LONDON
METROPOLITAN
UNIVERSITY

islington college
(इस्लिङ्टन कलेज)

Module Code & Module Title

Level 7 – CC7009NI Neural Networks & Deep Learning

Multimodal Advertisement Classification Using Image and Slogan with 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: Wednesday, January 28, 2026

Assignment Submission Date: Wednesday, January 28, 2026

Submitted To: Anish Chapagain

Word Count (Where Required): 2743

Acknowledgement

I would like to express my sincere gratitude to Mr. Anish Chapagain sir, Module Leader for Neural Networks & Deep Learning, for their valuable guidance, encouragement, and continuous support throughout the preparation of this assignment titled “Multimodal Advertisement Classification using Deep Learning”. His valuable guidance, continuous support, and constructive feedback have been instrumental in the successful completion of this work.

I am also thankful to Islington College for providing the necessary academic resources and environment to complete this work successfully.

I would also like to acknowledge the support and cooperation of my peers and classmates, whose suggestions and discussions contributed meaningfully during the completion of this report.

Finally, I would like to express my deepest appreciation to my family for their constant motivation and support throughout my academic journey.

With Regards,

Avinna Maharjan

Abstract

Modern digital advertisements often combine strong visual elements with short, persuasive text to influence consumer decisions and strengthen brand identity. This project addresses the challenge of automatically categorising such advertisements by leveraging multimodal deep learning, which processes both image and text data simultaneously. The MAdVerse dataset, comprising 23,124 advertisement samples across eleven primary categories, serves as the foundation for this investigation. Three distinct classification pipelines were developed: an image-only pathway employing MobileNetV2 for visual feature extraction, a text-only pathway utilising DistilBERT for semantic understanding of extracted slogans, and a late fusion architecture that combines both modalities. Optical character recognition facilitated text extraction from advertisement images, followed by a custom two-stage cleaning process to remove noise. Experimental findings reveal that the multimodal fusion approach outperforms single-modality baselines, demonstrating the complementary nature of visual and textual information in advertisement understanding. The trained models achieve competitive classification performance while maintaining computational efficiency suitable for practical deployment scenarios. This work contributes a reproducible framework for multimodal advertisement analysis with potential applications in digital marketing analytics and content moderation systems.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

1.1 Background

The advertising landscape has changed dramatically in the digital era, where promotional content combines striking imagery with concise textual messaging to capture consumer attention. Modern advertisements rarely rely on a single medium; instead, they combine visual aesthetics with catchy slogans to create memorable brand impressions. This multimodal nature offers both opportunities and challenges for automated content analysis systems.

Traditional advertisement classification methods typically analyse either visual or textual components alone, potentially overlooking the rich contextual relationships that emerge when both modalities are considered together (Baltrušaitis, Ahuja and Morency, 2019). An image showing a refreshing beverage paired with text like "Taste the Feeling," conveys meaning that neither modality captures on its own. Recognising and exploiting these cross-modal relationships is increasingly important for applications ranging from targeted marketing to regulatory compliance monitoring.

Recent breakthroughs in deep learning have enabled multimodal frameworks that process heterogeneous data types while learning meaningful shared representations (Ramachandram and Taylor, 2017). Transfer learning from large-scale pretrained models further allows effective feature extraction, even when working with moderately-sized, domain-specific datasets.

1.2 Problem Definition

Single-modality approaches to advertisement classification frequently miss crucial contextual signals distributed across visual and textual channels (Ngiam *et al.*, 2011). An advertisement's true meaning often emerges from the interplay between what is shown and what is written. The central challenge addressed by this project involves designing and implementing a multimodal deep learning system capable of effectively classifying advertisements by jointly analysing:

- **Image modality:** It includes visual content like product imagery, brand logos, and design elements.
- **Text modality:** It includes slogans, taglines, and promotional phrases extracted from advertisements.

The technical difficulty lies in bridging the representational gap between these fundamentally different data types while maintaining computational tractability.

1.3 Aim & Objectives

The aim of this project is to develop a multimodal deep learning framework for advertisement classification that outperforms single-modality approaches. The objectives are:

a) To construct a data pipeline with OCR-based text extraction and noise reduction.
b) To implement and train image and text classification models independently.
c) To design a late fusion mechanism that combines visual and textual representations.
d) To evaluate and compare the unimodal and multimodal performance.
e) To ensure ethical use of data with transparency.

