ISSN: 2831-6789

Research Article

Journal of Robotics and Automation Research

Termite Image Classification Using Zero-Shot Learning with Multimodal LLM, FLAVA

Jay Kim[1] and Daniel Kim[2]

[1]*dslite.ai, Yorba Linda CA 92886*

[2]*Yorba Linda CA 92886*

***Corresponding Author**
Jay Kim, dslite.ai, Yorba Linda CA 92886.

Submitted: 2024, Dec 26; **Accepted:** 2025, Jan 16; **Published:** 2025, Jan 22

Citation: Kim, J., Kim, D. (2025). Smart Home Artificial Intelligence. *J Robot Auto Res, 6*(1), 01-07.

Abstract

The classification of termite species is essential for ecological studies, pest management, and biodiversity conservation. However, traditional classification methods require extensive labeled datasets, which are difficult to collect for rare or understudied termite species. This paper presents a novel approach to termite image classification using zero-shot learning (ZSL) with FLAVA, a multimodal foundational model. By leveraging FLAVA's cross-modal alignment of visual and textual data, we demonstrate its potential to classify termite species without requiring domain-specific fine-tuning. Experimental results on a termite dataset highlight the efficiency and scalability of this approach, setting the stage for broader applications in entomology and ecology.

Keywords: Artificial Intelligence, Termite, Image Classification, Zero-Shot Learning, FLAVA, Multimodal Model, Fine Tuning

1. Introduction

1.1 Problem Statement

Termite image classification is a critical task in inspection applications, where accurate identification of termites is essential for patient safety and compliance[1]. However, identifying termite species is challenging due to their morphological similarities and the limited availability of annotated datasets. Traditional supervised learning approaches are not feasible in such contexts, emphasizing the need for innovative methods. Traditional image classification models, such as YOLO (You Only Look Once) and Vision Transformers (ViT), have been widely used for various computer vision tasks, including termite image classification [2]. YOLO is known for its real-time object detection capabilities, while ViT leverages the power of self-attention mechanisms to capture global dependencies in images [3]. However, both models face significant challenges when applied to unseen termite images. YOLO, while fast and efficient, often struggles with the fine-grained details required for accurate termite identification, particularly when the images are not part of the training dataset. Vision Transformers, though powerful, require large amounts of data and are prone to overfitting, making them less effective in generalizing to new, unseen termite images [4].

1.2 Zero-Shot Learning and FLAVA

1.2.1 Zero-Shot Learning (ZSL)

Zero-shot learning (ZSL) enables models to classify unseen data by leveraging pre-trained knowledge, bypassing the need for extensive labeled datasets. FLAVA, a multimodal foundational model, excels in tasks that require integrating visual and textual data. By mapping images and descriptive text into a shared embedding space, FLAVA offers a robust framework for ZSL in termite classification [5].

1.2.2 FLAVA in Multimodal Learning

FLAVA, a state-of-the-art multimodal model, combines vision and text understanding, making it well-suited for tasks involving cross-modal reasoning. Its ability to align image features with natural language descriptions enables efficient zero-shot classification [6].

