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](#_1.1_Background_of)
   2. [Statement of the Problem](#_1.2_Statement_of)
3. **Literature Review**
   1. [Overview of Computer Vision](#_2.1_Overview_of)
   2. [Overview of Multimodal Large Language Models (LLMs)](#_2.2_Overview_of)
   3. [Previous Studies on the Impact of LLMs on Computer Vision](#_2.3_Previous_Studies)

­­

1. **Methodology**
   1. [Description of the Open Source / Paid Models to be Used](#_3.2_Description_of)
   2. [Description of the Standardized Application to Interface the Models](#_3.1_Description_of)
   3. [Explanation of the Testing Procedures](#_3.3_Explanation_of)
   4. [Evaluation Metrics and Performance Assessment](#_3.3_Evaluation_Metrics)
2. **Results**
   1. Presentation of the Results from the Tests
   2. Comparison of the performance of Computer Vision and LLMs
3. **Discussion**
   1. Discussion of the Results in the Context of the Objectives of the Study
4. **Conclusions**
   1. Summary of Findings
   2. Recommendations for Future Research

**Acknowledgements**

**References**

**Appendices**

# 0 Abstract

---TO BE DEFINED---

# Introduction

## 1.1 Background of the Study

The field of artificial intelligence has seen remarkable advances in the last decades, particularly in the areas of large language models (LLMs) and computer vision (CV). While distinct, these two areas have begun to intersect during the last years thanks to the introduction of Multimodal Large Language Models, MLLMs. MLLM is a type of deep learning model that expands the capabilities of Large Language Models (LLMs) with the ability to understand and process not just text but multimodal information such as text, images, audio and video.

Immagine che contiene testo, diagramma, schermata, linea

Descrizione generata automaticamente

*A timeline of representative MLLMs - source [4]*

Recent research has shown the potential of MLLMs in various applications, from image captioning to visual question answering, marking a significant shift in the landscape of AI*.* Nevertheless, the field is still under researched because of its novelty: 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.

## 1.2 Statement of the Problem

Consequently, the objective of this thesis is to delve into the intersection of Computer Vision and Multimodal Large Language Models by applying the latter in a classical computer vision problem: “zero-shot classification”. Therefore this study is looking forward to answering:

*“How well a MLLM, thought to be a <<chatbot>>, can perform in zero-shot image classification?”*

While zero-shot image classification is the computer vision task of classifying images into one of several classes, these types of models are demonstrated to have a wide range of applications. We will not get into their general-purpose capabilities, but rather evaluate their performance on the specific task stated. To use an analogy, we are trying to shoot down a bird using a cannon. By doing so we can measure how much our cannon is capable of precision. Finally, we will compare the performance with other AI methods to contextualize the obtained results.

\n

# 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].

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamenteThe inception of the field is often attributed to Lawrence Robert, who established the first common methodologies during the 1960s. Since then, the field has seen tremendous growth and evolution. Up until 2012 methods were primarily rule-based, requiring explicit programming to extract features and make decision rules. The evolution of the deep learning field concretized in a “no-way-back” paper “*ImageNet Classification with Deep Convolutional Neural Networks*” where CNNs were used to disrupt the state of art of image classification in the ImageNet challenge. There was a paradigm shift in the methodology employed by the computer algorithms. Instead of explicitly specifying how to link the data output deep learning is aimed to specify an architecture able to extrapolate from the data who to do so.

There has been an evolution in the methods

**2010s**

* In 2012, the first deep learning-based image recognition system, AlexNet, was introduced by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. AlexNet significantly outperformed its predecessors in the ImageNet Challenge, making a turning point in the field of computer vision.
* In 2018, the first Generative Adversarial Network (GAN) based image generation system, StyleGAN, was introduced by NVIDIA. StyleGAN could create high quality synthetic images [[[1812.04948] A Style-Based Generator Architecture for Generative Adversarial Networks (arxiv.org)](https://arxiv.org/abs/1812.04948)].
* In 2020, the first large-scale image recognition system based on transformers, known as Vision Transformer (ViT), was introduced by Alexey Dosovitskiy and Thomas Kipf. ViT applies a Transformer Encoder directly to sequences of image patches for classification. [[[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arxiv.org)](https://arxiv.org/abs/2010.11929)]

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)].

