A Study of Method for Metadata Extraction via LLM and Scene Graph Generation

Byunghyun Kim, Dayeong Kim, Hyewon Seok, *Dongwook Lee, Seolyoung Jung
Kyungpook National Univ. *DataStreams

wlfjddlgoqn@knu.ac.kr, rlaekdud@knu.ac.kr, hws2008@knu.ac.kr

*dwlee@datastreams.co.kr, sunflower@knu.ac.kr

Abstract

The significance of metadata in information retrieval and data management is rapidly increasing, particularly in the realms of academic research and digital content curation. To enhance usability of metadata extraction of research papers and images, we propose an advanced method that surpasses existing techniques 2x accuracy by leveraging Large Language Model (LLM) and deeper classifiable Scene Graph Generation (SGG).

I. Introduction

As the volume of digital content expands exponentially, traditional methods of metadata extraction [1] struggle to keep pace, necessitating more sophisticated and scalable approaches. The emergence of Large Language Models (LLMs) has revolutionized various domains of artificial intelligence, offering unprecedented capabilities in understanding and summarizing with text. Similarly, advancements in scene graph generation (SGG) [2] have enabled deeper comprehension of visual content, which is perfect for metadata of image. We propose LLM to get metadata of paper, also deeper scene graph generation that can classify nuanced details like, human emotions, races, ages, and even specific animal breeds [3].

II. Text based Method

Traditional methods of metadata extraction from academic papers [1] often face challenges in accurately and efficiently capturing key information due to the limitations inherent in simple text processing techniques. These challenges are particularly pronounced when converting PDF files to text as this process can introduce errors such as misread characters (e.g., ‘hello’ converted to ‘heo’), or formatting issues which conventional text processing tools struggle to handle. To overcome these difficulties, we employ enhanced inference capabilities of Llama2 [4] that can understand text even when characters are corrupted or converted incorrectly. Our approach begins with a preprocessing step where PDF files are converted into text. This text is then inputted into an Llama2 with a specifically crafted prompt designed to extract essential metadata components such as the title, authors, abstract, references, and research domain. Llama2’s effectiveness is significantly enhanced through prompt-engineering techniques, which refine the queries made to the model to better suit the nuances of academic metadata. Additionally, we optimize the extraction process by adjusting the temperature and batch size of the model.

As illustrated in Table 1, employing Llama2-7B significantly enhances metadata extraction accuracy across 246 research papers, achieving nearly a 2.5 times improvement over traditional technologies. This demonstrates the forte of Llama2 academic understanding which is suitable to extract paper metadata.

Result	DBLPCheck	Llama2-7B
Authors found	40.24% (99)	91.46% (225)
Authors found +other	33.33% (82)	86.99% (214)
Research Domain	-	85.77% (211)

Table 1. Compare extraction accuracy

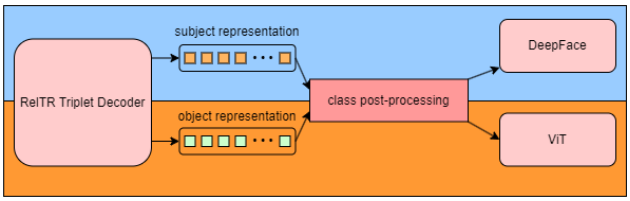

Figure 1. Deeper Scene Graph Architecture

III. Image based Method

For image metadata, we extend the utility of traditional tags with scene graph generation, a powerful tool that can classification and understanding of the relationships and elements within images. It is very useful for image searching or even in actively researched fields [5], [6]. We implement the RelTR [2] model, a state-of-the-art scene graph generator known for its efficiency in classifying objects within images. However, the standard implementation of RelTR includes limited classes related to human subjects (e.g., 'men', 'man') and animal breeds (e.g., ‘dog’) which can lead to lack of image explanation. To address this limitation, we integrate DeepFace [7] technology to classify nuanced human characteristics such as race, age, and emotions. For animal breeds, we utilize an ImageNet-based detection models [8], [9] that specializes in identifying specific breeds from images (e.g., ‘Dandie Dinmont’, ‘Tabby’). This approach marginally extends the run-time, but significantly enhances the scope of classifiable labels. Also we constructed a specialized small test set designed to evaluate our model capabilities, featuring more specific labels that refine general categories into distinct entities, such as categorizing 'dog' into 'golden retriever' in existing large-scale dataset like Open Images [10]. Model is released in https://github.com/KBH00/Deeper_RelTR.