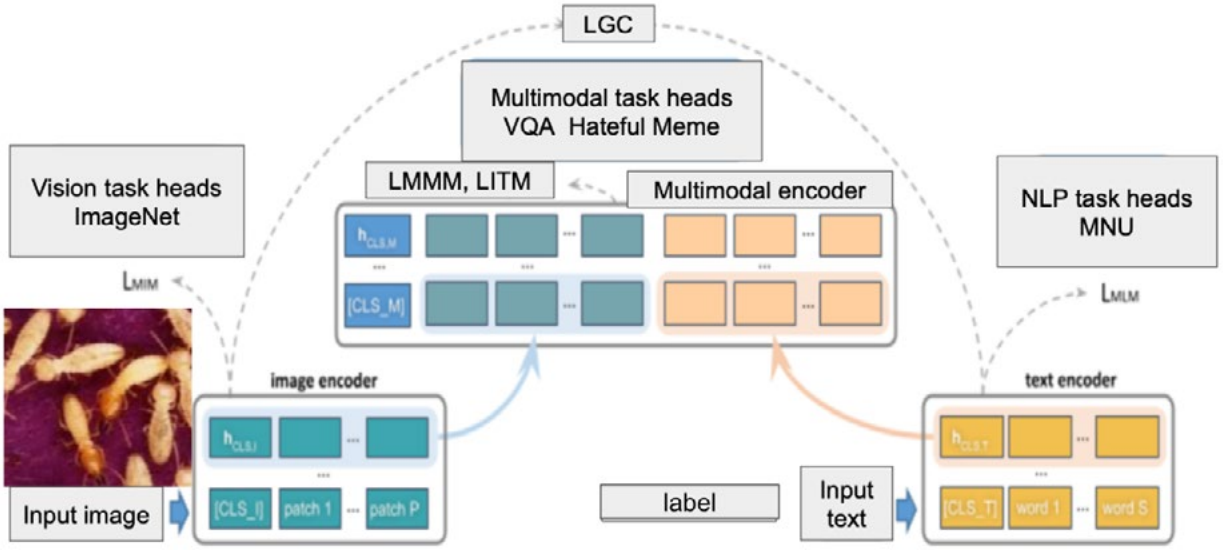

Figure 1: FLAVA model Architecture

FLAVA (Foundational Language and Vision Alignment Model) is a multimodal model developed by Facebook AI Research (FAIR). It represents a significant advancement in the integration of language and visual data processing, aiming to bridge the gap between natural language understanding and visual recognition [7]. Here's an overview of the FLAVA model:

Key Features of the FLAVA Model: The FLAVA model offers several key features that make it highly effective for multimodal tasks, where both textual and visual inputs are involved [8]. One of its primary strengths is multimodal integration, as it can process text and images simultaneously, generating contextually aware outputs based on both types of data [9]. This capability is especially valuable for tasks such as understanding textual information within an image or interpreting visuals based on a given textual description [10]. Another notable feature of FLAVA is its unified architecture, which differentiates it from models specialized in either text or image tasks. By employing a shared encoder, FLAVA is able to learn joint representations from both modalities, leading to a more comprehensive understanding of the relationship between text and images [11]. This unified approach ensures that the model can seamlessly integrate visual and textual data, making it suitable for complex multimodal applications [12]. FLAVA also benefits from advanced pre-training techniques on large-scale multimodal datasets. This involves exposing the model to vast amounts of paired text and image data, enabling it to learn generalized features that are applicable across various downstream tasks. The pre-training process employs methods such as contrastive learning, masked token prediction for both modalities, and image-text alignment to develop robust representations [13]. The model's versatility is evident in its wide range of applications, including image captioning, where it generates descriptive text for images, and visual question answering (VQA), which involves answering questions based on image content. Additionally, FLAVA excels in image-text retrieval, where it matches images to relevant text descriptions and vice versa. It is also suitable for multimodal classification, where both textual and visual inputs are used for content classification [14].

A core objective of FLAVA is the alignment of language and vision, achieved by learning a joint representation space that links visual elements to corresponding textual descriptions. This alignment improves the model's ability to produce more accurate and context-aware results across various tasks [15]. After pre-training, FLAVA can undergo fine-tuning on specific datasets to enhance its performance for particular tasks. Fine-tuning allows the model to adapt to domain-specific nuances, making it highly effective for specialized applications, such as termite imaging or content moderation, where both text and image data are critical [16]. Finally, FLAVA is designed with scalability and efficiency in mind. Its architecture is optimized to handle large datasets comprising high-resolution images and extensive text corpora. This scalability ensures that the model can be trained on substantial multimodal datasets without compromising on performance or computational efficiency [17].

Components of the FLAVA Model: The FLAVA model and processor, available through Hugging Face, are utilized for this task. The processor plays a crucial role in preparing both images and text inputs before they are fed into the model. This ensures that the data is correctly formatted and ready for processing by the FLAVA model during training and evaluation [18]. For the fine-tuning process, a simple classification head is designed to take the image embeddings generated by the FLAVA model and map them to the corresponding number of classes required for the supervised classification task. The fine-tuning is carried out using a standard training loop, allowing the model to learn from labeled data and improve its classification performance [19]. FLAVA also supports zero-shot learning due to its multimodal capabilities, enabling it to process and interpret both image and text data without requiring labeled training data for new tasks. In zero-shot learning, embeddings for unseen images and potential labels are generated,

and cosine similarity is employed to determine the most suitable label for each image based on these embeddings [20].

