LONDON METROPOLITAN UNIVERSITY

Islington college
(इस्लिङ्टन कलेज)

Module Code & Module Title
Level 7 – CC7009NI Neural Networks & Deep Learning

Assessment Type
60% Individual Coursework
Semester
2025 Autumn
Credit: 20 Semester Long Module

Student Name: Avinna Maharjan

London Met ID: 20048839

College ID: NP01MS7A240005

Assignment Due Date: Sunday, December 21, 2025

Assignment Submission Date: Sunday, December 21, 2025

Submitted To: Anish Chapagain

Word Count (Where Required): 563

I confirm that I understand my coursework needs to be submitted online via MySecondTeacher classroom under the relevant module page before the deadline in order for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.

Table of Contents

Table of Figures

1 Introduction

Digital advertisements increasingly rely on a combination of visual elements (images) and short textual slogans to influence consumer perception, make decision and communicate persuasive messages. Understanding advertisement content automatically is valuable for applications such as targeted marketing, content moderation, brand analytics, and ethical ad monitoring. Traditional machine learning approaches often process either text or images independently, which limits their ability to capture the complementary information present across modalities. Recent advances in multimodal deep learning show that jointly modelling heterogeneous data sources leads to more robust representations and improved classification performance (Baltrušaitis, Ahuja and Morency, 2019).

This project proposes a multimodal deep learning system that jointly analyses advertisement images and their associated slogans to classify ads into predefined categories. By integrating visual and textual features, the system aims to demonstrate improved performance and a deeper representation of advertisement semantics, aligning with modern multimodal learning principles in deep learning.

2 Problem Statement

Single-modality approaches often overlook contextual cues present across modalities, leading to reduced classification accuracy (Ngiam *et al.*, 2011). The objective of this project is to design and implement a multimodal deep learning model that classifies advertisements using both:

- Images (visual modality), and
- Slogans or short text descriptions (text modality).

The challenge lies in effectively combining heterogeneous data types while maintaining computational efficiency and ethical AI considerations.

2.1 Real World Applications

Multimodal advertisement classification has several real-world applications, including:

a) **Digital Marketing Analytics:** Automatic categorization of advertisements for campaign optimization.
b) **Ad Content Moderation:** Identifying sensitive or misleading advertisements.
c) **Recommendation Systems:** Improving ad targeting by understanding multimodal intent.
d) **Ethical AI Auditing:** Monitoring biased or harmful messaging in advertisements.

3 Chosen Modality and Dataset

3.1 Modality Selection

This project follows a multimodal learning approach, integrating:

a) **Image modality:** Visual content of advertisements.

b) **Text modality:** Associated slogans or short captions.

3.2 Dataset Description

The MAdVerse dataset is used for training and evaluation.

Source: Zenodo MAdVerse

Content: The dataset contains advertisement images paired with slogans and category labels (Sagar *et al.*, 2024).

Advantages:

a) Publicly available and well-documented

b) Designed specifically for multimodal advertisement understanding

c) Suitable size for training on limited hardware using transfer learning

4 Exploratory Data Analysis

Dataset Overview

The dataset includes 52,443 advertisement images sourced from google images, social media platforms and digital newspapers (Sagar *et al.*, 2024). Each advertisement sample includes an image and annotations stored in structured JSON files, which contain hierarchical category labels covering 11 primary categories, 51 sub-categories, and 524 fine-grained brand labels. Additionally, the dataset spans over 11 languages, with English, Hindi, and Marathi.

