Bachelor Thesis Proposal: Label-Free Evaluation of Image–Text Fusion for Unsupervised Clustering

Shubhangi

October 2026

Problem Statement

Modern pre-trained vision language models such as CLIP provide semantic representations of text and images that can be used for downstream tasks without training an encoder from scratch. When both these image and text representations are created for a data sample, they can be clustered individually or combined into a multi-modal representation. But combining modalities does not necessarily improve clustering. A strong image representation may be degraded by noisy or weak text information, while complementary text and image representations may produce better cluster structure when fused. In an unsupervised setting, where ground truth labels are not available for model selection, it is difficult to determine in advance whether fusion is beneficial.

This thesis will investigate when pretrained image and text representations are suitable to be combined for unsupervised clustering and whether this decision can be supported using label-free properties of the embedding spaces.

Focus will be placed on frozen pre-trained encoders and simple, interpretable fusion strategies rather than training new multi-modal models. GT labels will be used only for final evaluation.

Tasks

The student is expected to carry out the following tasks:

- Conduct a structured literature review on pretrained vision-language representations, multimodal clustering, embedding-quality evaluation, and recent approaches combining CLIP representations with clustering.
- Develop a modular and reproducible evaluation pipeline for extracting image and text embeddings, combining them using fusion strategies, clustering the resulting representations and then measuring clustering quality.
- Evaluate suitable pretrained representation models, with CLIP as the main vision-language baseline and image- and text-specific encoders such as DINOv2 and Sentence-BERT as comparison models.
- Perform experiments on at least two public paired image–text datasets. Recommended datasets include: N24News, containing news articles and associated images across 24 semantic categories and UPMC Food-101, containing paired food images and textual information across 101 categories.
- Compare clustering based on image embeddings alone, text embeddings alone, and simple multi-modal combinations such as normalized concatenation, weighted fusion, and shared-space averaging for compatible vision-language models.
- Use a standard clustering algorithm such as K-means as the primary clustering method and evaluate clustering results using established external measures such as Normalized Mutual Information, Adjusted Rand Index, and clustering accuracy.
- Investigate a small set of label-free representation-quality measures, including internal clustering quality, effective rank, clustering stability, and cross-modal neighbourhood agreement. Analyse whether these label-free properties are predictive of whether multimodal fusion improves or degrades clustering compared with the best individual modality.
- Develop a simple and interpretable label-free diagnostic or ranking method for selecting between image-only, text-only, and fused representations.

- Compare the proposed diagnostic against simple baselines such as always selecting the fused representation, internal clustering indices, effective rank, or clustering stability alone.

Expected Outcomes

- A reproducible benchmark for evaluating pretrained image, text, and fused representations for unsupervised clustering.
- A systematic comparison of representative pretrained models and simple image–text fusion strategies on public multimodal datasets
- An analysis of the conditions under which image–text fusion improves or harms clustering quality.
- A controlled study of the effect of noisy or incorrect multimodal correspondence on clustering.
- A simple label-free diagnostic for selecting between unimodal and fused representations without access to class labels.
- Practical recommendations for using pretrained vision-language embeddings in unsupervised multimodal clustering.
- A complete thesis, documented source code, experiment configurations, and reproducibility instructions.

The expected contribution goes beyond applying clustering algorithms to pretrained embeddings. The thesis should provide a controlled experimental analysis of multimodal fusion and investigate whether the usefulness of fusion can be predicted from the unlabeled embedding spaces themselves.

Contact

If you are interested, please send an email to shubhangi@dbs.ifi.lmu.de with your CV and transcript, and briefly describe your interest in this topic or any modifications to suit your research direction.

References

- Q. Garrido, R. Balestriero, L. Najman, and Y. LeCun, "RankMe: Assessing the Downstream Performance of Pretrained Self-Supervised Representations by Their Rank," *Proceedings of the 40th International Conference on Machine Learning*, 2023.
- A. Tsitsulin et al., "Unsupervised Embedding Quality Evaluation," 2023.
- S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of Neural Network Representations Revisited," *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- Z. Ma et al., "MoDE: CLIP Data Experts via Clustering," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Zeng et al., "CLIP-Enhanced Contrastive Clustering for Remote Sensing Image Retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.