Seeing Apples as Apples, Not as a Soup of Matrices: Image Classification Through the Composition of Tailored CLIP Task Vectors on Isolated Visual Attributes

Student: Evam Kaushik
Student ID: a1909167
Supervisor: Dr. Cristian Rodriguez-Opazo
Co-Supervisor: Dr. Bernard Evans

1 Introduction

The central question this research seeks to address is: *"Can all the knowledge, which can be encoded as language, be learned through non-interactive, observational means—i.e., by simply consuming labeled examples without any direct intervention or manipulation of the system from which the data arise?"*[1]

This inquiry leads to an intriguing thought experiment. Consider a scenario where someone has never encountered an apple. They are told everything about it in exhaustive detail: the shape, the color, the tactile feel of the skin, the crunchiness of its flesh, the sweetness of its juice, and the scent it emits. Armed with this linguistic description alone, can they form a mental image of an apple? More challenging still, what if this person has aphantasia—the inability to visualize mental images? Can they still piece together the concept of an apple from pure linguistic detail and, if their vision were suddenly granted (or restored), would they be able to recognize an actual apple in a photograph or in real life? Similarly, consider a person who has been blind since birth, taught all the concepts of the world through language. If their sight were restored, would they be able to visually identify and describe entities based solely on the verbal knowledge they have accumulated?

Bridging Philosophical Inquiry to Practical Methodology: To explore this profound question within the realm of machine learning, we leverage state-of-the-art models that embody a similar paradigm of learning from language and visual data. The Contrastive Language-Image Pretraining (CLIP) model [1] is designed to understand and associate images with their corresponding textual descriptions, effectively learning from a vast corpus of image-text pairs. This aligns closely with the thought experiment, where the model must form internal representations of visual concepts based solely on linguistic information paired with images, without direct manipulation or interaction with the visual data.

However, while CLIP excels at associating images with text, its internal representations remain largely opaque, posing challenges for interpretability. This is where Low-Rank Adaptations (LoRA) [2] come into play. LoRA provides a parameter-efficient method to adapt large pre-trained models like CLIP by injecting low-rank updates into the transformer layers. By integrating LoRA with CLIP, we can create modular, attribute-specific adaptations that enhance both the interpretability and performance of the model. This modularity is crucial for dissecting and understanding how different visual attributes (such as color, texture, and shape) contribute to the model's overall comprehension and classification capabilities.

In essence, CLIP serves as the foundational model that embodies the learning paradigm under investigation, while LoRA offers the tools necessary to dissect and interpret the model's internal mechanisms in a granular and attribute-specific manner. This combination allows us to empirically test whether the model's internal representations of visual concepts align with human-interpretable attributes, thereby addressing the core philosophical question of whether all language-encodable knowledge can be learned through non-interactive, observational means.

Naturally, a subsequent line of inquiry emerges: *"How many directions exist in language with respect to semantic concepts? How many of these learned concepts are human-interpretable (e.g., shapes, colors, textures)? Does a language model internally build a hierarchy of primitive, compositional features in a manner analogous to humans?"* These questions become especially relevant as we focus on visual attributes. By training models on image-associated data, we seek to determine the degree to which language-grounded representations overlap with human-understandable visual concepts.

[1] By non-interactive, we refer to learning from passively observed examples (such as watching a documentary or reading detailed descriptions) rather than actively exploring or experimenting with the environment. For instance, this would mean learning about car manufacturing solely by watching factory documentaries instead of visiting the factory, stopping the assembly line, and interacting with the workers. Similarly, a typical machine learning model 'learns' objects by seeing labeled images of those objects without physically handling them, as humans often do.

In pursuit of this, our research aims to enhance both the interpretability and the performance of CLIP [1] by integrating multiple LoRA models [2]. Specifically, we propose training several LoRA models—each focused on different visual compositional attributes (e.g., color, texture, size)—to produce attribute-specific task vectors [13]. These compositional visual features are chosen to be directly human-interpretable, thereby enabling us to examine whether a model's internal representations align with the hierarchical, concept-based structure that human learners naturally develop.

