## **Kanchana T**

**2022UG000062**

**MultiModel Machine Learning**

**Assignment - 1**

**Multimodal Image Classification Using Traditional Feature Extraction**

## **1. Introduction**

The task of multimodal learning combines multiple sources of information (e.g., images and text) to improve classification performance. In this project, we explore multimodal classification using the **MS-COCO 2017 dataset**. Each image is paired with human-annotated captions and multiple object category labels.

The primary goal is to compare **unimodal image features**, **unimodal text features**, and **multimodal fused features** for multi-label classification. Instead of deep neural models, we deliberately use **traditional feature extraction** (Canny edges for images, Word2Vec embeddings for text) and classical machine learning (Random Forest classifiers).

* Benchmark unimodal vs multimodal performance.
* Critically analyze strengths and limitations of feature choices and fusion strategy.
* Provide insights into how traditional pipelines handle complex multimodal data.

## **2. Preprocessing**

### **Dataset**

* **Images:** 5,000 validation images from MS-COCO 2017.
* **Captions:** COCO captions (captions\_val2017.json).
* **Labels:** Object categories from COCO instances (instances\_val2017.json).

### **Image Preprocessing**

1. Images resized to **64×64**.
2. Converted to grayscale.
3. **Canny edge detection** applied (threshold1=100, threshold2=200).
4. Flattened to 1D vectors → **4096-dimensional feature vectors**.
5. Normalized to [0,1].

### **Text Preprocessing**

1. Captions lowercased, punctuation removed.  
   Tokenized using **NLTK**.
2. Trained a **Word2Vec model (vector size=300, window=5)** on all captions.
3. For each caption → mean of word embeddings.
4. For each image → mean of its caption embeddings → **300-dimensional vector**.

### **Labels**

* 80 object categories extracted from COCO instances.
* Multi-label binarization applied → label vectors of size **80 per image**.

### **Data Splits**

* **Train:** 70% (3500 samples)
* **Validation:** 10% (499 samples)
* **Test:** 20% (1001 samples)

## **3. Methodology**

1. **Unimodal Image:** Canny edge features only.
2. **Unimodal Text:** Word2Vec caption embeddings only.
3. **Multimodal Fused:** Concatenation of image + text features → **4396 dimensions**.

### **Classifier**

* **Random Forest (per class):** One binary classifier trained per category.
* Parameters: n\_estimators=100, random\_state=42, n\_jobs=-1.

### **Evaluation Metrics**

* **Hamming Score:** Average per-label correctness per sample.
* **Subset Accuracy:** Fraction of samples with all labels exactly correct.
* **Precision, Recall, F1-score:** Computed per class.
* **Confusion Matrices:** Visualized per category.

## **4. Results**

| **Modality** | **Validation Hamming** | **Test Hamming** | **Validation Subset Accuracy** | **Test Subset Accuracy** |
| --- | --- | --- | --- | --- |
| Unimodal Image | 0.9658 | 0.9648 | 0.0140 | 0.0120 |
| Unimodal Text | 0.9740 | 0.9730 | 0.1463 | 0.1159 |
| Multimodal Fused | 0.9727 | 0.9721 | 0.1022 | 0.0899 |

* **Image-only:** High Hamming but very low exact match; edges fail to capture high-level semantics.  
  **Text-only:** Best exact-match accuracy; captions provide strong object cues.
* **Fused:** Slightly improves over image-only but does not surpass text-only.

## **5. Discussion**

**Strengths**

* Simple, interpretable, and reproducible pipeline.
* Demonstrates value of captions in multimodal tasks.
* Provides clear comparison between unimodal and fused setups.

**Limitations**

1. **Weak image features:** Canny edges discard texture, color, and object semantics.
2. **Caption limitations:** Not all objects appear in captions → missing labels.
3. **Fusion strategy:** Simple concatenation fails to model cross-modal interactions.
4. **Imbalance issue:** Many COCO categories are rare; Random Forests struggle without rebalancing.  
   **Evaluation bias:** Hamming score inflates performance due to dominance of negative labels.

**Future improvements**

* Replace edges with **CNN features** (e.g., ResNet, VGG).
* Use **pretrained embeddings** (BERT, CLIP) instead of Word2Vec.
* Apply **advanced fusion** (attention, gating, transformers).
* Handle imbalance (weighted loss, resampling).
* Evaluate with **mAP and macro-F1**, not only Hamming score.

## **6. Conclusion**

* **Captions (text features)** are more effective than edges (image features) for predicting object categories.
* **Fusion via concatenation** does not guarantee improvements over unimodal text performance.
* Classical approaches can provide interpretable baselines, but they fall short for complex datasets like COCO.

The work highlights both the potential and the limitations of multimodal pipelines using simple features. Future improvements should adopt deep embeddings and more advanced fusion mechanisms for meaningful gains.