International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Figure Detection and Comparison

Ardhra Ann Denny[a], *Jeevika Sojan*[a], *Reethu George*[a], *Sandra Theresa Mathew*[a],*Sivadas T Nair*[b]

[a] UG Student, Artificial Intelligence and Data Science, Viswajyothi College of Engineering and Technology, Muvattupuzha, 686661
[b] Assistant Professor, Artificial Intelligence and Data Science, Viswajyothi College of Engineering and Technology, Muvattupuzha, 686661

ABSTRACT:

In this paper, we present an innovative approach that combines advanced neural networks and OpenCV to transform scanned academic materials into dynamic digital content. Our method seamlessly extracts images and labeled figures from various academic sources, including answer sheets and textbooks. By leveraging images from both textbooks and answer sheets, our system can generate detailed descriptive summaries for each image. These summaries act as helpful tools in understanding the content of the images and facilitating deeper comprehension of the underlying academic material. Then we use a comparison method to evaluate how well student answers align with the provided academic materials. By analyzing the cosine similarity between the two, we can get a sense of students' understanding levels and identify areas that need more attention. This combination of neural networks, computer vision tools, and cosine similarity evaluation marks a significant advancement in how we approach digital education and assessment.

Keywords: Gemini Vision API, YOLOv8, Neural Network, OpenCV, Context Attention Block (CAB), Spatial Attention (SA), Gated Recurrent Unit

1. Introduction

We have introduced an approach aimed at revolutionizing traditional educational materials into immersive digital resources. At the core of our methodology lies a model, intricately engineered to adeptly extract images from an extensive range of educational sources, spanning from conventional textbooks to intricate answer sheets. The process commences with the extraction of figures embedded within answer sheets, facilitated by a finely tuned image detection model. This model collaborates with a sophisticated question analysis component, which meticulously discerns whether a particular answer warrants the inclusion of a corresponding figure. Upon determination, the image detection model, powered by the advanced YOLOv8 framework, impeccably isolates, and extracts the relevant figure from the answer sheet. Facilitated by a meticulously curated dataset comprising 150-200 labelled images, each meticulously annotated with a single class label, the model undergoes rigorous training to ensure unparalleled performance and accuracy. Moreover, the versatility of this model extends to encompass the extraction of figures from answer keys presented in PDF format, further augmenting its applicability across diverse academic scenarios.

To make it easier to compare the extracted figures, we use an image captioning or descriptive model that taps into the capabilities of APIs like GPT-4 and Gemini Vision. These powerful APIs generate detailed, accurate descriptions of the extracted figures, allowing for a thorough assessment of content correlation. By using cosine similarities, we methodically compare the descriptive summaries to identify subtle nuances and variations, enabling a nuanced evaluation of how well the educational materials align with student responses. This meticulous process not only enhances the accessibility and engagement of educational content but also paves the way for a transformative shift in academic assessment and content evaluation.

2. Literature review

2.1. Real time object detection using YOLOv8-CAB

The study introduces a new way to improve the accuracy of the YOLOv8 model in detecting objects, especially small objects in different types of images. By adding the Context Attention Block (CAB) and changing the Coarse-to-Fine (C2F) block, the proposed method enhances feature extraction without making the model more complex. It also modifies the Spatial Attention (SA) module to speed up the detection process. The resulting YOLOv8-CAB model significantly improves the detection of smaller objects by using multi-scale feature maps and repetitive feedback from the CAB block.

Testing on the COCO dataset shows a remarkable improvement over standard YOLO models, achieving a mean average precision of 97% in detection rate, which is a 1% increase compared to traditional models. This advancement paves the way for further improvements in real-time object detection techniques, especially in scenarios where rapid and precise detection is required.

2.2. Image captioning model

Image captioning is a task that combines computer vision and natural language processing. It goes beyond just identifying objects and provides detailed descriptions of images. The paper presents an attention-based Encoder-Decoder architecture for image captioning. It combines features from Xception, a convolutional neural network (CNN), and object features from YOLOv4, an object detection model. The goal is to generate accurate and descriptive sentences for images by leveraging semantic understanding and precise description generation. The methodology involves pre-processing the data, extracting relevant features, and decoding captions using an attention mechanism and Gated Recurrent Unit (GRU). The model utilizes object features to improve the quality of the generated captions and introduces a novel "importance factor" for encoding the layout of objects in the images.