Prior Research

CLIP Model and LoRAs

Contrastive Language-Image Pretraining (CLIP) [1]is a neural network that can learn the associations between images and text using a large number of image-and-text pairs for training images. It combines visual representations with natural language descriptions so that zero-shot classification can be performed. It has demonstrated success in learning visual representations from natural language supervision. However, its interpretability is a hinderance due to its internal representations.

Low-Rank Adaptations (LoRA)[2] offers a parameter-efficient method for adapting large language models by injecting low-rank updates into the transformer layers. By applying LoRA models onto CLIP layers, we can learn model weights θ for specific compositionality tasks.

Compositionality in Visual Recognition

Compositionality refers to task vectors combining to form new models. Task vectors embed specific knowledge that is learned during fine-tuning which, when composed, retain the ability to address multiple tasks efficiently.

Following the recent advancements in model merging and knowledge fusion [11], we now know that task vectors, which are built by subtracting the weights of a pre-trained model from the weights of the same model after fine-tuning on a task, play an important part in preserving specific knowledge and building upon concepts [13]. Since task vectors can be combined together through arithmetic operations such as negation and addition, it suggests that we can combine task vectors made of attribute-specific LoRA models to compose complex visual concepts from simpler primitives. We can gain further control over the emergence of each primitive by utilising anisotropic scaling of task vectors [14].
The Neural Algebra of Classifiers[3] is a framework to combine basic visual features through Boolean algebra operations, allowing us to modularly build concepts of increasing complexity. We suspect that there exists a strong correlation between the level of simplicity a concept can be broken down into, especially into visual attributes, and its intrinsic dimensions[15].

Lastly, Zero-Shot Learning[4] the task of classifying unseen classes, is an effective approach mitigate the limitations of supervised learning pertaining to novel or rare classes that the user has not trained on. In the scope of the project, such a learning pattern is relevant as it allows the model to identify and differentiate the unusual configurations of given attributes thereby making use of the learnt visual primitives. By proposing the model attributes in terms of these primitives, we give the model a combinatorial perspective of visual attributes, which aids in performing on new classes. This approach attempts to address the question of what the model is 'seeing' based on the activations of the LoRA models, thereby improving explainability and cognitive alignment of the model to humans.

Proposed Methodology

Attributes Extraction and Clustering: Creating Compositional Dictionary Dataset [3]

- Collecting a list of attribute names.
- Performing preprocessings such as lemmitization.
- Project attributes into a vector space using an embedding model like BERT, Word2Vec etc.
- Cluster attributes based on semantic similarity.

We then select the clusters with the most suitable examples of human-parsable features like colours, shapes, and textures.

Model Architecture

We propose integrating multiple LoRA models into the transformer layers of multiple CLIP models with each specialising in a specific visual attribute identified from the clustering process. Some of the examples include:

- **Color LoRA Model**: Not only one of the simplest human-parsable primitive, but also one with enough evidence to suggest that incorporating colour information can enhance the performance of deep neural networks in image classification tasks[6].
- **Shape LoRA Model**: Primitive shapes are learnable low-level features that can be utilised to aid deep neural networks in image classification[7].

Training Procedure

Once we have prepared a dataset of visual attributes, the training process then consists of the following steps:

1. **Attribute-Specific Training**: Training each LoRA model on its respective visual attribute using the annotated datasets and creating its specific task vector.
2. **Anisotropic Task Vector Combinations**: These task vectors are then combined anisotropically to represent compex tasks. [12]

Datasets

We will utilize attribute-annotated datasets such as Caltech-UCSD Birds-200-2011[5], Stanford Cars[8], Labeled Faces in the Wild (LFW)[9], and The Visual Genome Dataset[10].

These datasets provide the necessary annotations to train the LoRA models to produce desired task vectors effectively.

Evaluation Metrics

Evaluation of will be based on:

- **Interpretability**: A qualitative analysis of the model output corresponding with the associated task vector will be used to infer its degree of involvement in performing classification [16].
- **Classification Performance**: Model accuracy is measured for standard and zero-shot classification compared to baseline CLIP.

