RE-STRUCTURING CLIP'S LANGUAGE CAPABILITIES

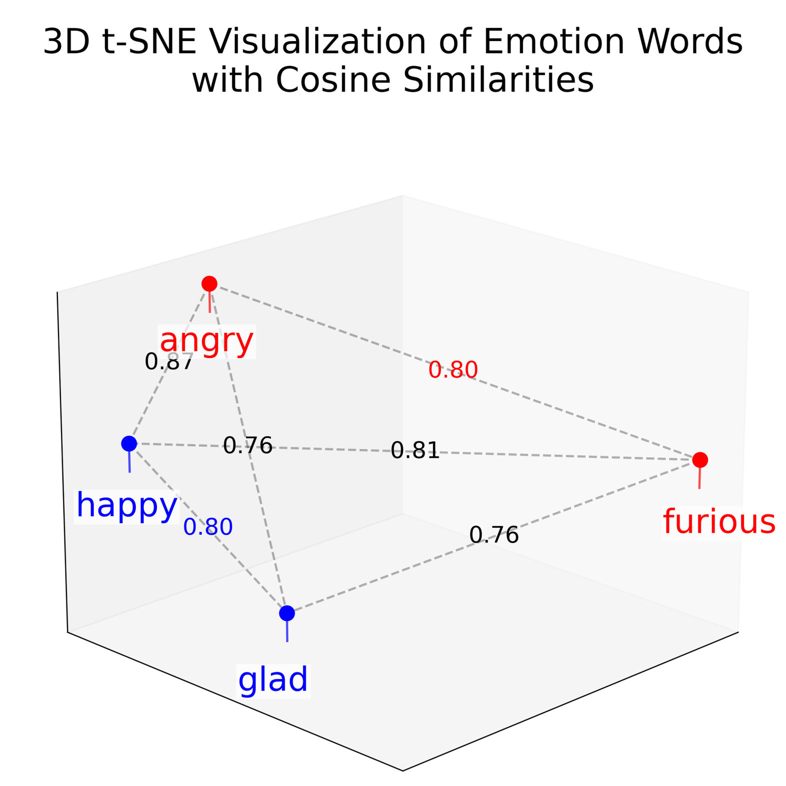
Zhiqi Gao and Frederic Sala

Department of Computer Sciences, University of Wisconsin-Madison

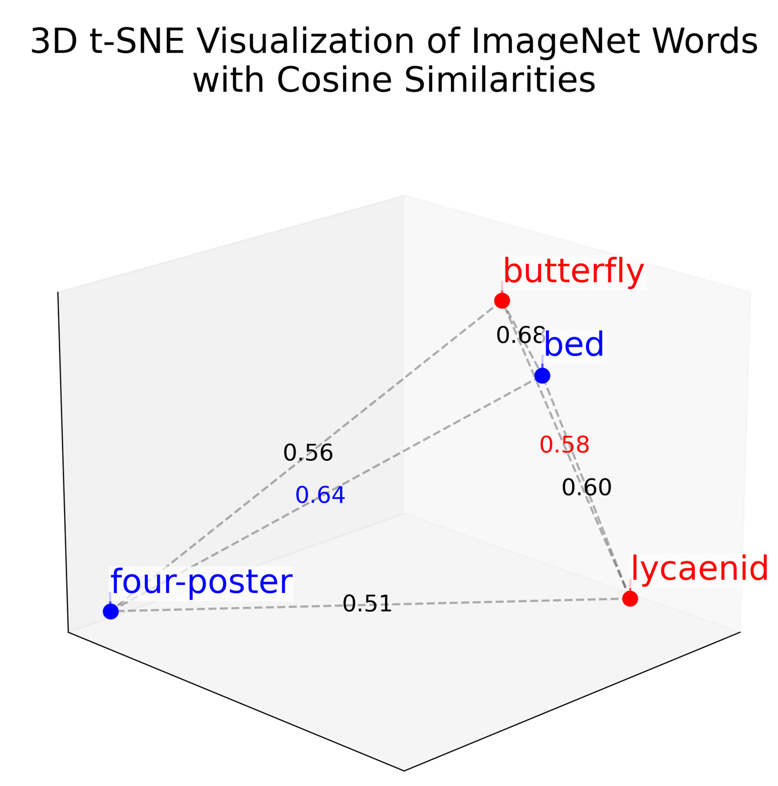
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Background & Motivation

CLIP's text encoder is tuned for image-text alignment, *not* language structure, making it sensitive to linguistic variations. For example, synonyms and antonyms **do not** behave as desired:



"angry" is closer to "happy" than "glad" is to "happy"



"butterfly" is closer to "bed" instead of its hyponym "lycaenid"

Question: Can we modify CLIP in a way that brings back its **structural understanding of language**, while still maintaining its alignment with image representations?

