4 Track 4: Text-Guided Image Clustering

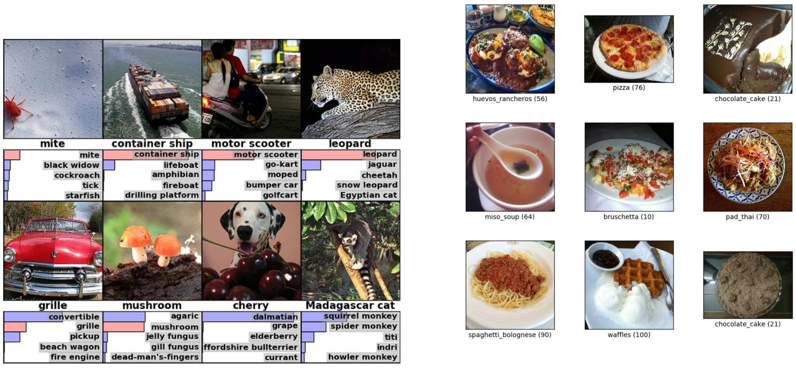
4.1 Introduction

The objective of this assignment is to explore whether transforming images into textual representations can enhance clustering performance. This involves extracting a variety of classical and deep learning-based image features, generating text descriptions from images, and analyzing clustering capability. The clustering performance will be evaluated using the Adjusted Rand Index (ARI) metric.
Contest and Dataset at Kaggle.

4.2 Dataset

A subset of **30 classes**, with 500 images per class, will be selected from two distinct datasets:

- **ImageNet (Non-Competitive Section)**: A large-scale dataset comprising millions of images spanning thousands of categories.
- **Food-101 (Competitive Section)**: A dataset consisting of 101,000 images across 101 food categories, commonly used for fine-grained image classification.



(a) Sample from ImageNet (b) Sample from Food-101

Figure 2: Example images from datasets used in this assignment.

4.3 Non-Competitive Section

4.3.1 Preprocessing

To standardize the data, the following preprocessing steps will be applied:

- **Dataset Splits:**
  - **Train:** 60% - features + labels provided.
  - **Validation:** 20% - features + labels provided.
  - **Test:** 20% - features only (labels hidden).
- Set a fixed random seed *(782)* to ensure reproducibility.
- Normalize pixel values to a standard range.
- The dataset combined is 1GB+. In case you feel that a smaller sample is sufficient, feel free to report and continue.

4.3.2 Feature Extraction

Feature extraction techniques will be categorized into classical and deep learning-based methods.

Classical Feature Extraction Participants will implement and analyze the effectiveness of the following classical feature extraction techniques:

- **SIFT (512D)**: A keypoint-based representation capturing distinct regions in an image.
- **HOG (256D-512D)**: Extracts shape and texture information based on gradient orientations.
- **Color Histogram (64D-128D)**: Captures color distribution across different channels.
- **Canny Edge Detection**: Identifies object boundaries by detecting edges.
- **Local Binary Patterns (LBP)**: A texture descriptor based on pixel intensity differences.

Deep Learning-Based Feature Extraction A pre-trained CNN will be used to extract embeddings from images:

- **ResNet-50 (2048D)**: Extract features from the penultimate layer's average-pooled feature map.

Analysis Task: Compare the clustering performance of classical features alone and in combination with deep learning features.

4.4 Generating Text Descriptions and Features

To introduce text-guided clustering, the following steps will be performed:

1. Generate image captions using **BLIP Image Captioning Base** Model.
2. Extract textual embeddings from these captions using **SBERT (768D)**.

Analysis Task: Compare clustering performance when using only image-based features versus text-based features.

4.5 Clustering

Clustering will be conducted using:

- **K-Means Clustering** (baseline method).
- **One additional clustering technique** (DBSCAN, Agglomerative, or Spectral Clustering).

Experiments:

- Clustering based on only image features.
- Clustering based on only text features.

Visualization: Apply t-SNE to visualize cluster separability.

5 Competitive Section

The competitive section encourages participants to experiment with advanced techniques while adhering to runtime and memory constraints. All the models combined should train and run in **10 hours**, and size (including pre-trained) should be less than **4 GB / 1 Billion parameters**. Exceptions can be discussed on case-by-case basis.

5.1 Guidelines

Participants are free to explore:

- Alternative CNN architectures such as EfficientNet or MobileNet.
- Advanced textual representations including VQA Prompts, BLIP and T5 embeddings.
- Multi-modal fusion techniques for clustering.
- Alternative clustering algorithms beyond the baseline methods.

5.2 Submission and Evaluation

- Submissions will be ranked based on **ARI scores** on a given test dataset.
- Innovation in feature engineering, clustering techniques, and fusion strategies is encouraged.

6 Submission Requirements

Final submissions must include:

- Source code for feature extraction, clustering, and evaluation.
- A comprehensive report covering methodology, experiments, and observations.
- Precomputed feature datasets in CSV format to facilitate reproducibility.
- Clustering visualizations, including t-SNE plots, to illustrate cluster separability.

Final Note: The report should provide insights from comparing classical, deep learning, and text-guided features in image clustering.

6.1 References

For additional resources on feature extraction, clustering, and multi-modal approaches, refer to:

- Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research.
- Feature Extraction in Image Processing.
- Image Captioning via BLIP (Kaggle Notebook).
- Exploring Text-Guided Image Clustering (Arxiv).
- **Lightweight Image Captioning Models**:
  - ViT-GPT2 Image Captioning (Hugging Face)
  - BLIP-Base Image Captioning.

- **Lightweight Visual Question Answering (VQA) Models**:
  - BLIP-VQA (Hugging Face).
  - MMF (Multi-Modal Framework by Facebook Research).

Good luck, and happy clustering!