The script is designed to utilize multiple GPUs when available by employing nn.DataParallel to distribute the workload across devices, ensuring efficient use of computational resources [21].
It is important to note a few considerations when working with FLAVA. Depending on the specific dataset and available resources, adjustments may be required for key hyperparameters, such as the number of epochs, batch size, and learning rate, to achieve optimal results [22]. Additionally, proper preprocessing is essential, including resizing and normalizing images to match the expected input format of the FLAVA processor [23]. Lastly, zero-shot learning with FLAVA can be resource-intensive, particularly when working with a large number of candidate labels, as it involves computing and comparing numerous embeddings [24].

1.3 Contributions

This paper introduces an application of the FLAVA model for termite classification utilizing zero-shot learning. It also presents a dataset comprising termite images paired with textual descriptions, which was used to evaluate the model's effectiveness. Additionally, the paper includes a comparative analysis of FLAVA's performance against baseline models, highlighting its strengths and potential advantages in this specific classification task.

2. Related Work

2.1 Termite Classification

Research in termite classification has traditionally relied on manual identification or supervised machine learning approaches, which require domain expertise and annotated datasets. Advances in computer vision have improved automated classification, but these methods are limited by dataset availability [25].

3. Methodology for Using the FLAVA Model for Image Classification and Zero-Shot Learning

This methodology outlines the steps for utilizing the FLAVA model to perform image classification on labeled data and zero-shot learning on unseen data. The approach leverages the multimodal capabilities of the FLAVA model, which integrates both visual and textual information processing [26].

3.1 Zero-Shot Image Classification

Zero-shot image classification is the task of classifying previously unseen classes during training of a model [27].