\n

## 2.2 Overview of Multimodal Large Language Models - MLLMs

From Convolutional Neural Networks to Large Language Models

Immagine che contiene testo, diagramma, schermata, Piano

Descrizione generata automaticamenteThe advent of deep learning was marked by the success of “AlexNet” in 2012 what these showcased the capabilities of deep learning. That was just the beginning of a new era for computer algorithms enabled by to the advent of big data (larger datasets and easier collection and storage), hardware advances (Graphic Processing Units and parallelization) and improved software techniques (new models and tooboxes).

A key milestone for the advent of Large Language Models was the advent of a new deep learning architecture known as the “Transformer”. Formalized in the paper “*Attention Is All You Need*” (2017) by a research team from Google, the architecture exploited a series of attention mechanism to compute weight between words. This enabled a series of advances in natural language processing that synthetized the Large Language Models (LLMs), architectures that are demonstrating unprecedent ability in general-purpose language generation. [https://arxiv.org/abs/1706.03762]

In 2020 the Transformer architecture was applied to image-recognition in the ViT model (Vision Transformer) with excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train. [https://arxiv.org/pdf/2010.11929.pdf].

One year later the link between vision and language was to be closed with the paper “*Learning Transferable Visual Models From Natural Language Supervision*” by OpenAI. The research team proposed CLIP, a neural network with the aim of learning visual concepts from natural language supervision. [https://arxiv.org/abs/2103.00020]

Immagine che contiene testo, diagramma, schermata, Piano

Descrizione generata automaticamente

It wasn’t long needed to expand models capabilities and connect image encoder to LLMs giving the birth of Multimodal Large Language Models.

Sources to look inside:

* <https://arxiv.org/pdf/2306.13549v2.pdf>
* <https://arxiv.org/pdf/2401.13601.pdf>

Multimodal Large Language Models represent a significant advancement in the field of artificial intelligence, augmenting traditional LLMs to support multimodal inputs or outputs. These models leverage the capabilities of LLMs such as reasoning and decision-making, enabling to perform a diverse range of multimodal tasks.

Immagine che contiene testo, schermata, diagramma, Piano

Descrizione generata automaticamenteThe development of ­­­MLLMs 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 MLLMs. [3]

**Architecture**

A typical MLLM can be abstracted into three modules, i.e. a pre-trained modality encoder, a pre-trained LLM and a modality interface to connect them. Drawing an analogy to humans, modality encoders such as image/audio encoders are human eyes/ears that receive and pre-process optical/acoustic signals, while LLMs are like human brains that understand and reason with the processed signals. In between, the modality interface serves to align different modalities[4].

++

/ esempi

* LLaVA

## 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*](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*](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*](https://arxiv.org/abs/2401.06209)*]*

\n

# Methodology

For a zero-shot classification task there are two main choices to be done are: which MLLM to use and which dataset to perform the classification on. Clearly, to benchmark the results obtained by MLLMs, the study will rely on previously existing research of zero-shot capabilities by computer vision (CV) , as a replica won’t be valuable. We will just have the foresight of choosing a dataset with extensive CV literature performed on it.

There are various trade offs to be made on the choice of the dataset. As our objective is to understand how well MLLMs are able to perform classification it is suitable for our dataset to embrace generality and not to be specific. On the other hand, we need a dataset to be among the classical datasets of computer vision to exploit a wide literature of computer vision. Hence the choice had to be “*ImageNet*”.

The last trade-off required for this study was the precision of the results of the study against the price (time and money). As unfunded research it would have been unsustainable to use the whole dataset for evaluation, so the study relies on a subset of the dataset called “mini-ImageNet”. Mini-ImageNet was proposed in NeurIPS (2016) and it consists of 50k training images and 10k testing images, evenly distributed across 100 classes []. It has to be noticed that the selected images were not resized to any particular size and are left to be the same size as they are in the ImageNet dataset [].

