Amrita School of Computing, Amritapuri

Department of Computer Science and Engineering (AI)

B. Tech 2021-2025 CSE(AI)

Project Phase 1

"Deep Vision: Automated Image Captioning and Multi-Modal Knowledge Graph Construction"

Group No: AI_AB2

Team Members

Sayanthika K R - AM.EN.U4AIE21058

Adithya R K Nambiar - AM.EN.U4AIE21105

Alap S Suresh - AM.EN.U4AIE21110

Gokul Krishna B R - AM.EN.U4AIE21131

August 2024

PROJEC T GUIDE	**COORDINATOR**	**COORDINATOR**
Dr. Lekshmi S Nair	Raji Ramachandran	Ms. Nisha K S

Domain

In the project "Deep Vision: Automated Image Captioning and Multi-Modal Knowledge Graph," we focus on essential fields such as computer vision, deep learning, natural language processing, multi-modal data integration, and knowledge graphs. Computer vision is what allows machines to see, through the interpretation and understanding of visual information; deep learning contributes complex layers for algorithms related to image processing and analysis. NLP is important to transform those visual interpretations into textual content. This will make the interaction between humans and machines more direct. Multi-modal data integration will allow more powerful analyses by bringing together different kinds of data like images and texts for meaningful representation. The construction of knowledge graphs goes a step further, providing structures to represent entities and relations explicitly, allowing to learn more about the data.

The significance of this finding lies in its potential to drive transformation across various domains. In fields ranging from content management, social media, this information has the potential to enrich a user's understanding of an image. Automated image captioning will make its presence felt where it is sorely needed. Leveraging a combination of visual and textual information over multimodal knowledge graphs offers even more refinement to systematic evaluation, improves sophisticated automated reporting generation, and enhances other decision-making capabilities. In addition to fixing current limitations in AI, this work crucially opens the door for future advancements by exploring many more possibilities with artificial intelligence, making it an exciting field for further research. The influence of these technologies as we build them will be massive and will prove to be an asset in many industries, such as e-commerce, healthcare, and more.

Problem

The main objective of this project is to translate visual information in images or scenes into human language descriptions by models generating natural descriptive captions, with appropriate understanding and expression. This project also has as an objective, the construction of knowledge graphs to render the context across related entities within images and thus greatly assist in capturing complex visual information in a structured format. It is such a two-pronged approach that will help in improving understanding and better accessibility of visual information, hence making it possible to strive for closer human-machine interactions. Integrating these capabilities shall help in the better explanation of scenes and improvement in the automated generation of content.

Dataset

The project will train on the COCO (Common Objects in Context) dataset, a largescale object detection, segmentation, and captioning dataset that contains over 330,000 images with bounding boxes around objects; detailed segmentations for precise localization, e.g., persons or cars marked out into their individual pixels mask-wise; key points as points of interest, usually major joints like eyes and nose positioned accurately to help set up skeleton rigs within the human body class (like other common categories person evaluations, bike parameters, etc.); captions multiple per each image describing different aspects of the same scene aiding internally afterward. The 80 object categories from COCO are a good set of diverse types of objects found in common environments, thus matching the objectives behind this project. It is rich, abundant, and wide in content types for its annotations; the semantics are described in different languages or other encodings with terms. We therefore argue for its use while developing models that integrate deep learningbased technologies for automatic image captioning and multi-modal knowledge graph construction. The dataset includes training, validation, and test splits to help improve the performance of models that are compatible with multiple tasks like object detection, segmentation, image captioning, or even key point detections. The COCO dataset is open source and an established standard in the computer vision/machine learning community, which will make it easy for researchers to compare their results with this work or see where they stand when collaborating on a similar problem statement. This means the project will be able to work from a base of knowledge and with existing techniques, opening possibilities for extended understanding and generation-related problems.

Hypothesis

The hypothesis of the project is to improve the relevance and generated description accuracy with the implementation of advanced deep learning models, especially transformer-based architectures, and compared with the traditional models in this project. More specifically, we will further expect these improvements in captioning to better facilitate more effective knowledge graph construction as richer and more precise textual descriptions enable better mappings of entities and their relations in images. This can subsequently be measured by comparing quality against existing benchmarks for captions generation and completeness of knowledge graphs. For example, these methods can be applied to develop more powerful image captioning algorithms based on advanced deep learning models. This will make the process of explaining contents within the images more accurate and relevant.