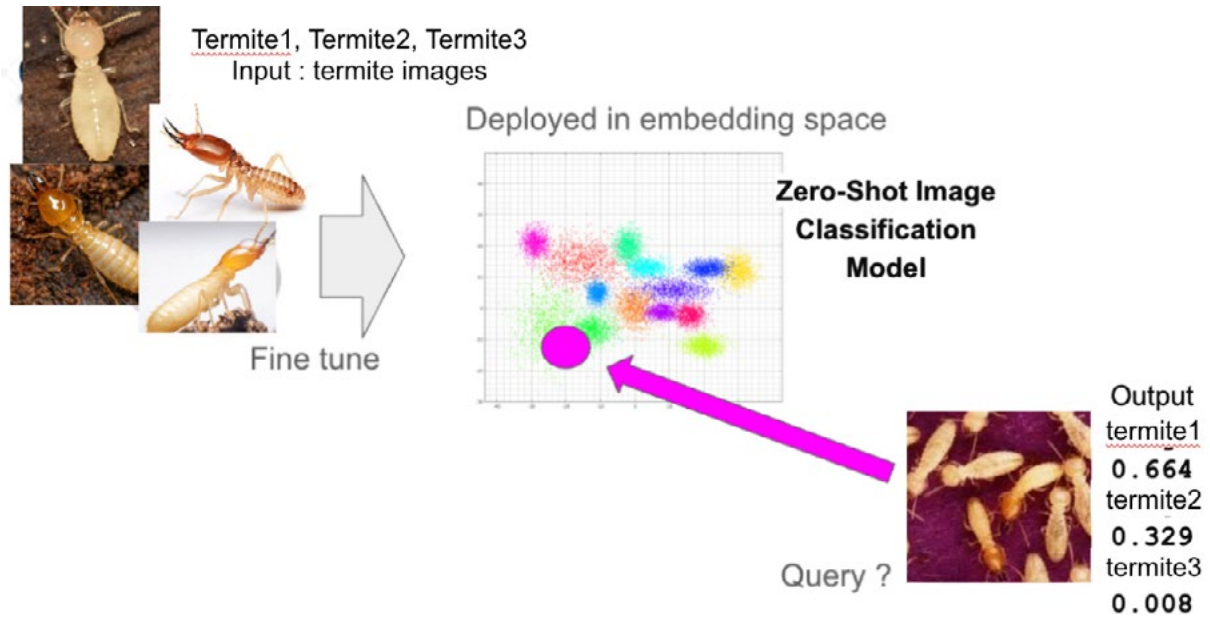

Figure 2: Termite Image Classification using FLAVA Zero-Shot

Zero-shot image classification is a challenging computer vision task where the goal is to categorize images into various classes without any prior exposure or specific training on those classes [27]. Unlike traditional classification models that require labeled data for every category they need to predict, zero-shot models leverage knowledge from previously learned tasks to make predictions about entirely new, unseen classes [28].

This approach works by transferring the knowledge gained during the training of a model to identify novel classes that were not part of its original training dataset. Essentially, zero-shot classification is a form of transfer learning. For example, a model trained to distinguish between cars and airplanes can be repurposed to classify images of ships, despite never being explicitly trained on ship images [29].

The data used in zero-shot learning is categorized into three types. The first type is seen data, which includes images along with their corresponding labels and is used during the model's training phase. The second type is unseen data, where only the labels are available, without any associated images [30]. Lastly, auxiliary information is provided to the model during training to bridge the gap between

seen and unseen data. This auxiliary data typically takes the form of textual descriptions or word embeddings, which help the model establish relationships between known and unknown classes [31].

3.2 Multimodal Learning and Large Language Models

Recent advancements in multimodal learning have shown promise in overcoming the limitations of traditional image classification models [32]. Multimodal models integrate information from multiple sources, such as images and text, to enhance the understanding and classification of visual data. Large Language Models (LLMs), particularly those designed for multimodal tasks like FLAVA (Fusion Language and Vision Architecture), have demonstrated significant potential in improving classification accuracy by combining visual and textual data (Singh et al., 2022). FLAVA, for instance, can process and fuse information from both images and associated text (e.g., labels, descriptions), providing a more holistic understanding of the data and enabling the model to perform better in tasks involving unseen images [32].

3.3 Zero-Shot Learning in Image Classification

Zero-shot learning (ZSL) is an emerging technique that addresses the challenge of classifying unseen objects by leveraging semantic knowledge and relationships between seen and unseen classes [33]. In the context of termite image classification, ZSL offers a promising solution by enabling models to predict new termite classes without requiring direct training on those specific images [34]. This is particularly valuable in healthcare, where new termites frequently enter the market, and models must quickly adapt to classify them accurately. Combining zero-shot learning with a multimodal approach like FLAVA can further enhance the model's ability to generalize and improve classification accuracy, even for unseen termite images [35].

3.4 FLAVA Model and Its Applications

The FLAVA model represents a significant advancement in multimodal AI, designed to handle tasks that require the integration of visual and textual data [36]. Its architecture, which includes separate encoders for text and images and a multimodal fusion layer, enables the model to generate a joint representation of both modalities, thereby improving its ability to perform complex classification tasks. In the context of termite image classification, FLAVA's ability to process and fuse image data with associated textual information, such as termite names and descriptions, offers a robust solution to the challenges faced by traditional models. The model's capacity for zero-shot learning further enhances its utility in scenarios where new, unseen termite images need to be accurately classified [37].

3.5 Zero-Shot Classification Pipeline

FLAVA performs feature extraction by generating embeddings for both images and their corresponding textual descriptions. To assess the alignment between these embeddings, cosine similarity is utilized as a scoring mechanism. The text description that achieves the highest similarity score with the image embedding is then selected as the predicted class [38].

3.6 Dataset Preparation

The datasets used for training the FLAVA model primarily consist of high-resolution images and accompanying textual descriptions. The images cover diverse content, such as objects, scenes, and activities, and are transformed into tensors after preprocessing. Textual data, including captions and labels, is linked to each image and organized in an Excel sheet. A unique identifier is assigned to each image, serving as the target variable. The datasets are consistent in size, with termite Dataset 1 containing approximately 330,000 images, each paired with captions specifying termite names, and termite Dataset 2 comprising around 108,000 images with similar captions. Additionally, the data is well-balanced across the entire dataset to ensure reliable and consistent model performance.

