

Group project - INFT2060

Til Naumann, Esther LE ROI, Maria GHOSH

October 21, 2025

Contents

Executive Summary	3
1 Introduction	4
2 Background and Description	5
2.1 Broader Context	5
2.2 Alternative Models	5
2.3 Model Description	5
3 Applications and Impact	5
3.1 Applications in Different Domains	5
3.2 Impact on Industry and Society	5
4 Experimental Evaluation	6
4.1 Evaluation Methodology	6
4.2 Results and Analysis	6
4.3 Discussion	6
5 Advantages and Limitations	6
6 Future Directions and Conclusion	6
6.1 Future Work	6
6.2 Conclusion	6
List of References	7
A Appendix A: Python Code	7
B Appendix B: Supplemental Material	7

Executive Summary

This report explores the capabilities and real-world adaptability of CLIP (Contrastive Language–Image Pretraining), a multimodal deep learning model developed by OpenAI in 2021. CLIP stands out for its ability to connect visual and textual understanding, learning from hundreds of millions of image–text pairs to recognize and describe new concepts without additional task-specific training. By testing it in two very different contexts, this project examines how far such general intelligence can go when applied to real challenges.

The first case study focuses on the healthcare domain, where CLIP was used to classify brain MRI scans and detect the presence of tumors. Using CLIP’s vision encoder as a frozen feature extractor and training a lightweight neural classifier on top, the model achieved a striking 96.99

The second case applies the same approach to the sustainability domain, using the Trash-Net dataset to categorize images of waste into classes such as paper, plastic, glass, metal, and cardboard. Despite the visual diversity and noise in this real-world dataset, CLIP achieved an accuracy of 92

Both implementations followed the same methodology: extracting image features with CLIP’s pre-trained vision transformer and training a compact classifier to adapt to domain-specific categories. Through iterative testing and model tuning, the results revealed how CLIP’s transfer learning capabilities can bridge tasks ranging from medical diagnostics to environmental sustainability.

Ultimately, this project highlights CLIP’s strength not only as a model, but as a foundation for applied intelligence — one that learns to see and understand in a way that feels closer to human reasoning. Yet, it also reminds us that accuracy alone is not the full measure of progress: the true challenge lies in how responsibly and meaningfully we choose to apply such technology.

1 Introduction

In recent years, artificial intelligence has made significant progress with the rise of multi-modal models. These models are systems that can understand and connect different types of data such as text, images, and audio. This ability to interpret information across multiple modalities has opened new opportunities for AI applications in diverse sectors, from healthcare to creative industries.

Developed by OpenAI in 2021, CLIP (Contrastive Language–Image Pretraining) represents an important breakthrough in this field. What is interesting about CLIP (Contrastive Language–Image Pretraining) is that unlike traditional image recognition models, it learns to link visual and textual representations by training on millions of (image, text) pairs collected from the internet. As a result, it learns by comparing correct and incorrect pairs since it can understand visual concepts described in natural language, rather than just relying on predefined labels. This enables CLIP to perform complex tasks such as zero-shot classification, image–text retrieval, and visual captioning, without requiring additional fine-tuning for each new dataset.

The purpose of this project is to analyze in depth the architecture, capabilities, and performance of CLIP, and to evaluate its potential impact across different industries. We selected two very different domains for this study: • Healthcare, through a scenario involving the detection of brain tumors from MRI scans paired with textual descriptions. Medical imaging generates vast amounts of visual data that require precise interpretation, and models like CLIP can assist radiologists. By linking textual medical reports with corresponding scans. It’s ability to understand the relationship between diagnosis and MRI images could support diagnostic accuracy and enable faster triage thus facilitating medical evaluation.

(A second domain to be confirmed — for instance, e-commerce, art and culture, or content moderation.)

Through practical experimentation, this report seeks to explore how CLIP can truly

make a difference when used as a visual search or decision-support tool. Beyond measuring its technical performance, the goal is to understand its real potential. By testing it ourselves, we hope to understand how such a model could support real human decisions and not just process data. The aim is to uncover both its strengths and its flaws. The analysis will discuss the model’s experimental performance, its strengths and limitations, and the ethical and technical challenges associated with applying CLIP in real-world environments.

2 Background and Description

2.1 Broader Context

Explain the wider context of the model, including relevant theory or existing research.

2.2 Alternative Models

Provide an overview of other related or competing models and how they compare.

2.3 Model Description

Describe the components of the given model and explain how they function on a high level.

3 Applications and Impact

3.1 Applications in Different Domains

Discuss practical applications of the model in at least two domains.

3.2 Impact on Industry and Society

Analyze the model’s potential influence on industries and broader societal implications.

4 Experimental Evaluation

4.1 Evaluation Methodology

Describe the experimental setup, evaluation criteria, and performance metrics.

4.2 Results and Analysis

Present the experimental results with appropriate tables, figures, and explanations.

4.3 Discussion

Interpret the findings and connect them to the model's intended purpose or hypotheses.

5 Advantages and Limitations

Provide a balanced discussion of the strengths and weaknesses of the model. Include any ethical, technical, or practical limitations.

6 Future Directions and Conclusion

6.1 Future Work

Identify future research opportunities, emerging challenges, or improvements for the model.

6.2 Conclusion

Summarize the main insights and conclusions drawn from your investigation.

List of References

Use proper citation style (e.g., APA, IEEE, or Harvard). Ensure all sources referenced in-text are included here.

A Appendix A: Python Code

Include the Python code used to produce your findings. Make sure the code is readable, well-commented, and reproducible.

B Appendix B: Supplemental Material

Include any extra material that supports your work — additional figures, extended data, or tables.