Experiments conducted on the MS COCO and Flickr30k datasets, which demonstrated a significant improvement in caption quality. The CIDEr score, a metric for evaluating caption quality, showed a 15.04% increase compared to previous approaches. Unlike previous approaches, this work integrates all object features directly, leveraging object layout information and yielding significant improvements.

3. Proposed work

We introduce a method for analysing figures in papers by utilizing cutting-edge techniques. Initially precise figure detection and extraction are accomplished using YOLOv8 ensuring identification of components in the document. Subsequent to this the process seamlessly incorporates figure with the Gemini Vision API allowing for the creation of descriptions for each figure extracted, improving document accessibility and comprehension. Additionally, a new approach involving cosine similarity is introduced for comparing figures. This proposed method not only automates figure analysis but also improves the clarity and usefulness of academic papers thereby contributing to advancements, in document processing and comprehension.

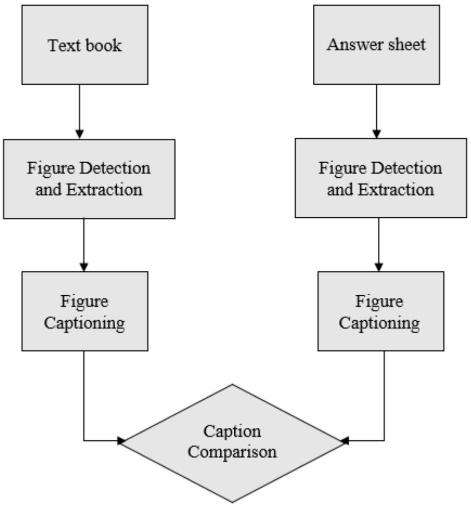

Fig. 1 – Overall process overview

3.1. Figure detection and extraction

Figure detection and extraction refer to the process of identifying and extracting graphical elements, such as images, diagrams, charts, and graphs, from digital documents. We introduce a new method for finding and extracting figures from digital documents. It uses the advanced capabilities of YOLOv8, a cutting-edge deep learning framework. This system can effectively identify figures in various document types, from scientific papers to technical reports. By leveraging YOLOv8, the proposed approach achieves exceptional accuracy in detecting figures, even in complex layouts with text. Additionally, the use of YOLOv8 streamlines the extraction process, ensuring that figures are precisely isolated and extracted for further analysis or manipulation. Extensive experimentation demonstrates the effectiveness of this method, showcasing its superiority over existing techniques and its potential to enhance applications like document indexing, content analysis, and information retrieval.

YOLOv8 is a cutting-edge approach to image detection tasks. Its key innovation is its ability to process entire images, rather than dividing them into grids or segments like traditional methods. By using a single neural network, YOLOv8 can simultaneously predict bounding boxes and class probabilities for objects within an image. This unified approach not only simplifies the detection process but also enables real-time detection with impressive speed and precision. YOLOv8 is particularly well-suited for applications that require rapid and accurate object detection, such as autonomous vehicles, surveillance systems, and medical imaging, thanks to its remarkable efficiency and accuracy.

YOLOv8 is an improved version of its previous models, with enhanced capabilities. It can extract features better, has a more efficient network architecture, and uses refined training strategies. These enhancements make YOLOv8 more versatile and able to handle a variety of object classes, sizes, and orientations. This makes it suitable for diverse detection scenarios. Additionally, YOLOv8 allows users to customize the model's configurations to suit their specific needs and optimize its performance for different domains and datasets. Overall, YOLOv8 represents a significant advancement in image detection technology, offering a powerful and efficient solution for a wide range of real-world applications.