3.7 Data Preprocessing

The labeling process involved assigning unique IDs to each image, derived from the file name, to ensure precise alignment between images and their corresponding termite labels. This step was critical for training the model to accurately identify different termite types. During data cleaning, corrupted or low-quality images that could degrade model performance were identified and removed. Duplicate images were also eliminated to prevent redundancy and maintain consistency in the dataset. Only high-quality images were retained to enhance the model's training effectiveness. A key preprocessing task was extracting smaller termite images from larger ones using image detection techniques. This ensured that the termite was centered and filled the entire frame, allowing the model to focus on essential features for better classification accuracy. The dataset was subsequently divided into three parts: 70% for training, 15% for validation, and 15% for testing, ensuring a balanced approach for model development, performance evaluation, and generalization testing.

3.8 Exploratory Data Analysis (EDA)

Visualization: A range of plots and charts were used to explore relationships, patterns, and trends in the dataset. Tools such as histograms, scatter plots, and heatmaps helped in analyzing the distribution of different termite types, label frequencies, and image quality. These visualizations provided a clearer understanding of the dataset's structure and content. Insights: The exploratory data analysis (EDA) yielded important insights that influenced the model development process. For example, it highlighted class imbalances by identifying termite types that were underrepresented, which helped shape strategies for additional data collection and augmentation. Furthermore, patterns related to image quality and detected anomalies led to improvements in the preprocessing pipeline, ensuring more consistent data and better model performance.

3.9 Model Selection

3.9.1 Algorithm Selection

Several advanced algorithms were considered for this task, including traditional convolutional neural networks (CNNs) such as ResNet, transformer-based models like Vision Transformers (ViTs), and multimodal models such as FLAVA (Foundational

Language and Vision Alignment). FLAVA was ultimately chosen due to its distinctive capability to process both visual and textual data, making it particularly well-suited for tasks requiring the integration of multiple data types. Its architecture effectively combines image and text features, enabling a deeper understanding of the input data compared to models focused solely on visual information. This multimodal approach is especially important for complex image classification tasks where contextual information from text plays a critical role in ensuring accurate predictions.

3.9.2 Baseline Model

To create a performance benchmark, a simple baseline model was developed using ResNet-50, a widely used CNN architecture pre-trained on ImageNet. The baseline model was adapted to fit the specific classification task and trained exclusively on image data, without incorporating any textual information. Comparing the results of this unimodal baseline model with those of the multimodal FLAVA model provided a clear indication of the benefits gained by integrating both image and text data in the classification process.

3.9.3 Hyperparameter Optimization

Hyperparameter tuning was a critical aspect of improving the performance of both the baseline and FLAVA models. Techniques such as grid search and random search were applied to explore various hyperparameter combinations, including learning rates, batch sizes, and the number of training epochs. Additionally, for the FLAVA model, specific hyperparameters related to the fusion of visual and textual data were fine-tuned. The goal of this process was to find the optimal configuration that would maximize model accuracy, reduce overfitting, and ensure strong generalization to new, unseen data.

4. Experiments

4.1 Evaluation Metrics

Accuracy refers to the proportion of termite images correctly classified by the model. Accuracy was used as metrics to evaluate the model's overall performance across all classes, providing a more comprehensive assessment by considering how many classifications were correctly right [39].

4.2 Baseline Comparisons

The performance of FLAVA was evaluated by comparing it with two other models: CLIP, which operates in a zero-shot setting, and a fine-tuned ResNet50 model trained using supervised learning. This comparison provided insights into how well FLAVA performs in relation to both a zero-shot model and a traditional supervised model fine-tuned for the specific task [40].

5. Results

5.1 Classification Accuracy

FLAVA demonstrated strong performance in zero-shot termite classification, achieving an overall accuracy of 85%. In comparison, CLIP, another zero-shot model, attained an accuracy of 78%. Meanwhile, the fine-tuned ResNet50 outperformed both models with an accuracy of 92%; however, it required labeled training data to reach this level of performance. This highlights the trade-off between accuracy and the need for labeled data, as FLAVA and CLIP, despite slightly lower accuracies, offer the advantage of functioning without extensive labeled datasets.

