Multimodal Idiomaticity Representation

1. Introduction

Understanding idioms is challenging for computational models. For instance, *"bad apple"* can mean a literal fruit or a corrupting individual. While humans use context and background knowledge to distinguish meanings, language models struggle, especially when integrating vision.

This project explores multimodal idiomaticity representation by selecting the image that best represents an idiom's intended meaning within a given sentence. By combining visual and linguistic understanding, the goal is to enhance computational comprehension of figurative language.

2. Methodology

Our approach integrates:

- **Data Augmentation with T5-small:** Generating diverse image captions to improve generalization.
- **Zero-shot CLIP Predictions:** Using CLIP's dual encoding to match text-based idioms with images without explicit training.
- **Fine-tuning CLIP:** Enhancing performance with original and augmented training data.

The pipeline included preprocessing, augmentation, and inference through zero-shot and fine-tuned models. Performance was evaluated using Accuracy, F1 Score, Precision, and Recall.

3. Results

Model	Accuracy	Precision	Recall	F1 Score
Zero-Shot CLIP model	33.33	33.33	100	50
CLIP with Fine Tuning (Original Train)	81.48	100	44.44	61.54
CLIP with Fine Tuning (Augmented Train) - unfreezing last few layers	40.74	23.08	33.33	27.27
CLIP with Fine Tuning (Augmented Train) - unfreezing last layer	55.56	42.86	100	60

Figure 1 : Model Performance metrics

From Figure 2, we can see that fine-tuning CLIP significantly improved performance. The zero-shot model had **33.33% accuracy** and **100% recall**, misclassifying all cases as idiomatic, leading to a low **F1 score (0.5)**. Fine-tuning on the original dataset boosted accuracy to **81.48%**, with an **F1 score of 61.54%**, showing strong improvements.

Augmented training data yielded mixed results. **Unfreezing just the last layer** retained **100% recall** but lowered accuracy to **55.56%**, as the model became biased toward idiomatic meanings. **Unfreezing multiple layers** further reduced accuracy to **40.74%**, due to overfitting on the augmented dataset. These findings suggest that while augmentation can enhance model robustness, excessive fine-tuning will lead to unintended biases.

Overall, fine-tuning on the **original dataset** produced the best results, excelling with common idioms but struggling with rarer or abstract ones. The model demonstrated a solid grasp of visually intuitive

idioms but had occasional difficulty interpreting those with more abstract meanings.

Figure 2: Model Prediction

4. Key Findings & Discussion

- LLM-Augmented Data Improves Robustness: T5-small generated diverse samples, making the model more adaptable to different idiom contexts.
- Fine-Tuning Outperforms Zero-Shot: Task-specific training significantly boosted accuracy and interpretability.
- Multimodal Learning Enhances Understanding: Combining text and images helped the model recognize nuances often missed by single-modality systems, particularly in idioms with clear visual associations.

5. Conclusion

This project demonstrates the potential of combining T5 and CLIP for multimodal idiom understanding. Fine-tuning significantly improved image-idiom matching, pushing models toward more context-aware and human-like interpretation of figurative language. Future work could explore larger LLMs for more diverse augmentation, as well as attention-based alignment mechanisms to better associate idiomatic phrases with visual representations. By continuing to refine multimodal approaches, we move one step closer to AI systems that understand language the way humans do.