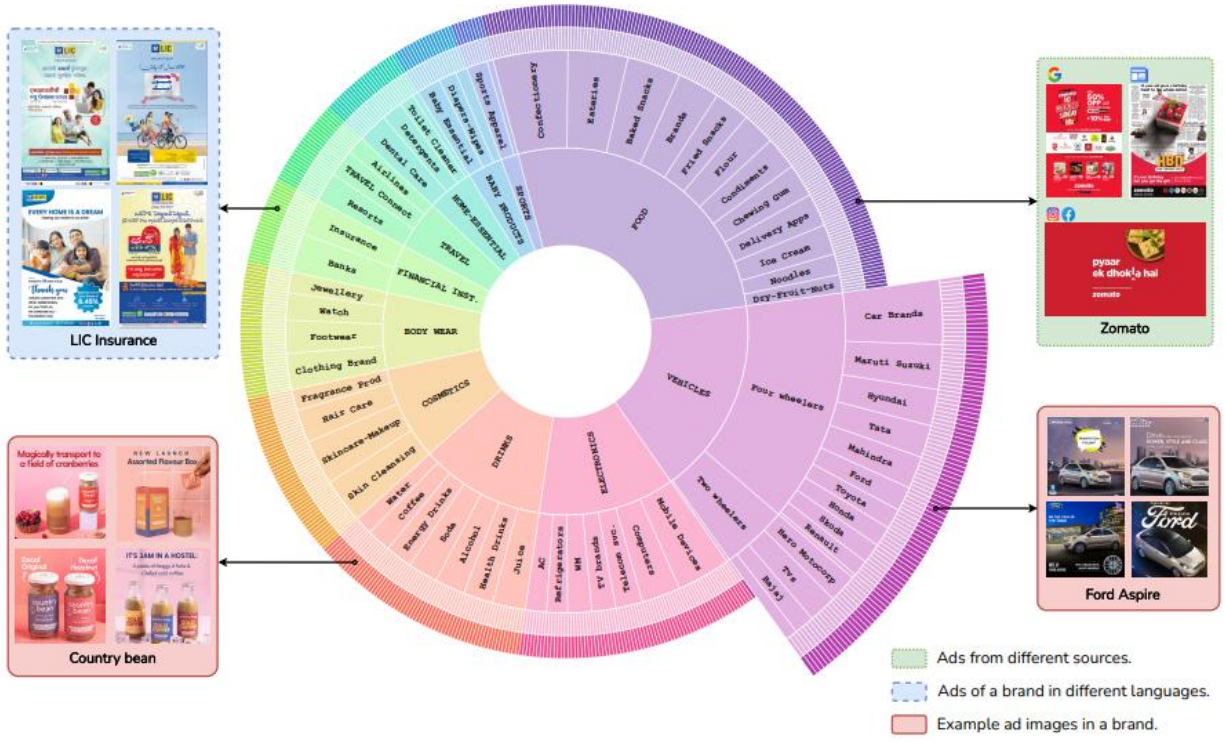

Figure 1: Hierarchical Taxonomy of the Dataset (Sagar *et al.*, 2024)

5 Model Architecture and Pre-trained Models

a) Image Encoder

A pre-trained convolutional neural network (CNN) such as MobileNetV2 will be used for image feature extraction. MobileNetV2 is pre-trained on ImageNet, which allows the model to leverage learned visual representations and reduces training time while still capturing relevant advertisement features and is computationally efficient and suitable for deployment on limited GPU resources (Howard *et al.*, 2017).

b) Text Encoder

For textual representation, DistilBERT, a lightweight transformer-based language model, will be used (Adoma, Henry and Chen, 2020). Transformers have demonstrated strong performance in natural language understanding tasks due to their self-attention mechanisms (Vaswani *et al.*, 2017). Its pre-trained weights allow the model to capture contextual meaning from short text inputs, such as slogans, which improves representation without requiring extensive training.

6 References

Adoma, A.F., Henry, N.-M. and Chen, W. (2020) 'Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition', in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, China: IEEE, pp. 117–121. Available at: https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379.

Baltrušaitis, T., Ahuja, C. and Morency, L.-P. (2019) 'Multimodal Machine Learning: A Survey and Taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp. 423–443. Available at: https://doi.org/10.1109/TPAMI.2018.2798607.

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017) 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications'. Available at: https://doi.org/10.48550/arXiv.1704.04861.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A. (2011) 'Multimodal Deep Learning', in. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, Washington, USA, pp. 689–696.

Sagar, A., Srivastava, R., R T, R., Kesav Venna, V. and Sarvadevabhatla, R.K. (2024) 'MAdVerse: A Hierarchical Dataset of Multi-Lingual Ads from Diverse Sources and Categories', in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp. 8072–8081. Available at: https://doi.org/10.1109/WACV57701.2024.00790.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. ukasz and Polosukhin, I. (2017) 'Attention is All you Need', in *Advances in Neural Information Processing Systems*. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Curran Associates, Inc. Available at: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (Accessed: 21 December 2025).