*)

No cenário dinâmico da inteligência artificial, a evolução dos modelos de linguagem desempenha um papel crucial na capacidade dos sistemas de compreender e gerar informações de maneira eficaz. Este artigo traça uma trilha pelos avanços notáveis em modelos de linguagem multimodais, com destaque para a transição desde as Recurrent Neural Networks (RNNs) até os atuais Large Language Models (LLMs) multimodais. Iniciaremos nossa jornada revisitando as limitações temporais das RNNs e, em seguida, exploraremos a revolução introduzida pelos Transformers, que permitiram o processamento eficiente de sequências extensas de dados.

À medida que avançamos, exploraremos a integração de modalidades diversas em Large Language Models(LLMs) multimodais. Um caso prático será apresentado para ilustrar a aplicação prática desses modelos, delineando um roadmap essencial para entender a evolução temporal e a crescente relevância dessas inovações na inteligência artificial. Este artigo visa fornecer uma visão abrangente e informada sobre o papel crucial dos Vision Transformers na convergência de linguagem e visão, impulsionando a próxima fase da revolução multimodal na inteligência artificial.

II. FUNDAMENTOS TEÓRICOS

A. Mecanismos e Técnicas (Heading 2)

Neste artigo, investigamos os mecanismos e técnicas que moldaram a evolução dos modelos de linguagem multimodais. Inicialmente, exploramos as limitações temporais inerentes às Recurrent Neural Networks (RNNs) e a revolução introduzida pelos Transformers, evidenciando sua capacidade de processar sequências extensas de dados de maneira eficiente. Uma ênfase especial é dada aos Vision Transformers, destacando como essas arquiteturas remodelaram a visão computacional, proporcionando avanços notáveis em desempenho e eficiência.

B. Possível solução para o Problema

Além disso, examinamos a integração de modalidades diversas em Large Language Models(LLMs) multimodais, identificando os Vision Transformers como peças-chave nesse processo. Este artigo apresenta um caso prático para ilustrar a aplicação desses modelos, delineando um roadmap essencial para compreender a evolução temporal e a crescente relevância dessas inovações na inteligência artificial. Assim, os Vision Transformers emergem como uma possível solução para superar as barreiras entre linguagem e visão, impulsionando uma fase significativa na revolução multimodal e proporcionando uma compreensão mais holística e contextual das informações.

III. REVISÃO BIBLIOGRÁFICA

A. O que é Vision Transformer (ViT)?

ViT é uma classe de modelos de aprendizado profundo que usa a arquitetura transformer, originalmente desenvolvida para dados sequenciais (como texto), e adapta para tarefas de reconhecimento de imagens. Diferente das redes neurais convolucionais (CNNs) convencionais que processam imagens em partes, ViTs consideram uma imagem inteira como uma sequência de partes e aplicam mecanismos de self-attention (permitem que o modelo avalie e destaque partes importantes de um input, como palavras em uma frase ou segmentos de uma imagem, com base na relevância dessas partes para o contexto ou tarefa específica) para entender as dependências globais entre essas partes.

B. Funcionalidade e Mecanismo

ViTs dividem uma imagem em partes de tamanho fixo, fazem uma incorporação linear de cada uma delas, adicionam embeddings de posição e alimentam a sequência resultante de vetores em um codificador transformer padrão. A autoatenção permite que o transformer pese a importância de diferentes partes na imagem, focando nas mais relevantes para a tarefa em questão. Em outras palavras, ViTs analisam imagens segmentando-as, transformando esses segmentos em vetores e usando um codificador transformer para focar nas partes mais relevantes para a tarefa específica.

A eficácia dos ViTs aumenta com o tamanho do modelo e o pré-treinamento extensivo em grandes conjuntos de dados.

C. Aplicações e Significado

Esses modelos têm mostrado grande promessa em várias tarefas de processamento de imagem, como classificação, detecção de objetos e segmentação, muitas vezes superando as CNNs tradicionais, especialmente em cenários com grande disponibilidade de dados e recursos computacionais, A capacidade de capturar dependências de longo alcance nos dados os torna particularmente eficazes em tarefas onde o entendimento contextual de toda a imagem é crucial.

D. Desenvolvimentos Recentes e Adaptações

Há um esforço para melhorar a eficiência dos ViTs, abordando seus altos custos computacionais e de memória. Técnicas como a introdução de módulos leves para adaptações específicas de tarefas têm sido desenvolvidas. Estas adaptações visam melhorar a transferibilidade de ViTs pré-treinados para vários domínios visuais com treinamento adicional mínimo.

Em resumo, os Vision Transformers representam uma mudança significativa na abordagem de tarefas relacionadas a imagens em aprendizado de máquina, oferecendo uma alternativa promissora às CNNs tradicionais. Sua capacidade de processar imagens como sequências e capturar dependências globais oferece uma ferramenta poderosa para tarefas complexas de reconhecimento visual. No entanto, sua eficiência e aplicabilidade em cenários com recursos limitados continuam sendo áreas de pesquisa ativa. O que buscamos com esse artigo é trazer mais clareza sobre esse assunto, uma vez que, seu potencial é tão alto quanto a evolução dos modelos de LLMs utilizados no mercado hoje.

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

Note that the equation is centered using a center tab stop. 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 . . ."

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].

IV. Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Authors and Affiliations

The template 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).

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) For papers with less than six authors: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

c) Deletion: Delete the author and affiliation lines for the extra authors.

B. 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. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.*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.

TABLE 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. (*Table footnote*)

Fig. 1. Example of a figure caption. (*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 *(Heading 5)*

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

The template will 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].

[1] HAMADI, Raby. Large Language Models Meet Computer Vision: A Brief Survey. arXiv preprint arXiv:2311.16673, 2023.

[2] YUE, Xiaoyu et al. Vision transformer with progressive sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 387-396.

[3] PARK, Namuk; KIM, Songkuk. How do vision transformers work?. arXiv preprint arXiv:2202.06709, 2022.

[4] THISANKE, Hans et al. Semantic segmentation using Vision Transformers: A survey. Engineering Applications of Artificial Intelligence, v. 126, p. 106669, 2023.

[5] CHEN, Shoufa et al. Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems, v. 35, p. 16664-16678, 2022.

[6] PARVAIZ, Arshi et al. Vision Transformers in medical computer vision—A contemplative retrospection. Engineering Applications of Artificial Intelligence, v. 122, p. 106126, 2023.

[7] WANG, Yikai et al. Multimodal token fusion for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 12186-12195.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.aaa

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.