Multimodal Machine Learning Assignment 1

2022UG000033 Nikita Agre

Multimodal Image Classification Using Traditional Feature Extraction

---

● Introduction

This report details a multimodal machine learning pipeline for image classification using the MS COCO 2017 validation set. The goal was to extract traditional features from images (Canny edges) and text (Word2Vec), fuse them, train a classifier, and compare multimodal performance against unimodal baselines.

● Dataset Description

The MS COCO 2017 Validation Set (5,000 images) was used.

- **Images:** Approximately 5,000 unique images.
- **Captions:** Each image is associated with at least 5 descriptive captions. (Identified 25014 captions in total)
- **Labels:** Object labels for images were extracted from the instance annotations.
- **Processed Data:** After linking images, captions, and labels, a total of 24,774 labelled samples were generated for training, validation, and testing. An additional 240 samples were identified as unlabelled and set aside for later prediction.
- **Classes:** 80 unique object classes.
- Highly **imbalanced** data

● Feature Extraction

Image Feature Extraction: Canny Edge Detection

It's a traditional method which captures the structural outlines of objects. It involves:

- **Noise Reduction:** Smoothing the image with a Gaussian filter to remove noise.
- **Gradient Calculation:** Finding the intensity gradients of the image.
- **Non-maximum Suppression:** Thinning the edges to retain only the most prominent ones.
- **Hysteresis Thresholding:** Applying a double threshold to distinguish strong edges from weak edges, and connecting weak edges to strong ones.

Sigle open-cv function "cv2.Canny" does this process.

Text Feature Extraction: Word2Vec Embeddings

Pre-processing Steps: To ensure meaningful embeddings, the raw text captions underwent the following steps:

- Lowercasing: All text was converted to lowercase.
- Punctuation Removal: Punctuation marks were removed.
- Tokenization: Sentences were broken down into individual words (tokens) using nltk.word_tokenize.

A pre-trained Word2Vec model 'word2vec-google-news-300' is used which provides 300-dimensional vectors for a large vocabulary of words. Since Word2Vec provides embeddings for individual words, and needed a single vector for an entire caption, the following aggregation strategy was used:

1. For each pre-processed caption, the Word2Vec embedding was retrieved for every word present in the model's vocabulary.
2. The average of all word embeddings in a caption was computed to create a single, fixed-size 300-dimensional vector representing that caption.

● Feature Fusion Strategy

Strategy used is **Concatenation.**

For each (image, caption) pair, the 4096-dimensional Canny edge feature vector and the 300-dimensional Word2Vec caption feature vector were simply concatenated (joined end-to-end). This resulted in a combined multimodal feature vector of 4396 dimensions for each sample. Concatenation is a straightforward and widely used early fusion technique. It creates a unified feature space, allowing the subsequent classification model to learn complex interactions between the image and text modalities directly from the combined representation.

Another strategy used is **elementwise addition** operation. Used zero-padding for the word2vec_caption_features so they match the canny_image_features dimension before the element-wise addition. This resulted in a fused feature vector of 4096 dimensions, maintaining the original dimensionality of the image features.

CSV file: Stored features (4396) and label of each image in csv file. And in another csv fused features (4096) and label is stored.

- **Image Classification Model**
  - **Task:** The classification task is **multi-label, multiclas**s classification, meaning an image can belong to multiple classes simultaneously.
  - **Dataset Split:** Data was split by unique images (**70% train, 10% validation, 20% test**) to **prevent information leakage**. There are 5 rows corresponding rows (captions) to one image. Splitting based on row could lead to captions for the same image appearing in both training and test sets, then evaluation would not make any sense.

Model: A simple Multilayer Perceptron (MLP) was employed.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 2048)	9,005,056
batch_normalization (BatchNormalization)	(None, 2048)	8,192
dropout (Dropout)	(None, 2048)	0
dense_1 (Dense)	(None, 1024)	2,098,176
batch_normalization_1 (BatchNormalization)	(None, 1024)	4,096
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 512)	524,800
batch_normalization_2 (BatchNormalization)	(None, 512)	2,048
dropout_2 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 256)	131,328
batch_normalization_3 (BatchNormalization)	(None, 256)	1,024
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 80)	20,560

Training Details:

- Optimizer: Adam optimizer with a learning rate of 1e-4.
- Loss Function: binary_crossentropy, which is standard for multi-label classification problems.
- Metrics: Precision, Recall, AUC (Area Under the Curve, specified for multi-label), and Binary Accuracy were tracked to evaluate performance.
- Batch Size: 64.
- Epochs: 30.