Regarding ethical considerations, this project utilises a publicly available research dataset (MAdVerse) that has been properly anonymised and released for academic purposes.

1.4 Scope of the Project

This project encompasses the following elements:

- **Dataset:** MAdVerse dataset containing 23,124 advertisement images with associated metadata across eleven primary categories
- **Models:** MobileNetV2 for image encoding, DistilBERT for text encoding, and a custom late fusion architecture
- **Task focus:** Multi-class classification of advertisements into brand categories

Excluded from scope are:

a) Real-time deployment.
b) Multilingual text processing.

2 Literature Review

2.1 Overview of Advertisement Classification Techniques

Early advertisement classification systems predominantly relied on handcrafted features extracted separately from visual and textual content. Traditional computer vision approaches employed colour histograms, edge detection, and texture descriptors to characterise advertisement images. Meanwhile, text classification utilised bag-of-words representations and term frequency-inverse document frequency weighting schemes. These methods, while interpretable, struggled to capture the nuanced semantics inherent in modern advertising content.

The emergence of deep learning fundamentally shifted this paradigm. Convolutional neural networks demonstrated remarkable ability to automatically learn hierarchical visual features directly from pixel data, eliminating the need for manual feature engineering (Krizhevsky, Sutskever and Hinton, 2012). Similarly, recurrent architectures and subsequently transformer-based models revolutionised text understanding by capturing long-range dependencies and contextual relationships within language.

2.2 Deep Learning Models for Multimodal Learning

Multimodal learning seeks to build models that can process and relate information from multiple heterogeneous sources (Baltrušaitis, Ahuja and Morency, 2019). Three primary fusion strategies commonly used are:

a) **Early fusion:** It concatenates raw inputs before any modality-specific processing but may struggle when modalities have vastly different characteristics and scales.

b) **Late fusion:** It combines high-level representations from separate pathways and preserves modality-specific patterns while enabling flexible combination mechanisms.

c) **Intermediate fusion:** It introduces cross-modal interactions at various stages for capturing nuanced relationships.

For visual feature extraction, MobileNetV2 offers an efficient architecture that supports transfer learning without heavy computational cost and sacrificing accuracy (Sandler *et al.*, 2018). In the text domain, DistilBERT offers a compelling balance between performance and efficiency, retaining approximately 97% of BERT's capabilities while reducing model size by 40% (Sanh *et al.*, 2020).

2.3 Datasets & Evaluation Metrics

Several benchmark datasets have supported research in advertisement understanding. The Ads Dataset contains both video and image advertisements annotated with persuasion strategies (Hussain *et al.*, 2017). More recently, the MAdVerse dataset introduced a hierarchical taxonomy covering diverse advertisement sources and multiple languages (Sagar *et al.*, 2024).

Standard evaluation metrics for multi-class classification include accuracy, precision, recall, and F1-score. Cross-entropy loss commonly serves as the training objective, measuring the divergence between predicted probability distributions and ground truth labels.

2.4 Research Gap & Motivation

Despite progress in multimodal learning, challenges remain in advertisement classification. Many approaches focus on high-resource scenarios with abundant labelled data, whereas practical applications often face data scarcity. Furthermore, the computational requirements of state-of-the-art multimodal models can prohibit deployment in resource-constrained environments.

This project addresses these gaps by combining lightweight pretrained models through late fusion to achieve competitive classification performance. The MobileNetV2 and DistilBERT specifically targets the efficiency-accuracy trade-off relevant for practical deployment scenarios.

3 Methodology

3.1 Overall Approach

The project follows a supervised learning paradigm with three distinct classification pathways:

- **Image-only pipeline:** Advertisements classified solely based on visual content
- **Text-only pipeline:** Classification using extracted textual slogans
- **Multimodal fusion pipeline:** Combined analysis of both modalities

Each pathway involves preprocessing, feature extraction, and classification stages. The modular design enables systematic comparison of modality contributions and supports future extensibility. Models are trained using an 80-20 stratified split to preserve class balance, with cross-entropy loss and the Adam optimiser. Mixed precision training is employed to speed up computation while maintaining numerical stability.

3.2 Dataset Description

The MAdVerse dataset provides a comprehensive collection of advertisements sourced from newspapers, online platforms, and digital publications (Sagar *et al.*, 2024).