Figure 2. RelTR and Ours (right)

label	RelTR	Ours
Human (Age×)	12	504
Animal (dog,cat,bird)	3	143
Animal (dog,cat,bird×)	7	204
Others	4	18

Table 2. Number of classifiable labels

Method	R@50 (V6)	R@50 (V6 +detailed)
RelTR	71.66	28.41
Ours	73.52	54.85

Table 3. Comparison on V6 and detailed test set

In figure 1, we illustrate our methodology which utilizes an end-to-end approach through the RelTR checkpoint (i.e., use existing model path). Here, triplets are detected using the Triplet Decoder. Subsequent post-processing enables the classification of detected classes. The post-processing process also, classifying which ImageNet classes [3] are included as subclasses among the existing RelTR classes, can allows the use of any ImageNet based models. During further branches, objects pertinent to human and animal classifications are selectively routed through the DeepFace and Vision Transformer (i.e., ViT), respectively. This two-stage process ensures that each element within the image is accurately tagged, enhancing the depth and utility of the generated metadata.

As in figure 2, that can classify man to 27 year old happy white man and dog to golden retriever. This outcome demonstrates that our more intricate scene graph acquires enhanced capability for elucidating images. Table 2 highlights the refined classification features of our methodology across three categories: Human (Age×), Animal Type, and Others (e.g., ‘vehicle’ to ‘bus’). For Person class, the model expands from 12 to 504 configurations by integrating with 6 races and 7 emotions, providing a detailed classification of human subjects even with estimated age.

In Table 3, our model initially performed similarly to RelTR on the original V6 [10] test set, achieving an R@50 of 73.52% compared to RelTR's 71.66%, with both models evaluated using partial matching to give score for partially detected labels. This higher score is because our system also possesses the capability to rectify misclassifications made by the RelTR model. Furthermore, on the V6 test set with more specific and detailed labels, our model's performance significantly improved to 54.85% R@50, substantially outperforming almost twice than RelTR’s 28.41%.

IV. Conclusion

Utilizing Llama2 for text has not only enhanced the accuracy of metadata extraction but also enriched the granularity of data obtained from academic papers. Follow-up research, such as fine-tuning the LLM, is possible to extract metadata not only from research papers but also from various texts. RelTR, supplemented by DeepFace and ImageNet-based models, can describe breeds and provide detailed human classifications. The class post-processing used here can significantly contribute to future work, especially in visual question answering and more detailed image generation studies. A multi-modal strategy that synergistically combines LLMs' language understanding capabilities with our scene graph model's visual insights could further leverage this holistic approach. Performance improvements can also be expected by leveraging Llama3 [11], panoptic scene graph generation [12], or additional research on related metrics and methodologies.

ACKNOWLEDGMENT

"This research was supported by the Korean MSIT (Ministry of Science and ICT), under the National Program for Excellence in SW)(2021-0-01082) supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation)"(2021-0-01082)

REFERENCES

[1] Marinai, S. "Metadata Extraction from PDF Papers," Proc. 10th Int. Conf. Doc. Anal. Recognit., Barcelona, Spain, July 2009, IEEE, doi:10.1109/ICDAR.2009.232.

[2] Cong, Y., et al. "RelTR: Relation Transformer for Scene Graph Generation," arXiv:2201.11460, Apr. 2023, doi:10.48550/arXiv.2201.11460.

[3] Deng, J., et al. "ImageNet: A Large-Scale Hierarchical Image Database," 2009 IEEE Conf. on Comp. Vision and Pattern Recognition, Miami, FL, June 2009, IEEE Xplore, doi:10.1109/CVPR.2009.5206848.

[4] Touvron, H., et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, Jul. 2023, doi:10.48550/arXiv.2307.09288.

[5] Hildebrandt, M., et al. "Scene Graph Reasoning for Visual Question Answering," arXiv:2007.01072, Jul. 2020, doi:10.48550/arXiv.2007.01072.

[6] Azade, F., et al. "SceneGenie: Scene Graph Guided Diffusion Models for Image Synthesis," arXiv:2304.14573, Apr. 2023, doi:10.48550/arXiv.2304.14573.

[7] Serengil, S.I., Ozpinar, A. "LightFace: A Hybrid Deep Face Recognition Framework," 2020 Innovations in Intelligent Systems and Applications Conf., Istanbul, Turkey, Oct. 2020, IEEE Xplore, doi:10.1109/ASYU50717.2020.9259802.

[8] Dosovitskiy, A., et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv:2010.11929 [cs.CV], 3 Jun. 2021, doi:10.48550/arXiv.2010.11929.

[9] He, K., et al. "Deep Residual Learning for Image Recognition." arXiv:1512.03385 [cs.CV], 10 Dec. 2015, doi:10.48550/arXiv.1512.03385.

[10] Kuznetsova, A., et al. 'The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale,' arXiv:1811.00982v2 [cs.CV], 21 Feb. 2020, doi:10.48550/arXiv.1811.00982.

[11] Llama3. (2024). Llama3 GitHub Repository. Available at: https://github.com/meta-llama/llama3a

[12] Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z. 'Panoptic Scene Graph Generation.' arXiv:2207.11247 [cs.CV], 22 Jul 2022, doi:10.48550/arXiv.2207.11247.