The Impact of Multimodal Large Language Models on Computer Vision

*Has GPT killed the “Vision Stars”?*

# Index

1. **Abstract**
2. **Introduction**
   1. Background of the Study
   2. Statement of the Problem
3. **Literature Review**
   1. Overview of Computer Vision
   2. Overview of Multimodal Large Language Models (LLMs)
   3. Previous Studies on the Impact of LLMs on Computer Vision
4. **Methodology**
   1. Description of the Open Source / Paid Models to be Used
   2. Description of the Application for Accessing (*Interfacing*) the Models
   3. Explanation of the Testing Procedures
   4. Evaluation Metrics and Performance Assessment
5. **Results**
   1. Presentation of the Results from the Tests
   2. Comparison of the performance of Computer Vision and LLMs
6. **Discussion**
   1. Discussion of the Results in the Context of the Objectives of the Study
7. **Conclusions**
   1. Summary of Findings
   2. Recommendations for Future Research

**Acknowledgements**

**References**

**Appendices**

# W - Index expansion

0 Abstract

* Start by stating your research question or objective.
* Briefly describe the methods you used to answer this question or achieve this objective.
* Summarize the main results of your research.
* Conclude with a statement about the implications of your findings.

1 Introduction

* 1.1 Background of the Study: Start with a broad overview of your field of study, then narrow down to the specific area your research focuses on. Include recent research or developments in the field.
* 1.2 Statement of the Problem: Clearly articulate the problem your research addresses. Explain why this problem is significant and why it is important that this problem is solved.
* 1.3 Objectiveness of the Study: List the specific objectives of your research. These should be clear, concise, and directly related to the problem statement.
* 1.4 Significance of the Study: Discuss the potential impact of your research on the field. This could include theoretical implications, practical applications, or policy implications.

2 Literature Review

* 2.1 Overview of Computer Vision: Provide a comprehensive review of the literature on Computer Vision. This should include a discussion of the following topics:
  + Definition and scope of Computer Vision: Explain what Computer Vision is, what its main goals are, and what types of problems it can solve.
  + History and development of Computer Vision: Trace the evolution of Computer Vision from its origins to the present day. Highlight the major milestones, breakthroughs, and challenges in the field.
  + Main subfields and applications of Computer Vision: Describe the main subfields of Computer Vision, such as image processing, object detection, face recognition, scene understanding, etc. Discuss how these subfields are related and how they are applied in various domains, such as medicine, security, entertainment, etc.
  + Current trends and challenges in Computer Vision: Identify the current trends and challenges in Computer Vision, such as deep learning, big data, multimodal data, adversarial attacks, etc. Discuss how these trends and challenges affect the field and what are the possible solutions or directions for future research.
* 2.2 Overview of Multimodal LLMs: Similarly, provide a comprehensive review of the literature on Multimodal LLMs. Discuss the following topics:
  + Definition and scope of Multimodal LLMs: Explain what Multimodal LLMs are, what their main goals are, and what types of problems they can solve.
  + History and development of Multimodal LLMs: Trace the evolution of Multimodal LLMs from their origins to the present day. Highlight the major milestones, breakthroughs, and challenges in the field.
  + Main types and architectures of Multimodal LLMs: Describe the main types and architectures of Multimodal LLMs, such as vision-and-language models, cross-modal models, multimodal fusion models, etc. Discuss how these types and architectures differ and how they are designed to handle multimodal data.
  + Main tasks and applications of Multimodal LLMs: Describe the main tasks and applications of Multimodal LLMs, such as image captioning, visual question answering, visual dialog, image-text retrieval, etc. Discuss how these tasks and applications are related and how they are applied in various domains, such as education, health, social media, etc.
  + Current trends and challenges in Multimodal LLMs: Identify the current trends and challenges in Multimodal LLMs, such as pre-training, fine-tuning, evaluation, generalization, explainability, etc. Discuss how these trends and challenges affect the field and what are the possible solutions or directions for future research.
* 2.3 Previous Studies on the Impact of LLMs on Computer Vision: Discuss previous research that has been conducted on the impact of LLMs on Computer Vision. This should include a summary of their findings and a discussion of how your research builds on these studies. You could organize this section by the following topics:
  + How LLMs can improve Computer Vision models: Discuss how LLMs can enhance the performance, robustness, and efficiency of Computer Vision models. Provide examples of studies that have shown the benefits of LLMs for Computer Vision tasks, such as object detection, face recognition, scene understanding, etc.
  + How LLMs can enable new Computer Vision tasks: Discuss how LLMs can enable new or novel Computer Vision tasks that were not possible or feasible before. Provide examples of studies that have demonstrated the potential of LLMs for Computer Vision tasks, such as image generation, image editing, image synthesis, etc.
  + How LLMs can pose new challenges or risks for Computer Vision: Discuss how LLMs can also pose new challenges or risks for Computer Vision, such as ethical, social, or security issues. Provide examples of studies that have raised the concerns or limitations of LLMs for Computer Vision, such as bias, fairness, privacy, safety, etc.