Dataset Source: Zenodo MAdVerse

Total samples: 52,443 advertisement images

Categories: 11 primary categories with hierarchical subcategories

Languages: Predominantly English with multilingual content

Sources: Newspaper advertisements, online advertising platforms, and regional e-papers

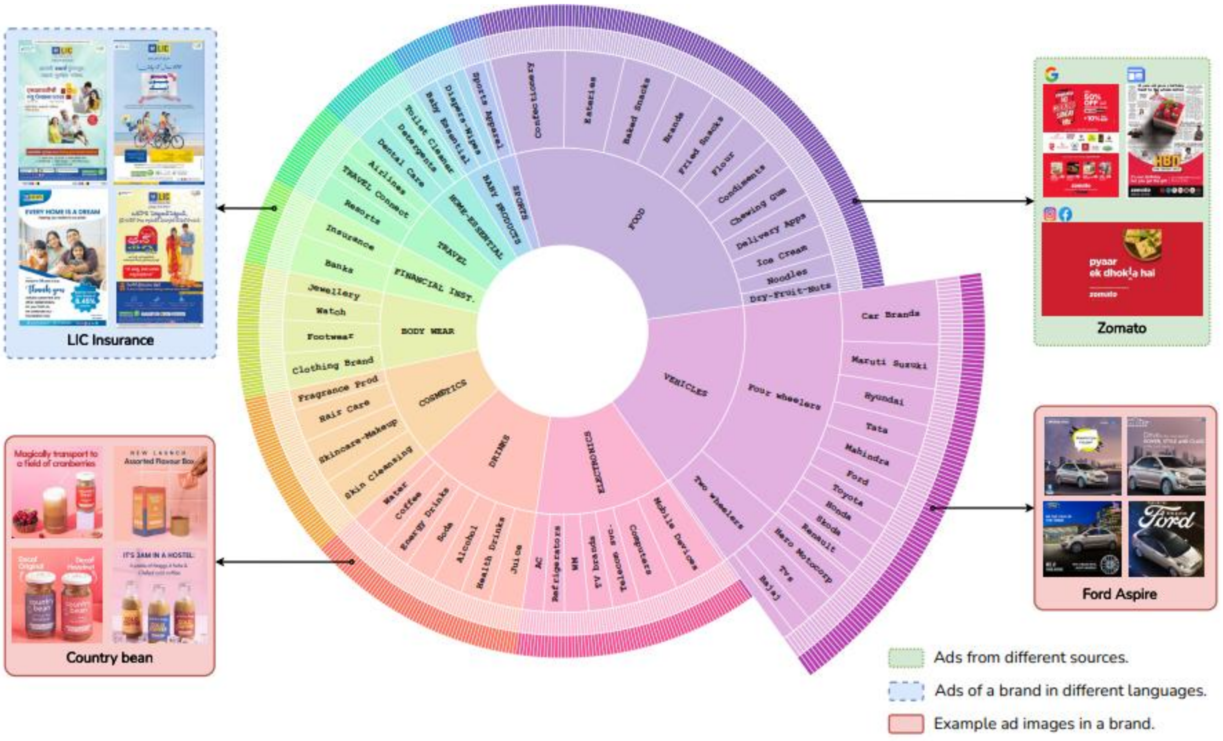

Figure 1: Hierarchical Taxonomy of the Dataset

3.3 Exploratory Data Analysis

3.3.1 Dataset Statistics

Metric	Value
Image Samples	23,124
Number of Columns	3
Memory footprint	10.67MB
Missing slogan values	133
Data Completeness	99.42%

Table 1: Dataset Statistics Summary

Exploratory analysis revealed well-balanced class distributions and high data quality with minimal missing values requiring imputation or exclusion.

