Bachelor Thesis Topic: Automated Semantic Discovery and Zero-Shot Labeling via VLM-Enhanced Clustering

1 Introduction and Background

In the classical era of machine learning, clustering on datasets like MNIST relied on low-level visual patterns (e.g., pixel gradients or shapes). While successful, these methods are "semantics-blind"—they group data based on mathematical distance without understanding the underlying concepts. Matching these clusters to pre-defined labels was a manual, post-hoc process that assumed the clustering logic perfectly mirrored human categorization, which is rarely true for complex, real-world images.

The emergence of Vision-Language Models (VLMs) such as CLIP has fundamentally shifted this paradigm. VLMs provide a semantically grounded latent space where visual features are pre-aligned with natural language descriptions. However, this new space introduces a novel challenge: Semantic Bias. A VLM might cluster a "Racing Car" and an "Ambulance" differently not because of their wheels, but because of their functional context.

This thesis investigates the transition from "visual grouping" to **"semantic discovery."** By applying density-based clustering (DBSCAN) within the VLM embedding space, we aim to explore whether the model's internal "world knowledge" can be leveraged to automatically organize unlabeled data and assign human-readable labels without any fine-tuning. Unlike the naive MNIST approach, this research focuses on the **alignment quality** between cluster centroids and hierarchical text prompts, and investigates how to use VLM as an **automatic evaluator** to refine clustering boundaries—effectively creating a self-correcting unsupervised learning pipeline.

2 Tasks & Goals

- **Literature Review**: Study the principles of Vision-Language Models and density-based clustering algorithms like DBSCAN.
- **Feature Extraction Pipeline**: Implement an embedding extraction module using at least two VLM backbones (e.g., CLIP-ViT and SigLIP).
- **Clustering Implementation**: Apply **DBSCAN** to the extracted embeddings. Unlike K-Means, DBSCAN does not require a pre-defined number of clusters, which is essential for discovering unknown categories.
- **Zero-Shot Labeling**: Develop a matching mechanism that compares the **centroid** of each discovered cluster against a pre-defined candidate label list (derived from dataset metadata) using cosine similarity.
- **Evaluation**:
    - Quantify clustering quality using **Adjusted Rand Index (ARI)** and **Normalized Mutual Information (NMI)** against ground-truth labels.
    - Assess the accuracy of the **Auto-Labeling** component by measuring the top-1 match rate between cluster centroids and correct category names.

- **Discussion**: Analyze "Semantic Purity"—investigating whether images within the same cluster share the same high-level concept even if they vary in visual appearance (e.g., different perspectives of a "Car").

4 Expected Outcomes

- A Python-based framework that automatically clusters and labels unlabeled image datasets.
- A comparative study demonstrating that VLM-based embeddings yield more "human-interpretable" clusters than traditional CNN features.
- A reproducible codebase utilizing `scikit-learn` for clustering and `OpenCLIP` for feature extraction.

5 Requirements

- Study in the field of Computer Science or Data Science.
- Proficiency in **Python** programming.
- Experience with machine learning libraries (`scikit-learn`, `PyTorch`).
- Basic understanding of **Vector Latent Spaces** and similarity metrics (Cosine Similarity).

Contact

If you are interested, please send your CV / self-introduction and transcripts to lan@dbs.ifi.lmu.de

References

- **[1] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996).** *A density-based algorithm for discovering clusters in large spatial databases with noise.* In KDD.
- [2] Radford, A., et al. (2021). *Learning Transferable Visual Models from Natural Language Supervision.* In Proceedings of ICML.
- **[3] Shao, J., Pu, L., et al. (2023).** *Deep Clustering with Concrete Vision-Language Antecedents.* In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).