Trained four different models**: 2 multimodal models** with concatenation and addition fusion technique and **2 unimodal models** (image only and caption only model)

Evaluated all four models by predicting labels for hold-out data which didn't have original labels and then printed image-caption with predicted labels like following:

Predicted: sink,toilet
Caption: "A bathroom being renovated with pipes in the wall"

Predicted: sink
Caption: "A church with a stain glassed window of the virgin mary."

Predicted: car,stop sign
Caption: "A stop sign on a post at a public street."

- **Evaluation and Performance Comparison**

Fusion Technique	**Model Type**	**Test Loss**	**Test Precision**	**Test Recall**	**Test AUC**	**Test Binary Accuracy**	**Classification Report**
Concatenation	Multimodal (Canny + W2V)	0.1629	0.5110	0.1331	0.5884	0.9618	**extremely poor performance for most classes**, with many having precision, recall, and F1-scores of 0.00. It largely fails to identify most objects. It only reliably identifies 'person'. The low AUC confirms that its ability to discriminate between the presence and absence of labels is only slightly better than random guessing
Elementwise addition with zero padding	Multimodal (Canny + W2V fused)	0.1624	0.5376	0.1211	0.5818	0.9629	Almost zero values of precision, recall, f-1 score for all classes.
-	Unimodal (Canny Image Only)	0.1611	0.4953	0.1224	0.5889	0.9622	Has 0.00 precision, recall, and f1-score for nearly all classes. Shows poorer discriminatory power. It only shows non-zero performance for few classes ('person', 'airplane', 'bus' etc)
-	**Unimodal (Word2Vec Caption Only)**	**0.0680**	**0.7986**	**0.5354**	**0.9446**	**0.9773**	**The AUC is exceptionally high, indicating excellent discriminatory power. Many other classes now show strong performance. So, this model is extremely effective for object classification.**

- **Confusion matrix per category** is also printed.
- **Inflated accuracy:** Since most image-caption pairs do not contain a specific object, predicting "negative" for most classes is leading to a high number of True Negatives (TN), artificially inflating accuracy. So, accuracy is misleading metric.
- Unimodal **(Word2Vec Caption Only):** Performed well, it strongly suggests that for the given task and chosen features, the textual descriptions provide much more discriminative information for object classification than the Canny edge features. The semantic content of captions is highly correlated with the object labels
- Unimodal **(Canny Image Only)**: this model largely failed to predict specific objects from edge information alone.
- The **multimodal model using concatenation** (Test AUC 0.5884) shows performance **on par with, or negligibly worse than, the unimodal Canny image-only model (Test AUC 0.5889)**. This is a critical finding, indicating that simply concatenating a powerful text feature with a weak image feature does not necessarily lead to improved performance over the image-only baseline, and can even introduce noise.
- This fusion strategy yielded the **lowest performance (Test AUC 0.5818)** among all models tested. This suggests that forcing a direct element-wise combination with zero-padding when modalities are so disparate and image features are so weak may actively hinder the learning process, possibly by creating a noisy combined representation that obscures any useful signal.
- Out of all models **caption only modal** with text features **outperformed**.

● Analysis and Discussion

Strengths:

- o Implemented 2 multimodal models comparing two different early fusion techniques.
- o Word2Vec proved highly effective in capturing the semantic meaning of captions as model with Word2Vec features performed well compare to all other models.
- o Addition fusion technique keeps the same dimensions whereas concatenation increases the dimensionality.
- o Concatenation keeps both modalities intact whereas addition of features tends to collapse modality specific info.

Limitations

- o Canny features were not semantically rich enough for robust object classification. As a result, the image-only model performed poorly, and the multimodal model's performance was significantly constrained by this weak visual input, unable to fully leverage the power of the combined approach. It completely disregards crucial visual cues like texture, color, and higher-level object parts.
- o Element-wise addition doesn't explicitly preserve which part of the signal came from which feature set and also it degraded the performance
- o Simple MLP model is not able to learn complex non-linear interactions.

● Improvement

- o Using non-linear fusion techniques can help to capture the cross-modal interaction learning.
- o Using advanced image feature extraction technique would improve the result. Like using pre-trained CNN model to extract image features would improve the result.
- o Using advanced model architectures
- o Addressing class imbalance.

● Conclusion

Implemented multimodal image classification pipeline, demonstrating **two early fusion strategies: concatenation and element-wise addition**. While the text modality, using Word2Vec, proved exceptionally effective, the chosen Canny edge features for images were a critical and limiting factor. The performance of both multimodal models, despite incorporating strong text features, was ultimately constrained by the weak image features, failing to surpass the

strong text-only baseline and, in the case of element-wise fusion, even underperforming the unimodal image model. This underscores that the quality and relevance of individual modality features are paramount, and effective fusion requires not only a suitable strategy but also strong, semantically rich inputs from all modalities to unlock the full potential of multimodal learning.