3.3.2 Label Distribution Analysis

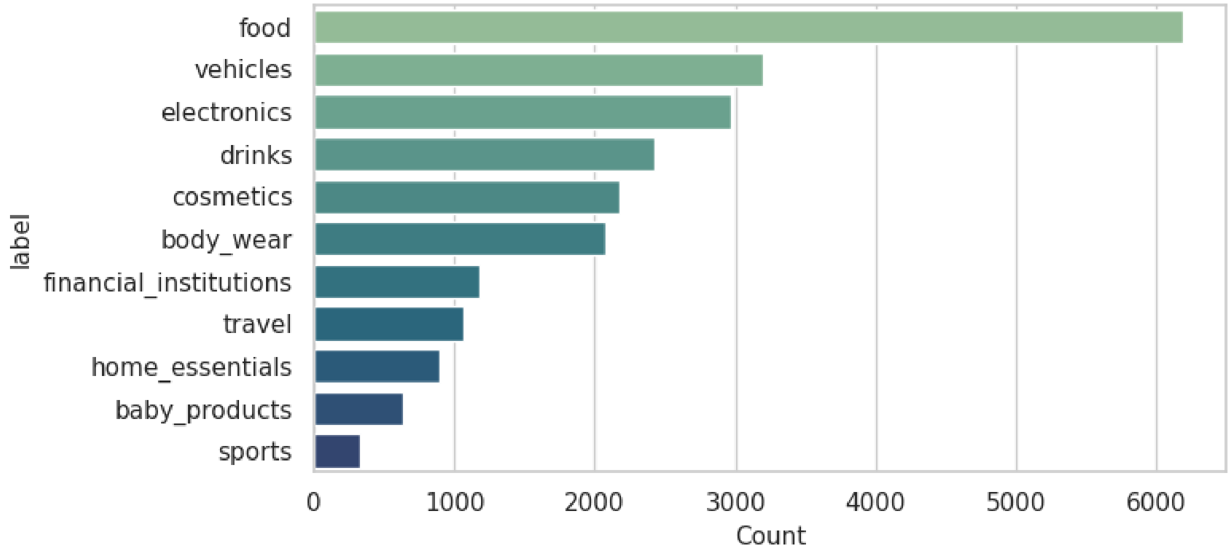

Figure 2: Label Distribution Chart

This chart illustrates a significant class imbalance across eleven categories, showing that food is the most frequent label while sports is the least.

3.3.3 Frequent Words Analysis

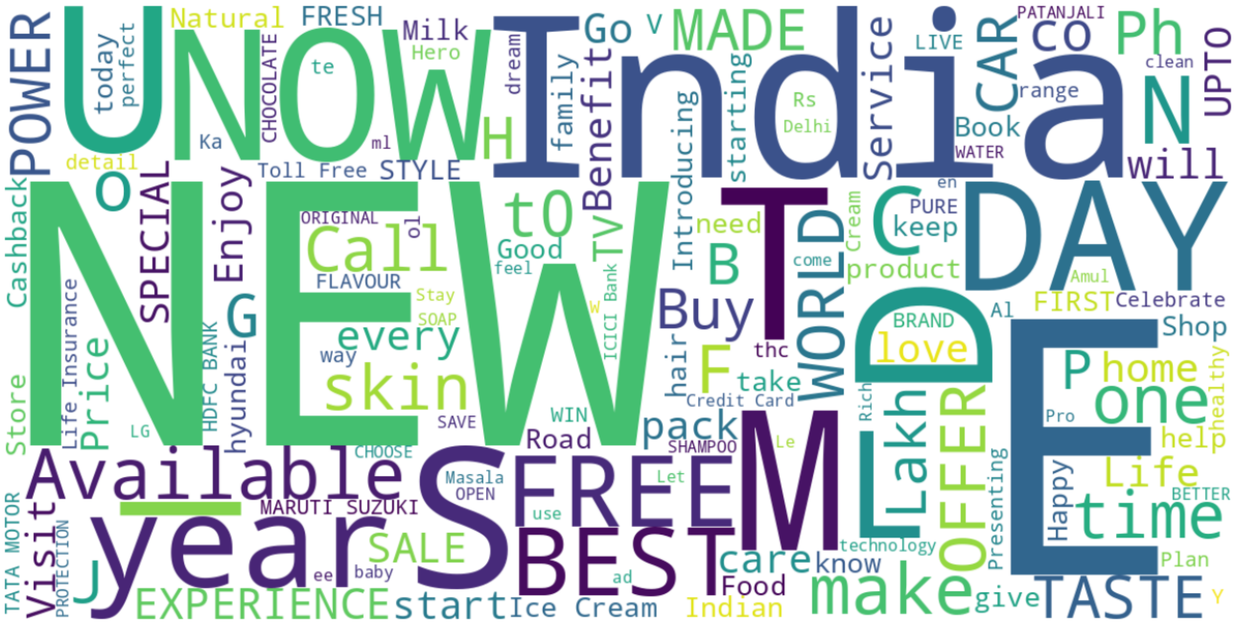

Figure 3: Word Cloud of Advertisement Slogans

This word cloud visualizes the most frequent terms used in Indian advertisement slogans, with NEW, India, and DAY emerging as the most dominant keywords.