Experimental Design

The experimental plan includes:

1. **Baseline Evaluation**: Record the standard CLIP model's performance on the selected datasets.
2. **LoRA Model Training and Task Vector Generation**: Training LoRA models on the selected attributes identified from the clusters to create attribute specific Task Vectors.
3. **Anisotropic Scaling of Task Vectors**: Creating enhanced models by anisotropically combining Task Vectors.
4. **Performance Comparison**: Compare the enhanced model's performance and against baseline CLIP.

Supervisor Engagement

Weekly meetings are scheduled with with Dr. Cristian Rodriguez-Opazo and Dr. Bernard Evans which has helped outline the scope of this project. A Discord channel is used for all communication, including paper presentation and sessions on knowledge transfer.

Expected Outcomes

We anticipate that the integration of attribute-specific LoRA models will:

- Create a compositional dictionary of several Task Vectors tuned to visual attributes.
- Improve the classification performance in zero-shot learning.
- Assess the model's ability to construct a hierarchy of visual primitives similar to human cognition.

By focusing on visual primitives and their compositionality, our method aims to address the challenges of zero-shot classification [4] and improve the overall robustness of visual recognition systems.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. *Learning Transferable Visual Models From Natural Language Supervision.* In ICML, 2021.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. *LoRA: Low-Rank Adaptation of Large Language Models.* In ICLR, 2021.

[3] Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. *Neural Algebra of Classifiers.* In CVPR, 2018.

[4] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata. *Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[5] P. Welinder, S. Branson, T. Mita, et al. *Caltech-UCSD Birds 200.* Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[6] A. C. Chou, H. C. Lu, and C. S. Chen, "ColorNet: Investigating the Importance of Color Information in Deep Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[7] J. Wang, Y. Sun, J. Shao, and X. Wu, "DeepShape: Deep Learned Shape Descriptor for 3D Shape Retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 155–162.

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. *3D Object Representations for Fine-Grained Categorization.* In 3D Representation and Recognition Workshop, ICCV, 2013.

[9] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[10] Ranjay Krishna, et al. *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.* International Journal of Computer Vision (IJCV), 2017.

[11] F. Wan, X. Huang, D. Cai, X. Quan, W. Bi, and S. Shi, "Knowledge Fusion of Large Language Models," arXiv, Jan. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.10491.

[12] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing Models with Task Arithmetic," ICLR 2023. [Online]. Available: https://arxiv.org/abs/2212.04089.

[13] E. Yang, Z. Wang, L. Shen, S. Liu, G. Guo, X. Wang, and D. Tao, "AdaMerging: Adaptive Model Merging for Multi-Task Learning," arXiv, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2310.02575.

[14] F. Z. Zhang, P. Albert, C. Rodriguez-Opazo, A. van den Hengel, and E. Abbasnejad, "Knowledge Composition using Task Vectors with Learned Anisotropic Scaling," arXiv, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.02880.

[15] Yuanzhi Li, et al. *Measuring the Intrinsic Dimension of Objective Landscapes.* In ICLR, 2018.

[16] S. Menon and C. Vondrick, "Visual Classification via Description from Large Language Models," *arXiv preprint arXiv:2210.07183*, 2022.

A Non-Experimental/Non-Invasive Supervised Observation

Non-Experimental or Non-Invasive Supervised Observation refers to the process of learning the mapping of features of a data point without interacting with the data point or the system it inhabits, thereby avoiding potentially changing the configuration of the system.

For example, consider learning about the manufacturing process of a car through two different approaches:

1. **Non-Interactive Approach**: Watching extensive documentaries about car manufacturing.
2. **Interactive Approach**: Visiting the car factory, interacting with the assembly line by stopping and inspecting parts in real-time, and asking supervisors interactive questions.

In the non-interactive approach, the learner passively consumes information without altering the system or engaging directly with the process. In contrast, the interactive approach involves active participation, which can lead to a deeper and more nuanced understanding through direct manipulation and inquiry.