3 Methodology

* 3.1 Description of the Open Source / Paid Models to be Used: Describe the models you will be using in your research. Discuss why these models were chosen and their strengths and weaknesses.
* 3.2 Description of the Standardized Application for Accessing the Models: Explain how you will access and use these models. This should include a description of the software or applications you will use and any necessary training or preparation.
* 3.3 Explanation of the Testing Procedures: Detail the procedures you will follow to test the models. This should include a step-by-step description of the testing process and any measures you will take to ensure the validity and reliability of your results.
* 3.4 Evaluation Metrics and Performance Assessment: Define the metrics you will use to evaluate the performance of the models. Discuss how these metrics were chosen and how you will determine if a model is successful or not.

4 Results

* 4.1 Presentation of the Results from the Tests: Present the results of your tests. This should include a clear and concise summary of your data, as well as tables, graphs, or other visual aids to help illustrate your findings.
* 4.2 Comparison of the performance of Computer Vision and LLMs: Compare the performance of the Computer Vision models and the LLMs. Discuss which models performed better and why.

5 Discussion

* 5.1 Discussion of the Results in the Context of the Objectives of the Study: Discuss your results in relation to your research objectives. Did you achieve your objectives? If not, why not? What factors may have influenced your results?

6 Conclusion

* 6.1 Summary of the Findings: Summarize your main findings. This should be a clear and concise summary of what your research revealed about the impact of LLMs on Computer Vision.
* 6.2 Implications of the Study for the Field of AI and Computer Vision: Discuss the implications of your findings for the field. This could include theoretical implications, practical applications, or policy implications.
* 6.3 Recommendations for Future Research: Suggest areas for future research. This could include unanswered questions from your research or new questions that your research raised.

Acknowledgements

* Acknowledge the people who helped you with your research. This could include your advisor, other faculty members, and any organizations that provided funding or resources for your study.

References

* List all the sources you cited in your thesis. Be sure to format your references according to the citation style you are using.

Appendices

* Include any additional information that is relevant to your study but did not fit into the main body of your thesis. This could include raw data, additional graphs or tables, or interview transcripts.

# 0 Abstract

---TO BE DEFINED---

# Introduction

## 1.1 Background of the Study

The field of artificial intelligence (AI) has seen remarkable advances in recent years, particularly in the areas of large language models (LLMs) and computer vision. These two areas, while distinct, have begun to intersect with the advent of multimodal LLMs (MM-LLMs), models that can understand and generate information across different modes such as text and images. Recent research has shown the potential of these models in various applications, from image captioning to visual question answering, marking a significant shift in the landscape of AI.

## 1.2 Statement of the Problem

Despite these advancements, the full potential of multimodal LLMs in the field of computer vision remains largely unexplored. There is a need to understand how these models can be effectively utilized, the challenges that might arise, and the solutions to address these challenges. This research aims to delve into these issues, exploring the disruptive potential of multimodal LLMs in computer vision and their practical application in image classification.

# Literature Review

## 2.1 Overview of Computer Vision

Computer Vision is a field of Artificial Intelligence that aims to derive meaningful information from visual data from the real world, such as digital images or videos, by replicating the capabilities of human vision [1].

The advent of deep learning techniques has revolutionized this field, leading to significant advancements in tasks such as *image classification, object detection, and semantic segmentation*. Deep learning, a subset of machine learning has been particularly effective in computer vision due to its ability to learn from large amounts of data.

The development of Convolutional Neural Networks (CNNs), a class of deep learning models, has been instrumental in pushing the boundaries of what was previously possible in computer vision. Despite these advancements, traditional computer vision techniques continue to hold relevance, particularly in scenarios where deep learning models may not be fully optimized [[2](#ref_2)].

## 2.2 Overview of Multimodal Large Language Models (LLMs)

Multimodal Large Language Models (MM-LLMs) represent a significant advancement in the field of artificial intelligence, augmenting traditional LLMs to support multimodal inputs or outputs. These models leverage the reasoning and decision-maing capabilities inherent in LLMs, enabling to perform a diverse range of multimodal tasks *(the object of this paper/thesis is to analyse the performance of MM-LLMs in the image classification task: image to text)*. The development of MM-LLMs has been characterized by their applications, strengths, and weaknesses. For instance, they have been used in various domains, including autonomous driving. However, despite their success, there are still challenges such as the need for large amounts od data and computational resources. Recent research has focused on addressing these challenges, with efforts being made to optimize the training strategies of MM-LLMs. [3]