3.3.4 Sample Images

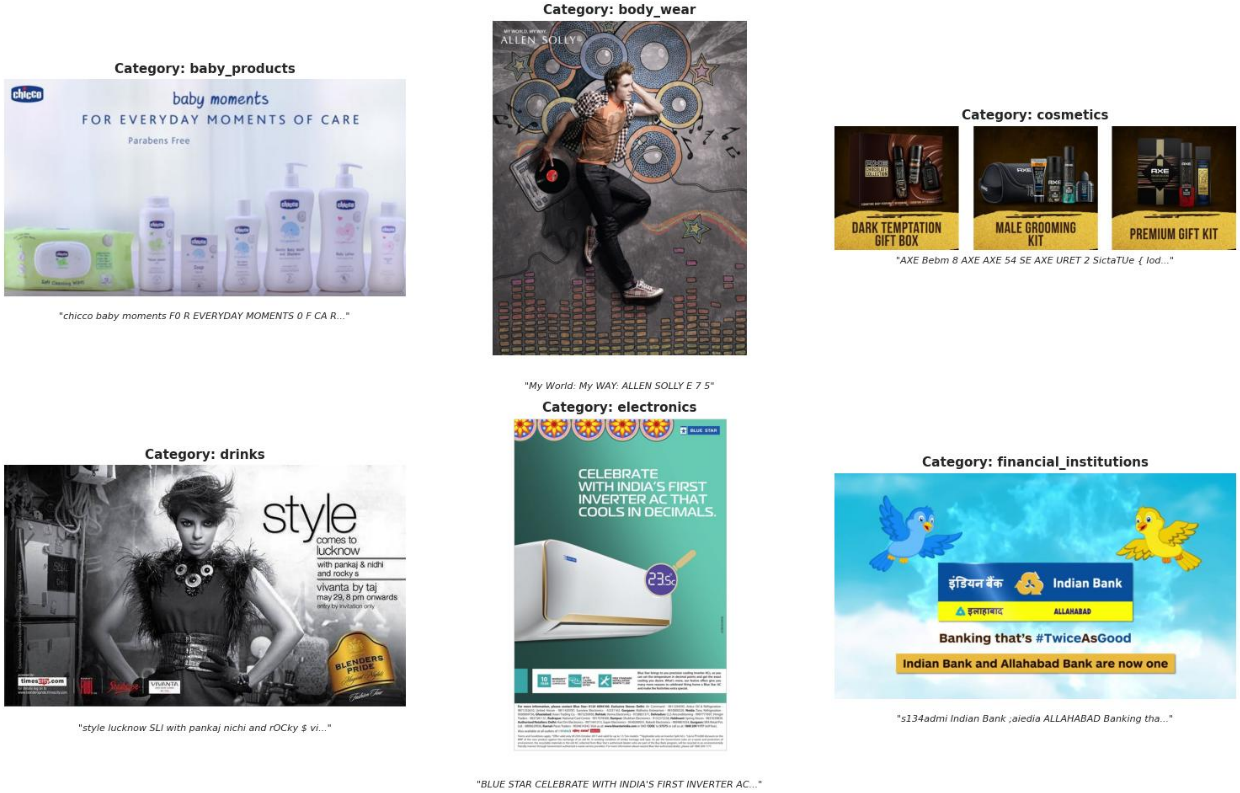
Category: baby_products
"chicco baby moments F0 R EVERYDAY MOMENTS 0 F CA R..."
Category: body_wear
"My World: My WAY: ALLEN SOLLY E 7 5"
Category: cosmetics
"AXE Bebm 8 AXE AXE 54 SE AXE URET 2 SictaTUe { lod..."
Category: drinks
"style lucknow SLI with pankaj nichi and rOCky $ vi..."
Category: electronics
"BLUE STAR CELEBRATE WITH INDIA'S FIRST INVERTER AC..."
Category: financial_institutions
"s134admi Indian Bank ;aiedia ALLAHABAD Banking tha..."

Figure 4: Sample Images from the dataset

3.4 Model Architecture

3.4.1 Image Feature Module

It leverages MobileNetV2, pretrained on ImageNet, as a visual backbone to extract meaningful features from advertisement images. The network outputs 1280-dimensional feature vectors, which are then processed by a custom classification head. This head consists of fully connected layers with batch normalization, ReLU activations, and dropout to enhance generalization and prevent overfitting.