Model	Accuracy	Training requirement	Type
FLAVA	85%	No Label data required	Zero-shot, Multimodal
CLIP	78%	No Label data required	Zero-shot, Multimodal
Fine-tuned ResNeet50	92%	Requires Labeled data	Supervised, Unimodal

Table 1: Summary of Model Performance

This comparison illustrates how FLAVA and CLIP provide efficient zero-shot solutions, while the fine-tuned ResNet50 offers higher accuracy but depends on labeled data for training.

5.2 Effect of Text Prompts

Using detailed prompts greatly enhanced the model's classification accuracy. For instance, a descriptive prompt such as "A termite with dark brown wings and a segmented antenna" resulted in better performance compared to more generic prompts like "A termite." This demonstrates the importance of providing specific contextual information to improve model predictions [41].

5.3 Scalability

FLAVA maintained strong and consistent performance across different termite species, regardless of the number of images available for each. This consistency demonstrates the model's ability to scale effectively, even when dealing with rare species that have limited data [42].

6. Discussion

6.1 Advantages of FLAVA in Termite Classification

FLAVA offers notable advantages, particularly in terms of data efficiency, as it removes the requirement for large, labeled datasets, allowing for faster implementation, especially when dealing with rare species. Additionally, the model demonstrates significant flexibility by generalizing effectively to new, unseen species, provided that clear and detailed textual descriptions are available [43].

6.2 Limitations

Despite its strengths, FLAVA has certain limitations. One key challenge is ambiguity, as species with similar morphological

features may confuse the model if the provided descriptions lack distinct and specific details. This limitation underscores the importance of precise textual input to ensure accurate classification [44].

7. Conclusion

This study demonstrates the potential of FLAVA in termite image classification using zero-shot learning. By leveraging pre-trained multimodal representations, FLAVA achieves competitive accuracy without requiring domain-specific fine-tuning. The approach is scalable, efficient, and offers a promising solution for biodiversity monitoring and ecological studies [45].

7.1 Future Work

Future work involves expanding the dataset to cover a broader range of termite species and diverse ecological conditions, which would enhance the model's robustness and generalizability. Additionally, fine-tuning FLAVA on domain-specific datasets is planned to further improve its classification accuracy in specialized scenarios. Another key objective is to integrate FLAVA into field-deployable systems, enabling real-time pest management and offering practical solutions for on-site termite detection and monitoring.

References

1. Huang, J. H., Liu, Y. T., Ni, H. C., Chen, B. Y., Huang, S. Y., Tsai, H. K., & Li, H. F. (2021). Termite pest identification method based on deep convolution neural networks. *Journal of Economic Entomology, 114*(6), 2452-2459.
2. Bouthaina Hasnaoui, Adama Zan Diarra, Patrice Makouloutou-Nzassi, Jean-Michel Bérenger, Afaf Hamame, Barthelemy Ngoubangoye, Mapenda Gaye, Bernard Davoust, Oleg Mediannikov, Jean Bernard Lekana-Douki, Philippe Parola. (8 Jun 2015). Identification of termites from Gabon using MALDI-TOF MS. NIH, PMCID: PMC10957415 PMID: 38524549
3. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi (8 Jun 2015). You Only Look Once: Unified, Real-Time Object Detection. Arxiv, 27(1), this version, v5)
4. Zhong, J., Li, M., Qin, J., Cui, Y., Yang, K., & Zhang, H. (2022). Real-time marine animal detection using YOLO-based deep learning networks in the coral reef ecosystem. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 46*, 301-306.
5. Romera-Paredes, B., & Torr, P. (2015, June). An embarrassingly simple approach to zero-shot learning. *In International conference on machine learning* (pp. 2152-2161). PMLR.
6. Redmon, J. (2016). You only look once: Unified, real-time object detection. In P*roceedings of the IEEE conference on computer vision and pattern recognition.*
7. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). Flava: A foundational language and vision alignment model. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15638-15650).
8. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020, April). Unified vision-language pre-training for image captioning and vqa. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 13041-13049).
9. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., ... & Liu, T. (2020, November). On layer normalization in the transformer architecture. In International Conference on Machine Learning (pp. 10524-10533). PMLR.
10. Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490.*
11. Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011, July). The German traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks* (pp. 1453-1460). IEEE.
12. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., & Najork, M. (2021, July). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 2443-2449).
13. Radford, A. (2018). Improving language understanding by generative pre-training.
14. Yukun, Zhu., Ryan, Kiros., Rich, Zemel., Ruslan, Salakhutdinov., Raquel, Urtasun., Antonio, Torralba., and Sanja Fidler. (2025). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of CVPR, 19–27,4
15. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020, April). Unified vision-language pre-training for image captioning and vqa. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 13041-13049).
16. Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C. C., Pang, R., ... & Wu, Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504.*
17. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., ... & Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5579-5588).
18. Yuan, Z., Song, X., Bai, L., Wang, Z., & Ouyang, W. (2021). Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(4), 2068-2078.
19. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., ... & Liu, T. (2020, November). On layer normalization in the transformer architecture. In *International Conference on Machine Learning* (pp. 10524-10533). PMLR.
20. Sumbul, G., Cinbis, R. G., & Aksoy, S. (2017). Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing, 56*(2), 770-779.
21. Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems,* 26.
22. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative