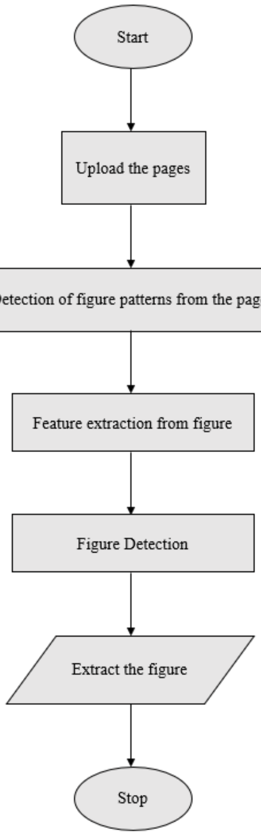
Start
Upload the pages
Detection of figure patterns from the pages
Feature extraction from figure
Figure Detection
Extract the figure
Stop

Fig. 2 – Figure detection and extraction

3.2. Figure captioning

Figure captioning is a task that involves generating textual descriptions or captions for images. In figure captioning, a model takes an input image and generates a descriptive sentence or phrase that accurately reflects the visual content depicted in the image. In this paper, we introduce an innovative approach for automatic figure captioning leveraging the Gemini Vision API. Figure captioning plays a crucial role in enhancing the interpretability and accessibility of visual content within scholarly documents. By integrating the Gemini Vision API, our system generates descriptive captions for figures extracted from digital documents, providing valuable context and insights to readers. The Gemini Vision API, renowned for its accuracy and versatility in image understanding tasks, enables our system to produce coherent and informative captions that accurately describe the content of each figure. Through extensive experimentation and evaluation, we demonstrate the effectiveness and reliability of our approach in generating high-quality captions for a diverse range of figures, thereby facilitating improved comprehension and usability of scholarly documents across various domains.

The Gemini Vision API provides a robust solution for automating figure captioning. It allows seamless integration of advanced image understanding capabilities into applications and systems. By leveraging cutting-edge deep learning techniques, the API analyses images and extracts meaningful visual features. This enables accurate recognition of objects, scenes, and context within the image. The rich understanding of image content forms the basis for generating descriptive captions that summarize the key elements and concepts depicted in the figure.

By using the Gemini Vision API, developers and researchers can make the process of creating informative captions for figures in academic papers and other visuals easier. The API's strong performance and flexibility allow applications to generate captions that are not just accurate, but also relevant to the context and flow naturally. The API's scalability and easy integration also make it possible to use figure captioning solutions across different platforms and areas, improving accessibility, understanding, and usefulness of visual content for various audiences.

3.3. Figure comparison

In this context, figure comparison specifically refers to comparing the descriptions or captions generated for different images. This comparison can involve assessing how well the captions match the meaning and relevance of their corresponding images. In this paper, we present an innovative approach for comparing figures using cosine similarity applied to their descriptions. Traditionally, figure comparison in scholarly documents has relied heavily on visual cues, which may overlook the nuanced semantic context of the content. By utilizing textual descriptions extracted from figures, our method enhances the comparison process by incorporating linguistic information, enabling a more comprehensive assessment of similarities and differences. We leverage advanced natural language processing techniques to extract informative captions or labels from figures. Then, we compute the cosine similarity between the descriptions, providing a quantitative measure of the semantic resemblance between the descriptions. This integration of textual information into the figure comparison framework enhances the clarity of the comparison results. It also helps in better understanding the data trends and relationships shown in documents.

In our approach it initializes a pre-trained sentence transformer model, which converts the textual descriptions of two figures into numerical representations called embeddings. These embeddings are then used to calculate the cosine similarity using the PyTorch implementation from the sentence-transformers library. The resulting cosine similarity score measures the semantic similarity between the descriptions, giving a reliable way to assess the resemblance between figures. By incorporating this approach into figure comparison processes, we can gain valuable insights into the content depicted in figures, advancing research in areas such as document analysis, content understanding, and knowledge discovery.