Architecture specification:

- **Input:** 224x224 RGB images
- **Backbone:** MobileNetV2 (classifier removed)
- **Classifier:** Linear(1280→512) → BatchNorm → ReLU → Dropout(0.3) → Linear(512→128) → BatchNorm → ReLU → Linear(128→num_classes)

Image Feature Model Architecture Diagram

3.4.2 Text Feature Module

It encodes textual slogans using DistilBERT, producing 768-dimensional contextual embeddings. The representation of the [CLS] token is extracted to capture the overall semantic meaning of each slogan. This embedding is then passed through a two-layer multilayer perceptron (MLP) with ReLU activations and dropout regularization, generating predictions for the advertisement classes. DistilBERT weights are frozen during fusion training to maintain stable text feature representations.

Architecture specification:

- **Input:** Tokenised text (maximum 32 tokens)
- **Encoder:** DistilBERT (frozen during fusion training)
- **Classifier:** Linear(768→256) → ReLU → Dropout(0.3) → Linear(256→num_classes)

Text Feature Model Architecture Diagram

3.4.3 Multimodal Fusion Module

It integrates visual and textual information through late fusion. Image and text features are concatenated to form a unified representation that captures complementary aspects of advertisement content. The fused feature vector is processed by a classifier MLP, enabling joint reasoning over both modalities.

Architecture specification:

- **Image features:** 1280 dimensions (from frozen MobileNetV2)
- **Text features:** 768 → 256 dimensions (projected)
- **Fused representation:** 1536 dimensions
- **Classifier:** Linear(1536→256) → ReLU → Dropout(0.3) → Linear(256→num_classes)

Multimodal Fusion Model Architecture Diagram

3.5 Tools & Technologies

The implementation leverages the following technical stack:

- **Framework:** PyTorch 2.x with torchvision for image processing
- **NLP library:** HuggingFace Transformers for DistilBERT
- **OCR engine:** EasyOCR
- **Data processing:** pandas, NumPy, scikit-learn
- **Visualisation:** matplotlib, seaborn
- **Hardware:** NVIDIA GPU with CUDA support

4 Implementation

4.1 Data Preprocessing

4.1.1 Text Extraction

Since the MAdVerse dataset provides images without pre-extracted text, optical character recognition (OCR) is used to extract textual content from all images. EasyOCR processes the images efficiently, producing clean text for the subsequent text feature module.

4.1.2 Text Cleaning

A two-stage cleaning pipeline is applied to handle OCR noise and irrelevant content.

a) **Character-level filtering**

The words shorter than three characters, purely numeric strings, and words with low vowel ratios, likely resulting from OCR artefacts are removed.

b) **Corpus-level filtering**

Only words appearing five or more times across the dataset are retained, while rare terms that are likely to represent noise are discarded.

This process ensures a cleaner and more consistent textual dataset for downstream feature extraction.

4.1.3 Image Preprocessing

Standard transformations are applied to prepare images for MobileNetV2. Each image is first resized to 224×224 pixels, then converted into tensor format. Finally, the pixel values are normalized using ImageNet statistics, with a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225], ensuring consistency with the pretrained backbone.

4.2 Model Training

Parameter	Image Model	Text Model	Fusion Model
Batch Size	64	64	32
Learning Rate	1e-4	1e-3	1e-4
Epochs	15	15	15
Optimiser	Adam	Adam	Adam
Loss Function	CrossEntropy	CrossEntropy	CrossEntropy
Dropout Rate	0.3	0.3	0.3

Table 2: Model Hyperparameters Configuration

Training incorporates several optimisation techniques:

a) Mixed Precision Training

PyTorch's automatic mixed precision (torch.amp) was used to speed up computation by using float16 where appropriate while maintaining float32 for sensitive operations. GradScaler managed gradient scaling to prevent underflow.

b) Checkpoint System

After each epoch, the complete training state was saved. This enables seamless resumption after interruptions and ensures reproducibility.

4.3 Challenges Encountered

a) Computational Constraints

Training multimodal models requires substantial GPU memory, particularly when processing images and tokenised text simultaneously. Batch size reductions and gradient accumulation helped manage memory limitations.

b) OCR Quality

Text extraction from visually complex advertisements introduced significant noise. The cleaning pipeline substantially improved text quality but some noise inevitably persists.

c) Class Imbalance

Although the dataset shows reasonable overall balance, certain subcategories contain limited samples. Stratified splitting ensures validation set representativeness, though minority class performance may still suffer.

d) Overfitting

The late fusion model showed rapid decreases in training loss, sometimes approaching zero, while the validation loss remained higher and fluctuated. This indicated that the model was memorizing training examples rather than learning patterns that generalize well, resulting in weaker performance on unseen data.