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

In the last year there has been a proliferation of Multimodal LLMs therefore there has been an analysis to select the most appropriate ones. First, we should distinguish between open and closed source [].

Various companies provide the service of inference against a black-box ml-algorithm through APIs and we’ll refer to those models as “closed-source”. Even though the nature of this paper is academical, hence inheritedly open-source the innovation in the field of LLMs is driven by private companies and the most advanced findings can be easily related to private companies. It’s not a mystery that architectures, such as GPT-4 Vision by OpenAI (Microsoft-funded), Gemini by Google and Claude-3 by Anthropic were the catalyst that made possible the current research fervour. Although the black-box nature of these models limits our results understanding and the research value, black-box Multimodal Large Language Models are best performing on many benchmarks, therefore this paper is forced to take them into account. It will be done by using as Claude-3 and, in a limited dataset due to financing frictions, GPT4-V.

On the other hand, the most valuable discussions from an academic perspective are achieved by the employment of the open-source options. And the study uses many different options:

* LLaVA, the most used and *de-facto* standard for Image-Text-to-Text
* Blip2, Visual Question Answering
* Kosmos2

[what are some reasons to open source AI? To prevent concentration of power in the hands of whose have developed it. We might say it’s undesirable to maintain the control of AI to a couple of companies. On the other hand there are near-term commercial incentives on the closed source side. Against open source there’s another long term argument which is: if eventually AI becohugely powerful ]

The selection of the large language model for this study was influenced by a resource constraint, leading to a series of trade-offs explained behind. The models under consideration fall into two categories: Open Source and Closed Source.

The study was conducted on various levels of profoundness, based on resources counted as: time, money, and hardware.

**Open-Source Models**:

1. LLaVA: This model was not selected due to its large size, which exceeded the capabilities of both my personal computer and the freemium tier of Hugging Face’s services. *The financial cost associated with upgrading to a higher tier was deemed prohibitive.*
2. CLIP: it is not a Language Learning Model (LLM) so it won’t be part of our active research: it is noteworthy because many LLMs utilize it in the background hence it will be used as one of the performance metrics to understand MLLMs performance.
3. KOSMOS2: This model was executed using the freemium environment provided by Hugging Face through their “Spaces” service.

**Closed Source Models**:

1. GPT4 by OpenAI: This model was not selected due to the financial cost associated with its use.
2. Gemini by Google: This model was not available for use in Europe, and therefore could not be selected for this study.
3. Claude3 by Anthropic: This model was ultimately chosen for use in this study.

Each of these models was evaluated based on a variety of factors, including cost, availability, and compatibility with existing systems. The final selection, Claude3 by Anthropic, best met the needs and constraints of this study.

* Open Source (hf)
  + LLaVA
    - Big for freemium Hugging Face => $$, work in progress, limited to personal expenses.
  + CLIP
    - Not a LLM => relevant as most of LLMs use it in background.
  + KOSMOS2
    - Executed relying on freemium Hugging Face environment through their “Spaces” service.
* Closed Source
  + GPT4 by OpenAI
    - $$ 🡪 1.6k images
  + Gemini by Google
    - Not available in Europe
  + Claude3 by Anthropic
    - Picked 🡪 10k images as less pricy

Trade-offs made in picking the models:

* model:
  + Open-source vs Paid Models
  + Size (performance on the task) / Cost-of-computation
  + API vs a controlled machine

## 3.2 Description of the Standardized Application to Interface the Models

To fulfil the classification tasks to the best of the model’s capabilities an important phase is the prompt engineering. Before proceeding with the classification on a dataset I delved into various ways of phrasing the prompt for the zero-shot classification.

To fulfil this objective, I developed an application to interface various models. The application stack is made by a JavaScript frontend and a Flask backend to accesses the models. The models are made available through either Hugging Face “tensors” for open-source ones or by API calls.