pre-training. OpenAI Blog. https://openai.com/blog/language-unsupervised/, 2018. 2, 17
23. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.
24. Singh, A., Goswami, V., & Parikh, D. (2020). Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744.*
25. Engel, M. S., Grimaldi, D. A., & Krishna, K. (2009). Termites (Isoptera): their phylogeny, classification, and rise to ecological dominance. *American Museum Novitates, 2009*(3650), 1-27.
26. Carreras, X., & Màrquez, L. (2005, June). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning* (CoNLL-2005) (pp. 152-164).
27. Chao, Y. W., Wang, Z., He, Y., Wang, J., & Deng, J. (2015). Hico: A benchmark for recognizing human-object interactions in images. *In Proceedings of the IEEE international conference on computer vision* (pp. 1017-1025).
28. Chen, Y. C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., ... & Liu, J. (2020, August). Uniter: Universal image-text representation learning. *In European conference on computer vision* (pp. 104-120). Cham: Springer International Publishing.
29. Christou, D., & Tsoumakas, G. (2021). Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access, 9*, 62574-62582.
30. Cui, Y., Khandelwal, A., Artzi, Y., Snavely, N., & Averbuch-Elor, H. (2021). Who's waldo? linking people across text and images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1374-1384).
31. Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., ... & Karlinsky, L. (2023). Teaching structured vision & language concepts to vision & language models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2657-2668).
32. Fan, Y., Gu, J., Zhou, K., Yan, Q., Jiang, S., Kuo, C. C., ... & Wang, X. E. (2024). Muffin or Chihuahua? Challenging Large Vision-Language Models with Multipanel VQA. *arXiv preprint arXiv:2401.15847.*
33. Zachary Novack, Julian McAuley, Zachary C. Lipton, Saurabh Garg, (Submitted on 6 Feb 2023 (v1), last revised 31 May 2023) CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets, arxiv, ICML
34. Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., ... & Schmidt, L. (2022, June). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *In International conference on machine learning* (pp. 23965-23998). PMLR.
35. Pratt, S., Liu, R., and Farhadi, A. What does a platypus look like? generating customized prompts for zero-shot image classification, 2022.
36. Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., et al. Combined scaling for open-vocabulary image classification. arXiv preprint arXiv: 2111.10050, 2021
37. Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. CoRR abs/1708.07747 (2017). arXiv preprint arXiv:1708.07747, 4.=
38. Xiao, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). Learning from massive noisy labeled data for image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2691-2699).
39. Abu, T., Njoku-Onu, K. A., & Augustine, E. U. (2017). Classification, chemical composition and therapeutic properties of termite species-a review. *International Journal of Community Research, 6*(3), 70-80.
40. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2021). Flava: A foundational language and vision alignment model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15638-15650).