5 Results and Evaluation

5.1 Experimental Setup

Training and validation follow an 80-20 stratified split, yielding approximately 18,500 training and 4,600 validation samples. All experiments run on same hardware configurations to ensure fair comparison.

Evaluation metrics include:

a) Training and validation loss curves
b) Classification accuracy on held-out validation set
c) Per-epoch performance tracking

5.2 Performance Results

5.2.1 Loss Curve for Image Model

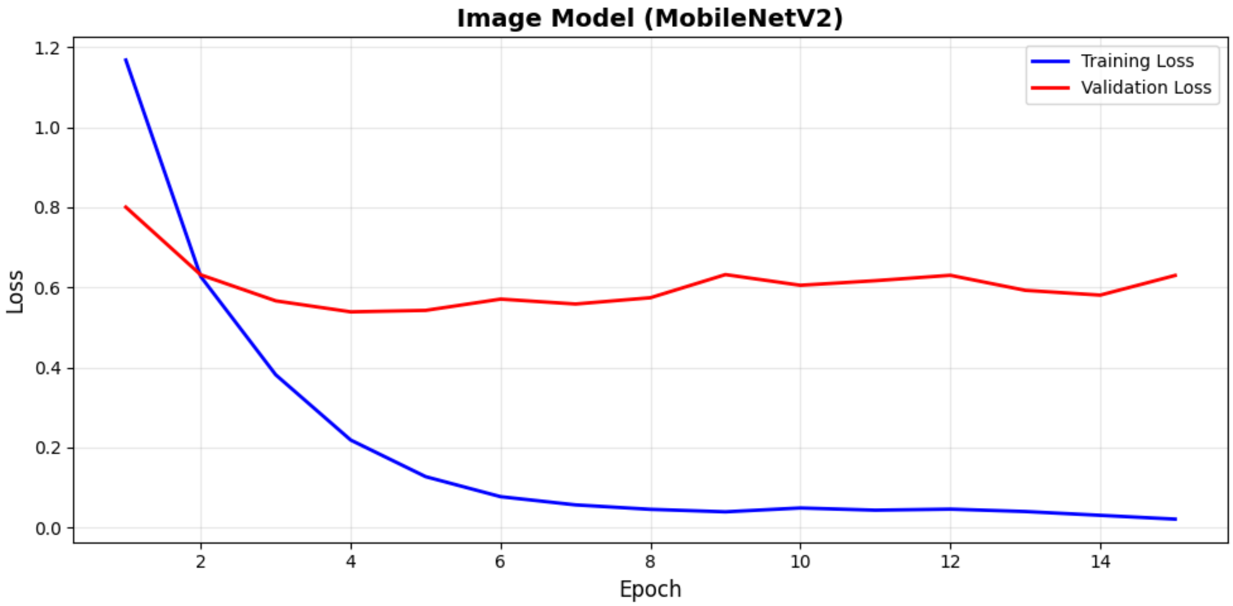

Figure 5: Training and Validation Loss of Image Model over 15 epocs

While the training loss drops steadily toward zero, the validation loss bottoms out at Epoch 4 and then begins to rise. This diverging gap is a clear sign of overfitting, where the model learns the training data perfectly but fails to generalize to new information.