Methodology: Fine-Tuning with a Semantic Loss

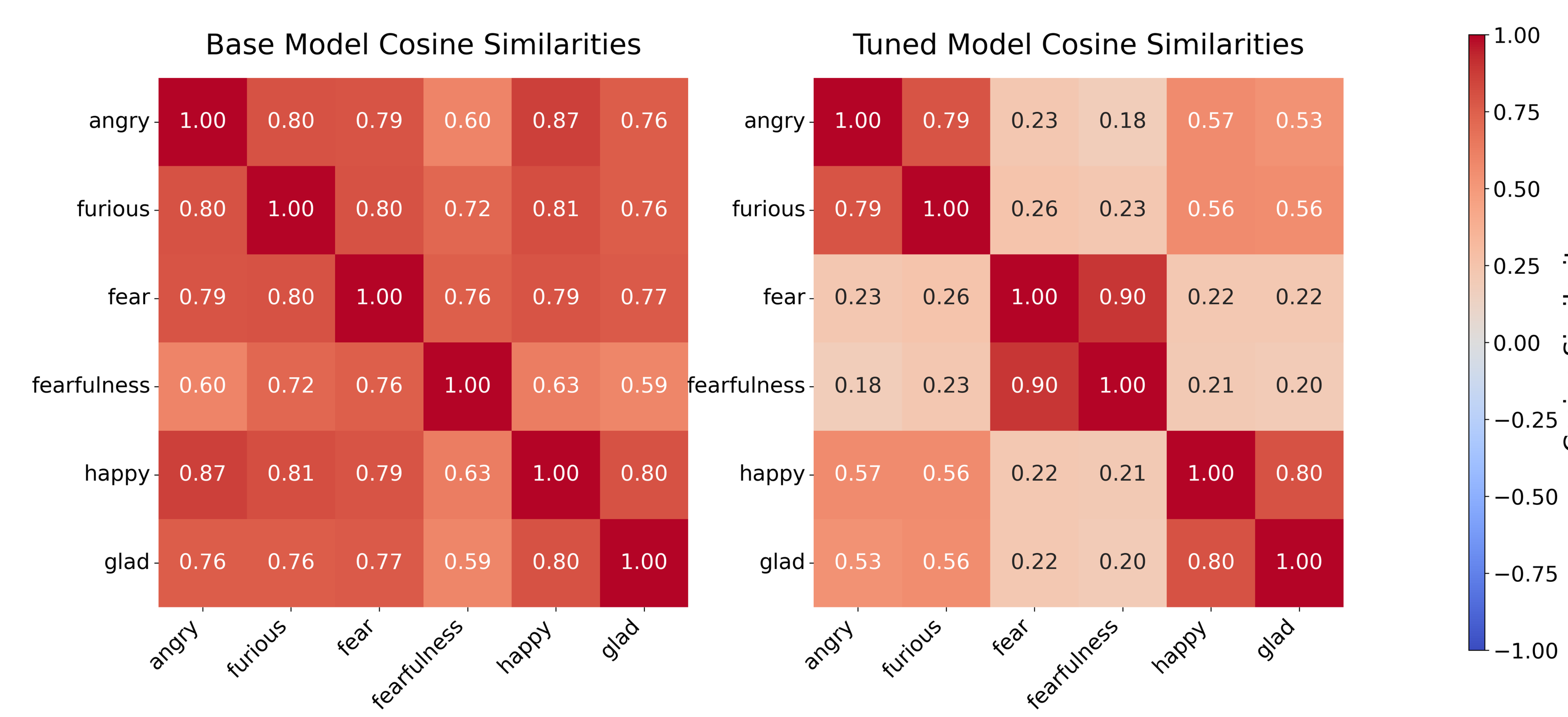
We fine-tune CLIP's text encoder by exploiting the **semantic hierarchy of WordNet** to rebuild its understanding of language structure with no image data and minimal computation overhead.

Our Goal: Craft a loss that with components corresponding to our two goals:

- Distance Loss ($\mathcal{L}_{\text{distance}}$): Reflect semantic relationships using **Wu-Palmer Similarity (s_{wup})**
- Regularization Loss ($\mathcal{L}_{\text{reg}}$): Prevents significant deviation

$$\mathcal{L} = \underbrace{\left(s_{wup}(w_i, w_j) - \cos\theta\left(M(w_i), M(w_j)\right)\right)^2}_{\mathcal{L}_{\text{distance}}} + \lambda \underbrace{\text{MSE}\left(M(w), M_0(w)\right)}_{\mathcal{L}_{\text{reg}}}$$

