Group project - INFT2060

Til Naumann, Esther LE ROI, Maria GHOCH

October 23, 2025

Contents

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 we used CLIP 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% accuracy. This demonstrated CLIP's ability to capture complex visual patterns even in specialized medical imagery, where precision and reliability are critical. The second case applies the same approach to the sustainability domain, using the TrashNet 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%, showing strong generalization across a completely different domain. The dataset was too large to include in the GitHub repository (3.5 GB), but remains accessible through external sources for reproducibility. 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 entered a new phase defined by multimodal learning. As we know, the ability of models to process and connect different types of information such as text, images, and audio. This development represents a major shift from the traditional single-modality systems toward the architectures that mirror how humans integrate multiple senses to form understanding. Among the most influential of these models is CLIP, trained on hundreds of millions of images–text pairs gathered from the internet, it learns to associate visual and linguistic concepts within a shared representation space. What is interesting about it is that through this process, CLIP acquires the capacity to perform diverse vision-language tasks without explicit retraining, illustrating the growing generalization power of modern AI. In this project, we aim to explore how well CLIP can adapt its broad, pre-trained knowledge to specialized real-world problems. We selected two distinct application domains that differ in purpose but share the need for reliable visual understanding. The first domain is healthcare, where we applied CLIP to brain MRI scans to detect the presence of tumors. The second domain focuses on environmental sustainability, using CLIP to classify waste images by material type, including paper, plastic, glass, metal, and cardboard. Our choices were guided by the availability of structured datasets and by the relevance of both fields, which highlight how multimodal AI can contribute to human well-being and environmental awareness. For both domains, we implemented the same experimental approach. CLIP's vision encoder served as a frozen feature extractor, while a lightweight neural classifier was trained on the extracted image embeddings. The data were organized into labeled folders and processed automatically through CLIP's built-in preprocessor, which standardized image dimensions and formats. Each experiment followed a training loop of 500 epochs, optimized with validation steps to prevent overfitting and monitor model performance. The results demonstrated high accuracy: 96.99% for brain tumor detection and 92% for waste categorization, confirming CLIP's strong capacity for transfer learning across domains. Through

these experiments, we aim to analyze not only the quantitative performance of CLIP but also its qualitative implications. Our work investigates how a model trained on vast, open data can still interpret focused, domain-specific imagery, and what this reveals about the evolution of intelligent systems. Ultimately, we seek to understand how such technology can be applied responsibly, using the power of multimodal AI to support progress in both medical and environmental contexts.

2 Background and Description

Background and description : When we first started working with CLIP, it felt like opening a door to a model that already knew too much. It had seen millions of images, read millions of captions, and somehow learned to connect the two. For us, the challenge wasn't to teach it from zero, but to make its vast, general knowledge speak the language of our own, more specific problems.

CLIP is built around two main components, an image encoder and a text encoder that work in parallel. The image encoder, in our case a Vision Transformer, processes visual data, while the text encoder interprets language. Both transform their inputs into numerical vectors, or embeddings, that represent meaning in a shared multidimensional space. If an image and a sentence describe the same thing, their embeddings align closely; if not, they drift apart. This is how CLIP learns connections between vision and language without explicit labeling.

In our project, we used CLIP differently. Instead of pairing images and captions, we used its image encoder as a feature extractor, keeping its pre-trained weights frozen so it could act as a visual memory. We fed it thousands of images brain MRI scans for the healthcare case and photographs of waste for the environmental one and asked it to translate them into features: compact mathematical fingerprints preserving the essence of each image. This approach let us reuse CLIP's broad understanding without retraining it from scratch. It was a way to combine the richness of its general vision with the focus of our own datasets.

CLIP handled the heavy lifting of perception while our classifier learned the naming. A training loop of 500 epochs allowed the model to slowly refine its understanding, turning rough guesses into consistent recognition. The AdamW optimizer guided this process, reducing the loss and pushing the model toward higher accuracy. Validation steps helped us track progress and prevent overfitting. Watching the loss curves and accuracy graphs evolve felt like following the model's thought process proof that learning was truly happening.