## 2.3 Previous Studies on the Impact of LLMs on Computer Vision

There have been several studies on the impact of LLMs on Computer Vision. For example, the VisionLLM framework provides a unified prespecrive for vision and language tasks by treating images as a foreign language and aligning vision-gentric tasks with language takss that can be flexibly defined and managed using language instructions. [<https://arxiv.org/abs/2305.11175>]

Another study demonstrated that LLMs can enhance the performance of computer vision tasks, providing new insights an directions for future research. [<https://arxiv.org/abs/2303.18223>]

However, it’s important to note that while LLMs have shown promise in enhancing computer vision takss, there are sill areas where traditional computer vision techniques may be more effective [<https://arxiv.org/abs/2401.06209>]

# Methodology

## 3.1 Description of the Open Source / Paid Models to be Used

The first problem to be solved was a structural one. Given a funding limitation the obvious structural choices were made in the open-source domain with the constraint of the hardware available.

The first objective to be accomplished was a basic interface to simplify interaction with the model in the pipeline hence I made a choice familiar to me: a web application to interact with various models.  
  
This allowed me to better grasp the problem of my personal hardware limitations that driven me to the second most important objectibe: after a finetuned prompt for the image classification task it was

Idea: best resource usage. Where possible extrapolate the image-encoder used in the MM-LLMs to perform zero-shot classification without relying on the whole structure. Ex: CLIP / ALIGN. This provides a more accurate analysis as it is not biased by potential errors in other parts of the structure.

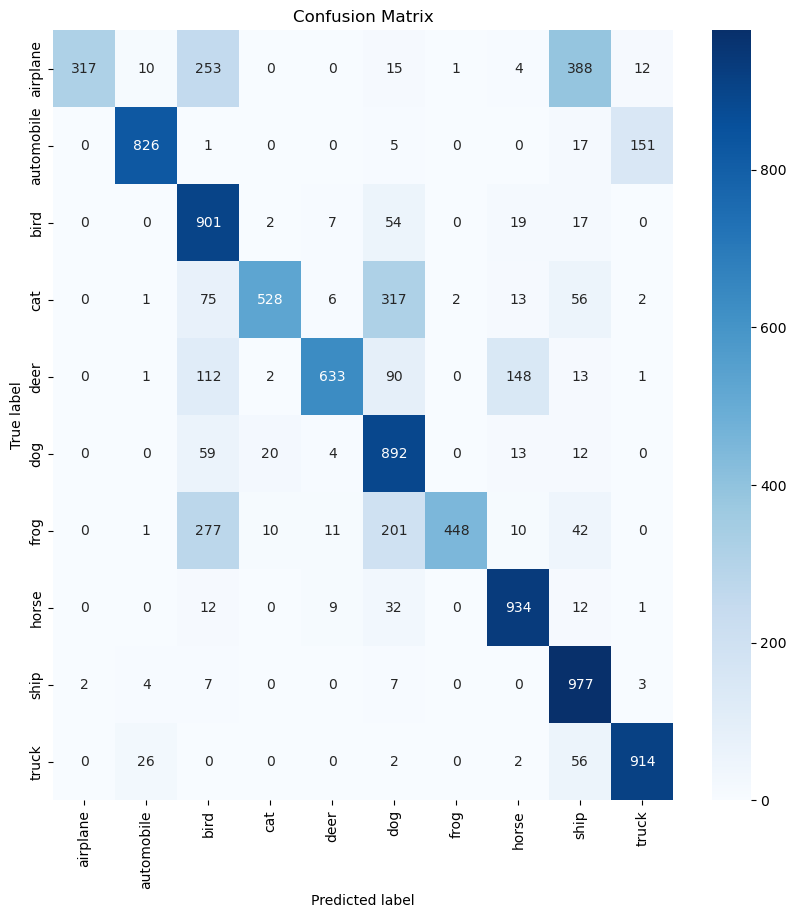
* 1. Description of the Standardized Application for Accessing the Models
  2. Explanation of the Testing Procedures