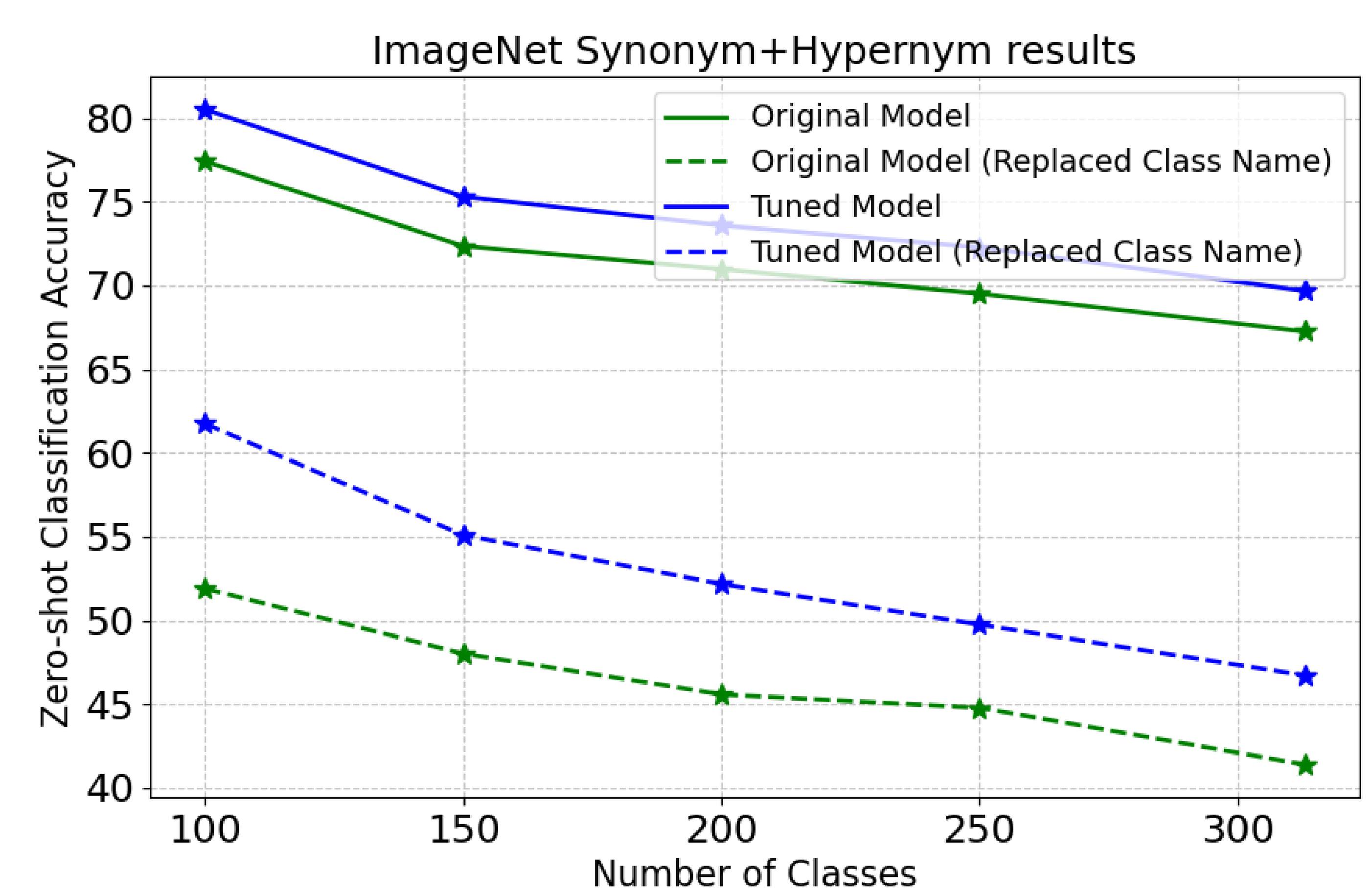
Examples



Our method helps **align the word vector space.**

Results: Zero-Shot Accuracy Gains

Our method yields **consistent classification accuracy improvement** with both settings in **ImageNet, OpenImage, and FER2013**.



Accuracy improves on original and synonym/hypernym-replaced class names.

Results: Generalization Abilities

Demonstrating Generalization: We show that a model **fine-tuned on text set A can improve the performance on task B**, which shows the model is not overfitting.
Specifically, we evaluate performance on the OpenImage subset with different models trained on ImageNet Texts

Classification accuracy comparison for different models

Model	**Synonyms - 93 classes**		**Hypernyms - 150 classes**	
	Orig. Acc	Repl. Acc	Orig. Acc	Repl. Acc
Original Model	75.95	46.37	72.78	25.73
OpenImage-Tuned	**78.78**	52.67	74.15	29.75
ImageNet Hypernym	77.74	52.16	74.95	**30.00**
ImageNet Synonym	77.78	**52.98**	75.00	28.56
ImageNet Mixed Set	**78.78**	52.56	**75.93**	29.64

Summary

1. **A Structure-Based Fine-Tuning Method for CLIP's Text Encoder Using Hierarchical Information**
2. **Improved Zero-Shot Classification Accuracy and Robustness to Linguistic Variations**

Future Directions

- **Scalability & Polysemy:** Challenges including a large polysemy portion and decreasing marginal gains in applying the method to the entire WordNet structure.
- **Image-Caption Datasets:** Adapt the methodology for image-caption datasets like LAION for broader applicability.
- **Limitations with Propositional Words:** Frameworks like CLIP struggle with terms such as *not*, *is a*, and *more/less than*, which is included in complex semantic relationships

Check our Blogpost!