Multi-modal knowledge graphs can be created from these upgraded captions, which will provide better productivity and coverage of the point-to-point edge relationship built with a better understanding. Moreover, the pliability and worth of the captions created need to be considered by evaluating success, in part through comparison with results from ground truth data, and at the same time, considering practical utility with respect to how constructed knowledge graphs address issues of the real world and its applications.

Existing system/ Literature survey

Title of the work	Journal details/product details	Contributions of the paper/product	Limitations
Transform and Tell: EntityAware News Image Captioning	The paper was presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020.	- Introduces an endtoend model for generating captions for news images that incorporate realworld knowledge and named entities. The - model significantly outperforms previous methods on the GoodNews dataset, achieving a fourfold improvement in the CIDEr score.	- Primarily focuses on named entity recognition and may not generalize well to other types of images or contexts outside of news articles. - The system relies on pretrained models, which may limit its adaptability to novel domains.

Boosting Entity-Aware Image Captioning with Multi-Modal Knowledge Graph	The paper is published in the IEEE Transactions on Multimedia, Vol. 26, 2024.	- Effectively associates visual objects with named entities and captures relationships between them, leading to more informative and accurate image captions.	- The method relies heavily on external knowledge bases, which may introduce bias or inaccuracies depending on the data sources. - This approach may struggle with entities or relationships not well-represented in the external knowledge base.
Deep Learning Approaches on Image Captioning: A Review	Published in ACM Computing Surveys, Vol. 56, No. 3, Article 62, in October 2023.	- Highlights the advancements and categorizing various approaches. - Discusses challenges such as object hallucination and missing context, suggesting directions to address these issues.	- It is limited by the rapid pace of advancements in deep learning, meaning newer methods may not be covered.

A Large-Scale Multi-Modal Knowledge Graph with Triplet Fact Grounding	The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)	- Constructs a new large-scale multimodal knowledge graph called ImgFact - . Shows ImgFact improves performance on link prediction and relation classification tasks compared to existing approaches.	- Some triplets in ImgFact are outdated compared to current realworld facts due to knowledge graph information not being updated in real-time. Some images - lack the presence of either the head or tail entity, possibly due to pretrained CLIP model used.
Sequence to Sequence – Video to Text	It is research work authored by scholars from various institutions including the University of Texas at Austin, The University of Massachusetts, Lowell, and the International Computer Science Institute, Berkeley.	- The model learns the temporal structure of video frames and generates grammatical sentences that describe the events depicted in the video.	- Challenges such as handling diverse and complex video content and improving the accuracy of generated descriptions can be inferred as potential limitations.

Research/ Product

Aim of the project will require novel and detailed language description-generation strategies to be built for high-level image traits interpretation, as well as constructing graphical semantic networks that can improve computer vision and natural language processing. With the help of this excellent dataset, COCO will generate more accurate and relevant captions to encourage understanding around visual content and its context. Although the new paradigm also addresses limitations of current automated image analysis concepts, it supports research on state-of-the-art concurrency for a framework integrating visual and textual data that is expected to drive additional applications in areas such as reporting automation, social media content management, or accessibility solutions. In the end, results and methodologies from this project can better inform future generations of AI systems who understand and interact with their environment like a human, bolstering advances in artificial intelligence techniques.

Novelty

The novelty of this project lies in the new two-prong approach, enhancing our understanding of images. It puts together two omnipotent approaches: generation of positive image descriptions and knowledge models for aggregating known data to act as references. Unlike most existing solutions that focus on one of these aspects, this project does stand out as an amalgamation of both. Such integration not only improves artifact accuracy, but also makes it possible for the system to colour images that have never been seen before, showing flexibility.

This is added through knowledge models, which add depth and create beautiful visual descriptions. Thereby, such insights and connections can be opened through this program that no other technique would be able to produce. It is this blend of structured knowledge and advanced deep learning techniques which is making this automated visualization and knowledge representation project vastly influential and special in the domain.

References

- Transform and Tell: Entity-Aware News Image Captioning
- Boosting Entity-Aware Image Captioning with Multi-Modal Knowledge Graph
- Deep Learning Approaches on Image Captioning: A Review
- A Large-Scale Multi-Modal Knowledge Graph with Triplet Fact Grounding
- Sequence to Sequence – Video to Text

Students' Name and Signature

Students' Name	Signature
Sayanthika K R	Santhy
Adithya R K Nambiar	
Alap S Suresh	
Gokul Krishna B R	

Guide's Name and Signature

Guide's Name	Signature
Dr. Lekshmi S Nair	

Date: 13/08/2024