By following this path, I came to two conclusions that shaped the testing procedure:

* I didn’t have enough local memory to run the lightest LLaVA model (either on GPU or CPU) on my personal device. As LLaVA is the most used open-source model for multimodality I tried to find different freemium computing platforms without positive results. Nevertheless, this work wasn’t useless as I discovered Hugging Face’s product named Spaces that allows free access to a 16GB RAM on their CPUs: product that I used with the model KOSMOS-2, the only Large Language Model with multimodality small enough to fit in this computation.
* The prompt to be followed had to contain two parts or phrases: one to specify the task to be executed and one to ask the model for the response given the image.

**Prompt engineering**

Open Source

Different models needed different formats for the specification of the task. The general approach followed was the merging of on one hand strongly highlight the task, object of the inference, while on the other hand let the model be itself a chatbot: hence not trying to force a single word output.

* For GPT4-V instead, after some tests through my interface, it was clear that it got also the capabilities to understand the task of “zero-shot classification”. Therefore the prompt for GPT4-V also had the requirement to be constrained to be one of the classes provided.

\n

## 3.3 Explanation of the Testing Procedures

The computation was executed outside of the “Standardized Application” but rather with ad-hoc scripts.

**Open-Source**

For open-source models the challenge was on the hardware to employ for executing the models. The best solution as a balance between quality and price was the service “*Spaces*” offered by the company Hugging Face. The service grants hardware on their machines to execute arbitrary code, hence I developed an ad hoc script in python to:

* retrieve image by image from the miniImageNet database
* feed the Multimodal Large Language Models with the image and the ad-hoc prompt
* store the computation result inside the online database “firestore” by Google.

The data was stored to a third-party company to overcome the limitation of non-persistency due to the use of a freemium version of the “*Space*” service.

**Closed-Source**

For closed-source models the computation process must, by default, outsource the inference to the company’s API. This means on one end relieving the problem of owning the deployment of the model and on the other end the process has higher costs of inference. I was able to run the API calls locally, by respecting the hourly and daily limitations, and store the results inside the “firestore” database as for the Open-Source model to have things as standardized as possible.

**BERT**

Because of the closed source nature of the classification task and the generative nature of LLMs even employing prompt-engineering was not enough to directly obtain closed domain responses, with various levels of adherence to the domain. To account for that, if after a first “strip” of punctuation and upper case / lower case indifference the model’s response was not part of the possible classes of the database I employed BERT model [ref] to classify the text provided.

## 3.3 Evaluation Metrics and Performance Assessment

The task object of the study is “*zero-shot classification*” and as it will be performed on a dataset containing 100 classes, it falls under the denomination of “*multi-class classification*”.

For each model used to classify the database we will produce a confusion matrix, since it encloses all the relevant information about performance, and a report containing various indicators. [5]

Immagine che contiene testo, schermata, numero

Descrizione generata automaticamente**Confusion Matrix**

The confusion matrix is a cross table that records the number of occurrences between two raters, the true/actual classification, and the predicted classification. The classes are listed in the same order in the rows as in the columns, therefore the correctly classified elements are located on the main diagonal from top left to bottom right and they correspond to the number of times the two raters agree. [5]

To maintain uniformity across the paper, the columns represent the predictions made by the model, while the rows indicate the actual classifications.

**Report**

\n

# Results

## 4.1 Presentation of the Results from the Tests

**Multimodal Large Language Models**

ImageNet – subset of the training split:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Ways** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **Support** |
| *GPT 4* | *100* | *0.97* | *0.93* | *0.95* | *0.93* | *1.6k* |
| Claude 3 (haiku) | 100 | 0.79 | 0.63 | 0.65 | 0.63 | 10k |
| Kosmos 2 | 100 | 0.69 | 0.30 | 0.34 | 0.30 | 10k |
| Kosmos 2 | 5 | 0.20 | 0.16 | 0.17 | 0.16 | 10k |
| Blip2 | 5 | 0.50 | 0.39 | 0.41 | 0.39 | 10k |
| llava v1.5 7b vicuna | 5 | 0.60 | 0.50 | 0.40 | 0.50 | 10k |
| llava v1.6 7b mistral | 5 | 0.78 | 0.56 | 0.61 | 0.56 | 10k |
| llava v1.6 13b vicuna | 5 | 0.89 | 0.79 | 0.81 | 0.79 | 1.5k |