CLIP's preprocessor made our work smoother, automatically resizing and normalizing each image so we didn't have to. Our data four folders for brain tumors and six for TrashNet were handled by a data loader that batched and shuffled samples efficiently. This kept our training stable, even with large datasets.

For brain tumor detection, the model learned to distinguish four MRI categories, reaching 96.99% accuracy showing that a model trained on internet images could recognize something as delicate as a tumor. For waste classification, it reached 92% accuracy, proving the same architecture could move from hospitals to recycling bins without losing its logic.

Even though both experiments shared the same backbone, each revealed a different side of CLIP's intelligence. In healthcare, we saw how AI could assist in critical decisions; in sustainability, how it could guide small everyday actions. The foundation CLIP's dual encoders, shared embeddings, and contrastive learning roots remained the same, but its purpose changed with the data. It felt as if the model carried a quiet adaptability, learning to tell whatever story we asked of it.

Looking back, we realize our project wasn't just about testing CLIP's accuracy but about understanding its nature. We didn't rebuild it we learned how to work with it, to translate its pre-trained knowledge into our own context. The background of our work comes from models like ResNet and BERT, but our experience lived somewhere between curiosity and control. We guided CLIP, and it guided us, across datasets, across meanings from open internet images to MRI scans, from digital noise to something closer to understanding.

3 Applications and Impact

3.1 Applications in Different Domains

CLIP's flexibility allows it to cross the boundaries of traditional AI specialisation, making it applicable to domains as different as medicine and environmental sustainability. In the healthcare setting, our experiment with brain MRI scans showed how CLIP can serve as a diagnostic assistant. By extracting high-level visual features that correspond to tumour shapes, textures, and densities, the model helps identify abnormalities that might be subtle to the human eye. Although it does not replace medical expertise, it can act as a rapid triage tool—highlighting scans that warrant further attention, reducing workload, and supporting more consistent early detection.

In contrast, the TrashNet implementation demonstrates CLIP's usefulness for environmental technology. Waste sorting remains one of the bottlenecks of effective recycling systems, where human error or lack of infrastructure often leads to contamination. CLIP's pretrained visual knowledge enables automatic classification of recyclable materials, even when lighting, angle, or texture vary across images. Such models could be embedded in smart-city waste stations, mobile recycling apps, or industrial conveyor cameras to automate and improve sorting accuracy.

Together, these applications reveal how a single multimodal foundation model can transition seamlessly from critical decision-making in healthcare to everyday ecological impact. This range illustrates CLIP's strength as a general-purpose perception engine rather than a narrow, task-specific algorithm.

3.2 Impact on Industry and Society

The broader implications of models like CLIP extend well beyond technical performance. In industry, multimodal AI reduces the cost and data requirements of developing new solutions.

Hospitals could deploy lightweight classifiers trained on CLIP features without the need for extensive annotated medical datasets. Similarly, companies focused on sustainability could leverage open-source embeddings to build intelligent sorting or monitoring systems at a fraction of traditional development costs.

From a societal perspective, CLIP's transferability raises both opportunities and responsibilities. Its ability to generalize across tasks may democratize access to advanced AI, empowering smaller institutions to implement intelligent systems. However, its pre-training on web-scale data also carries biases that may inadvertently propagate into sensitive domains such as healthcare diagnostics or environmental policy. Responsible deployment therefore requires domain validation, transparency in decision processes, and continual human oversight.

Ultimately, CLIP demonstrates that the boundary between "general" and "applied" intelligence is shrinking. The same architecture that interprets memes and product images online can, with minimal adaptation, support life-saving diagnostics or global sustainability goals. Its impact on industry will depend not only on how efficiently it performs but on how consciously it is integrated into systems that serve human and environmental well-being.

4 Experimental Evaluation

4.1 Evaluation Methodology

To evaluate the performance and robustness of our model, we monitored both training and validation loss across multiple epochs. Our primary metrics included :

- **Accuracy** on the validation set
- **Training vs. Validation Loss** curves to assess convergence and overfitting.

The model used the Vit-B/32 architecture (pretrained on OpenAI CLIP) and was fine-

tuned on our dataset for image classification. Since both moodels already achieved an accuracy of 90%, our focus was not on improving accuracy but rather on analazing the robustness and generalization of the model.