By incorporating textual descriptions, this approach enhances the interpretability and accuracy of comparisons, capturing nuanced meaning that visual analysis alone may miss. This allows for a more comprehensive understanding of the content depicted in figures, enabling researchers to discern subtle differences and similarities that contribute to deeper insights into the underlying data trends. Additionally, leveraging cosine similarity provides a robust and intuitive metric for quantifying the resemblance between figure descriptions. Unlike methods relying solely on visual features, cosine similarity offers a standardized measure that accounts for semantic similarities, facilitating more reliable and consistent comparisons across various figures and documents.

4. Experimental results

4.1. Dataset setup

We are preparing to train the YOLOv8 model for custom figure detection. To do this, we are creating a specialized dataset. This process involves compiling previous answer sheets, assignments, and relevant materials to form the training data. Next, we carefully annotate each image, labelling bounding boxes to highlight the areas of interest. By creating this dataset, we establish a comprehensive repository of annotated images, meticulously curated to enable the robust training of our figure detection model.

4.2. Result

Our figure detection model showed impressive performance, largely because we fine-tuned the pre-trained YOLOv8 model. This refinement boosted the model's accuracy in detecting figures within the scanned academic materials. As a result, our system successfully extracted figures from every page of the answer sheets, accurately identifying the class of each figure and providing precise bounding box coordinates for their localization and e-xtraction. This robust performance highlights the effectiveness of our approach in automating the figure extraction process, which improves document digitization workflow and enhances accessibility to visual content in answer sheets. Our study utilizes the Gemini Vision API to automatically generate detailed captions for figures in answer sheets. Once a figure is uploaded, the system promptly retrieves an informative one-paragraph description that captures the essence of the visual content, providing valuable insights into its content. This integration improves the accessibility and understanding of materials. When comparing different figures, the cosine similarity scores give a reliable way to measure how similar the meanings of the descriptions extracted from the images are. This score provides a comprehensive assessment of how closely the text generated from different images matches in meaning.

5. Conclusion

In this study, we have developed a comprehensive method to turn traditional academic materials into engaging digital resources. We used advanced neural networks and computer vision tools like OpenCV. Our approach can extract images and labeled figures from diverse academic sources, including answer sheets and textbooks. We then generate descriptive summaries for each extracted image, using the Gemini Vision API for automated figure captioning. Additionally, we employ cosine similarity to assess the correlation between the provided materials and student responses, uncovering subtle nuances and variations. This process enhances the accessibility and engagement of educational content, while also laying the foundation for transformative changes in academic assessment and content evaluation. This study paves the way for future research in refining and optimizing methodologies for figure extraction, captioning, and equation extraction. These techniques could be applied beyond just academic materials, like in e-learning platforms and digital libraries. Given the rapid progress in AI and computer vision, there are chances to integrate emerging technologies which could enable automated content creation and assessment in education.

Acknowledgements

We would like to express our sincere gratitude to Viswajyothi College of Engineering and Technology, Vazhakulam for their generous support. We would like to express our sincere gratitude and heartfelt thanks to Dr. Anita Brigit Mathew, Mrs. Mary Nirmala George and Mrs Anupriya Mohan of the Artificial Intelligence and Data Science Department for their motivation, assistance, and help. We also thank all the faculty members of the Artificial Intelligence and Data Science Department for their constant support and encouragement.

REFERENCES :

1. Talib, M., Al-Noori, A. H., & Suad, J. (2024). YOLOv8-CAB: Improved YOLOv8 for Real-time object detection. *Karbala International Journal of Modern Science*, *10*(1), 5.

2. Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, *9*(1), 20.

3. Fang, J., Feng, Z., & Cai, B. (2022). DrawnNet: offline hand-drawn diagram recognition based on keypoint prediction of aggregating geometric characteristics. *Entropy*, *24*(3), 425.

4. Gunawan, D., Sembiring, C. A., & Budiman, M. A. (2018, March). The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series* (Vol. 978, p. 012120). IOP Publishing.

5. Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016, April). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International conference on cyber and IT service management* (pp. 1-6). IEEE.

6. PL, C. (2019). A study on various image processing techniques. *International Journal of Emerging Technology and Innovative Engineering*, *5*(5).