ImageNet – subset of the validation split:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Ways** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **Support** |
|  |  |  |  |  |  |  |

**Classical Computer Vision**

Executing the same task with artificial intelligence for image classification.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Ways** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **Support** |
| ResNet50 | 100 | 0.96 | 0.94 | 0.95 | 0.94 | 10k |

## 4.2 Comparison of the performance of Computer Vision and LLMs

The pure numbers will have to be contextualised by comparing them with the performance of all the other models, the “*classical computer vision*”. In particular:

**ResNet**

Paper: <https://arxiv.org/abs/1512.03385>

|  |  |  |
| --- | --- | --- |
|  | Top-1 error | *Derived Accuracy* |
| ResNet-34 B | 21.84% | 78.16% |
| ResNet-50 | 20.74% | 79.26% |
| ResNet-101 | 19.87% | 80.17% |
| ResNet-152 | 19.38% | **80.62%** |

*Error rates % of single-model results on the ImageNet validation set*

**Vision Transformer**

Paper: <https://arxiv.org/abs/2010.11929>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ViT-H/14 | ViT-L/16 | ViT-L/32 | ViT-B/16 | ViT-B/32 |
| ViT | **88.08**% | 87.12% | 84.37% | 84.15% | 80.73% |

*Top1 accuracy (in %) of Vision Transformer on ImageNet* [ViT paper]

x

**CLIP**

Paper: <https://arxiv.org/abs/2103.00020>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Clip-Vit L/14 | Clip-Vit B/32 | Clip-Vit B/16 | Clip- RN101 | Clip-RN50 |
| ImageNet Accuracy | **75.3%** | 68.6% | 63.2% | 62.2% | 59.6% |

*Zero-shot performance of CLIP models over ImageNet dataset* [CLIP paper]

# Discussion

Given the previous numbers becomes essential to contextualize the loss of accuracy.

Let’s look at a serious of results deeply connected between one and another.

* LLaVA (CLIP ViT-L/14) 80%
* CLIP (ViT-L/14) 75.3%
* ViT-L/16 87.12%

The baseline is that it’s not a new state of the art model. Ref: <https://paperswithcode.com/sota/image-classification-on-imagenet>

There is indeed a loss of information starting from the vision transformer (ViT) and going through the text pipeline (LLM) the non-trivial news is:

* that Multimodal LLM do not decrease the performance of lighter models such as CLIP
  + maybe they could even increase the performance through a sort of reasoning

This model was first used and finetuded by the OpenAI team and connected to a

Considerations:

* The responses of ResNet / ViT / CLIP are constrained in a closed domain while the general purpose of the LLM has a double effect while performing image classification:
  + The first and most notable effect is the mismatch between the verbose output and the class: to construct a pipeline [image, text] => [label] we need to include a text-classification step that #1 increases the classification error and #2 includes complexity.
  + The verbose output exposes the classification process which could be a valuable information to take into consideration.

If the model is not able to perform a direct recognition of the class from the fisionomy, it might be able to derive it as the most probable option:

Classes: school\_bus, ashcan, Walker\_hound, spider\_web, file

Model Output: The object in the image is a dog, and it most likely belongs to the class "Walker\_hound," which is a specific breed of hound dog.

Or from deduction:

\n

# Conclusions

## 6.1 Summary of Findings

## 6.2 Recommendations for future research

* Fine Tuning a model to provide classification capabilities.

Expand the research on the realm of Object Detection.

\n

# 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>

[4] A Survey on Multimodal Large Language Models - [*https://arxiv.org/pdf/2306.13549v2.pdf*](https://arxiv.org/pdf/2306.13549v2.pdf)

[5] Metrics for Multi-Class Classification: an Overview - <https://arxiv.org/pdf/2008.05756.pdf>