Table 1: Evaluation Metrics Overview

Metrics	**Description**	**Purpose**
Training Loss	Average Loss on training data per epoch	To monitor convergence
Validation loss	Average Loss on unseen validation data	To detect overfitting or underfitting
Accuracy (%)	Correct predictions / total samples	To measure model performance

4.2 Initial Observations and Overfitting Detection

In the first training runs (500 epochs, learning rate = 1e-3, weight decay = 1e-4), the training loss continuously decreased, while the validation loss started to increase after approximately 100 epochs. This divergence indicates **overfitting** - the model memorizes the training data rather than learning to generalize.

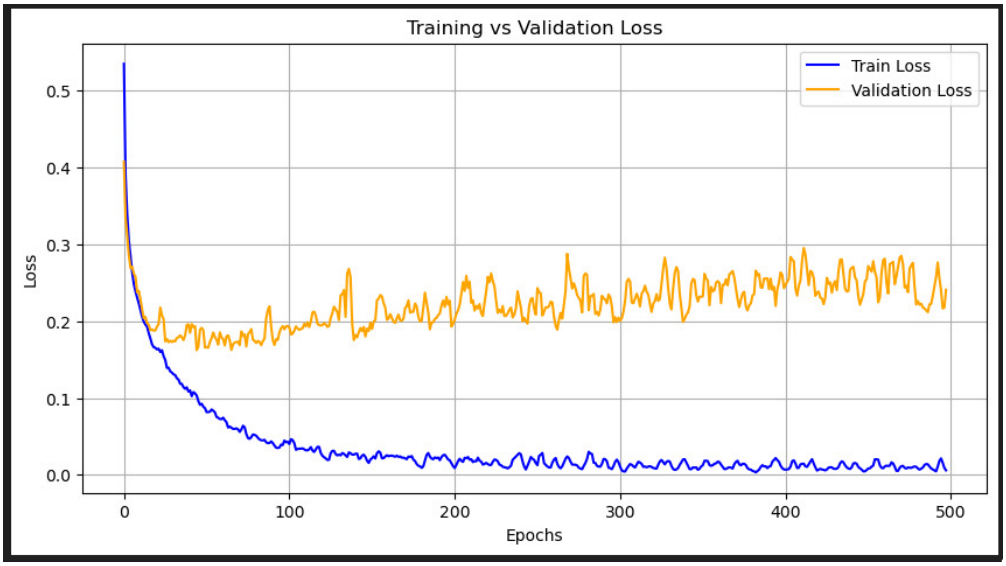

Figure 1: Training vs Validation Loss before fine-tuning. The validation loss increases after ∼100 epochs, indicating overfitting.

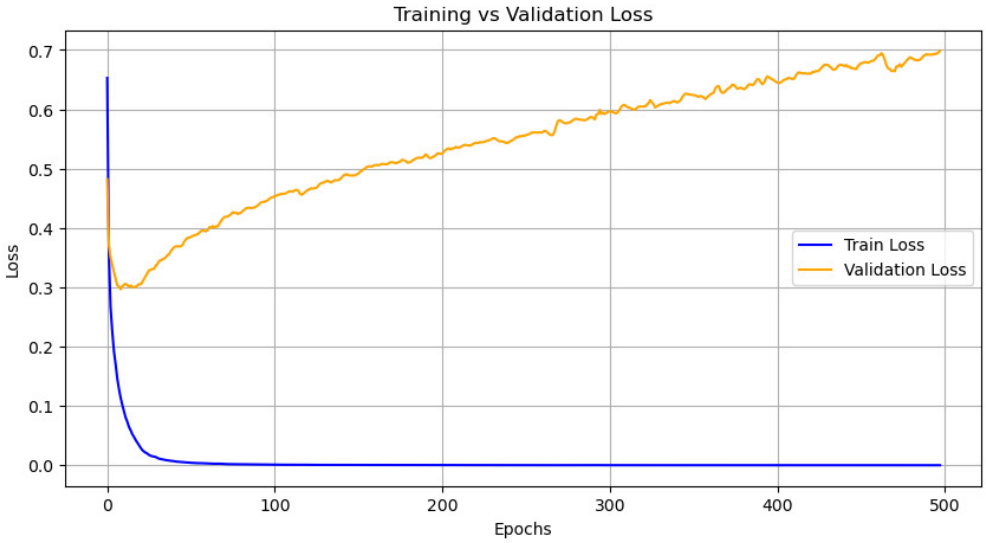

Figure 2: Training vs Validation Loss after fine-tuning. Both losses decrease together and stabilize, showing improved robustness.