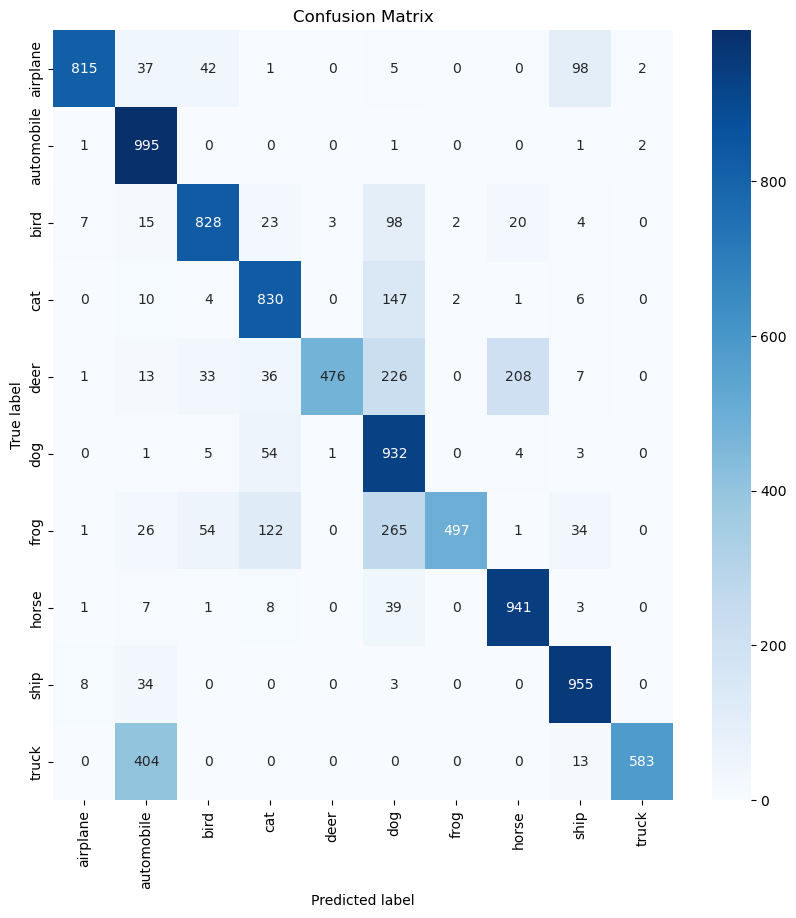
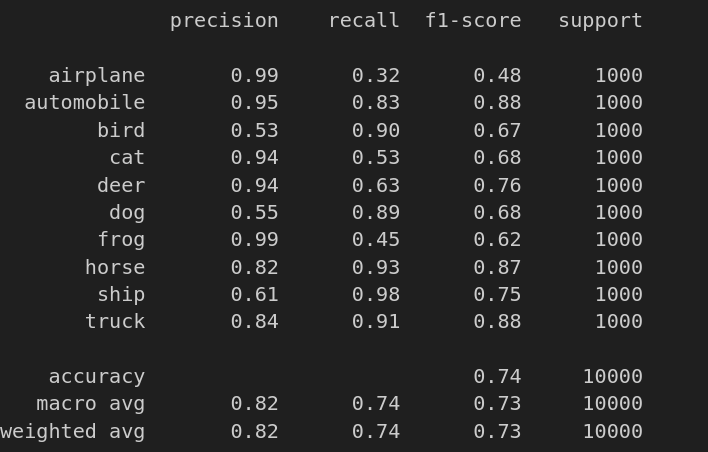
CLIP model:

- CIFAR 10

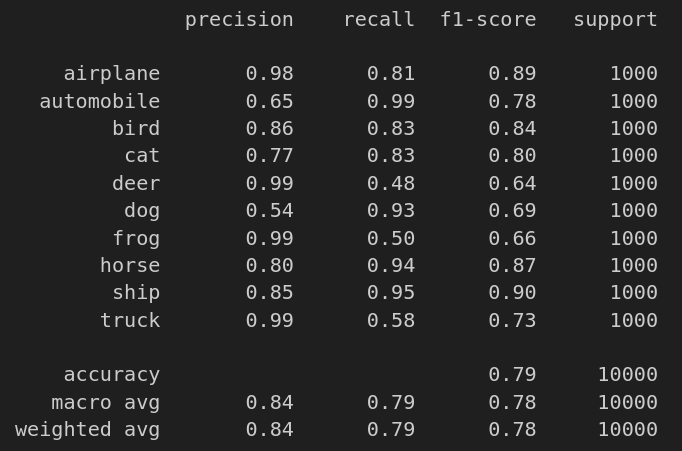
10k images (test set) → 10 minuti per evaluation.

**PROMPT: just the label.**





PROMPT: “the image contains a {label}”.



* 1. Evaluation Metrics and Performance Assessment

Dataset choice:

- Caltech 101

- Mini Image Net:

<https://paperswithcode.com/dataset/mini-imagenet>

https://www.kaggle.com/datasets/arjunashok33/miniimagenet

Model choice:

- KOSMOS-2

- CLIP

# References

[1] [x] What is Computer Vision? | IBM - <https://www.ibm.com/topics/computer-vision>

[2] [x] Deep Learning vs. Traditional Computer Vision | IMaR Technology Gateway <https://arxiv.org/ftp/arxiv/papers/1910/1910.13796.pdf>

[3] ---- <https://arxiv.org/abs/2401.13601>