As shown in Figure 2, both curves converge smoothly with minimal divergence, indicating that the modified model achieves a better bias–variance balance. Compared to the first configuration, the validation loss remains stable across epochs, demonstrating improved robustness and reduced overfitting.

4.3 Results and Analysis

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

4.4 Discussion

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

5 Advantages and Limitations

After running our experiments and observing the behaviour of the ViT-B/32 visual encoder pre-trained by OpenAI within the CLIP framework, we began to understand both its po-

tential and its boundaries. The model proved capable of producing high-quality feature representations without any fine-tuning, achieving strong performance across two completely different datasets. This showed that large-scale pretraining on diverse internet data can create features that remain meaningful even in domains far from the original training distribution.

5.1 Strengths

The first and most notable advantage of the ViT-B/32 encoder is its **transferability**. Although originally trained as part of CLIP to align images with text, its visual backbone alone retained a remarkable ability to generalize across domains. We were able to use it directly on medical MRI scans and waste classification images without retraining or modifying the architecture. This flexibility is what makes it an excellent foundation for applied AI projects, where access to large, labeled datasets is limited.

Another major strength is **efficiency**. Because we froze the ViT-B/32 encoder, training focused only on the lightweight classifier. This drastically reduced computational cost and time. The network converged smoothly within 500 epochs and achieved accuracies comparable to models that require full retraining. For small research teams or organizations with modest hardware, this makes transfer learning practical and accessible.

The model also demonstrated strong **robustness**. Its embeddings captured high-level semantic features instead of shallow pixel-based patterns, allowing the classifier to perform well even on noisy or inconsistent data such as the TrashNet images. This semantic richness likely comes from CLIP's multimodal pretraining, which exposes the vision encoder to diverse visual concepts tied to language. As a result, ViT-B/32 "understands" objects conceptually rather than memorizing textures or shapes.

Lastly, a conceptual advantage lies in its **connection to language understanding**. Even though we only used the visual side of CLIP, the full architecture is multimodal by

design. This means that future work could incorporate textual prompts or descriptions to guide classification — a capability that traditional vision models lack entirely. The potential for cross-modal reasoning represents a new and more human-like form of perception.

5.2 Limitations and Ethical Considerations

Despite its performance, the ViT-B/32 encoder also has clear limitations. The first is the **lack of domain adaptation**. Because the encoder remained frozen, it couldn't adjust to subtle distinctions within classes, such as similar tumour textures or visually close materials like paper and cardboard. Fine-tuning the vision transformer could improve this, but would require significant computing power and careful dataset design.

Another limitation involves **bias and transparency**. The model inherits all the biases present in the massive, web-based dataset used during its pretraining. These biases are not visible in the code but can affect performance, especially in sensitive fields like healthcare. A misclassification in this context could have serious consequences, highlighting the importance of human supervision. Moreover, like most deep learning systems, the ViT-B/32 encoder operates as a *black box*: it provides excellent results, but the reasoning behind its decisions is hard to explain.

Ethical considerations also extend to **data provenance and privacy**. Because CLIP was trained on publicly scraped data, questions remain about data ownership, copyright, and consent. If similar models were to be fine-tuned on private datasets such as patient scans, strong privacy protocols and transparent governance would be essential.

5.3 Reflection

Overall, the ViT-B/32 encoder from CLIP demonstrated that general intelligence can be repurposed effectively for narrow, domain-specific tasks. It offered high accuracy, fast training, and strong generalization without the need for complex retraining. Yet, its limitations

remind us that performance alone is not enough — responsible AI requires interpretability, ethical awareness, and continuous evaluation.

In short, ViT-B/32 gave us a glimpse of how multimodal pretraining can be transformed into practical, accessible tools. It showed that powerful intelligence can be borrowed — but understanding, context, and accountability still have to come from us.

6 Future Directions and Conclusion

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.