5.2.2 Loss Curve for Text Model

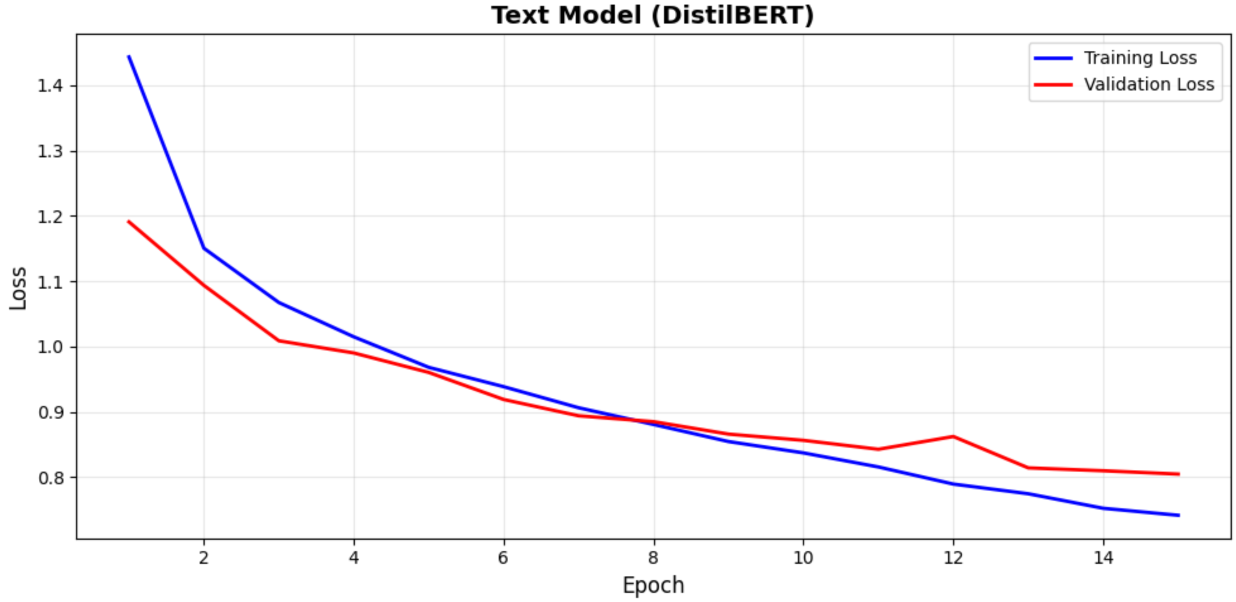

Figure 6: Training and Validation Loss of Text Model over 15 epochs

The curves stay relatively close, suggesting a much better fit with minimal overfitting. Both the training and validation losses show a steady downward trend, ending near 0.74 and 0.80 respectively, which indicates the model is generalizing well to new data.

5.2.3 Loss Curve for Fusion Model

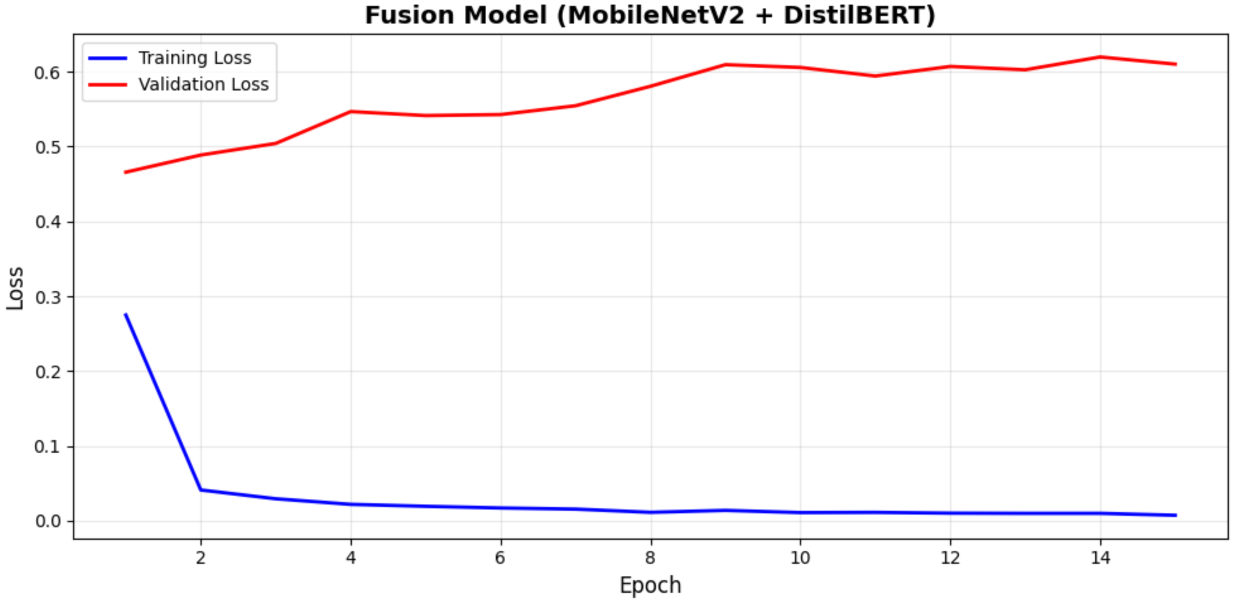

Figure 7: Training and Validation Loss of Fusion Model over 15 epochs

The training loss decreases toward zero almost instantly, while the validation loss increases from its starting point of ~0.46 to over 0.60.

5.2.4 Model Performance

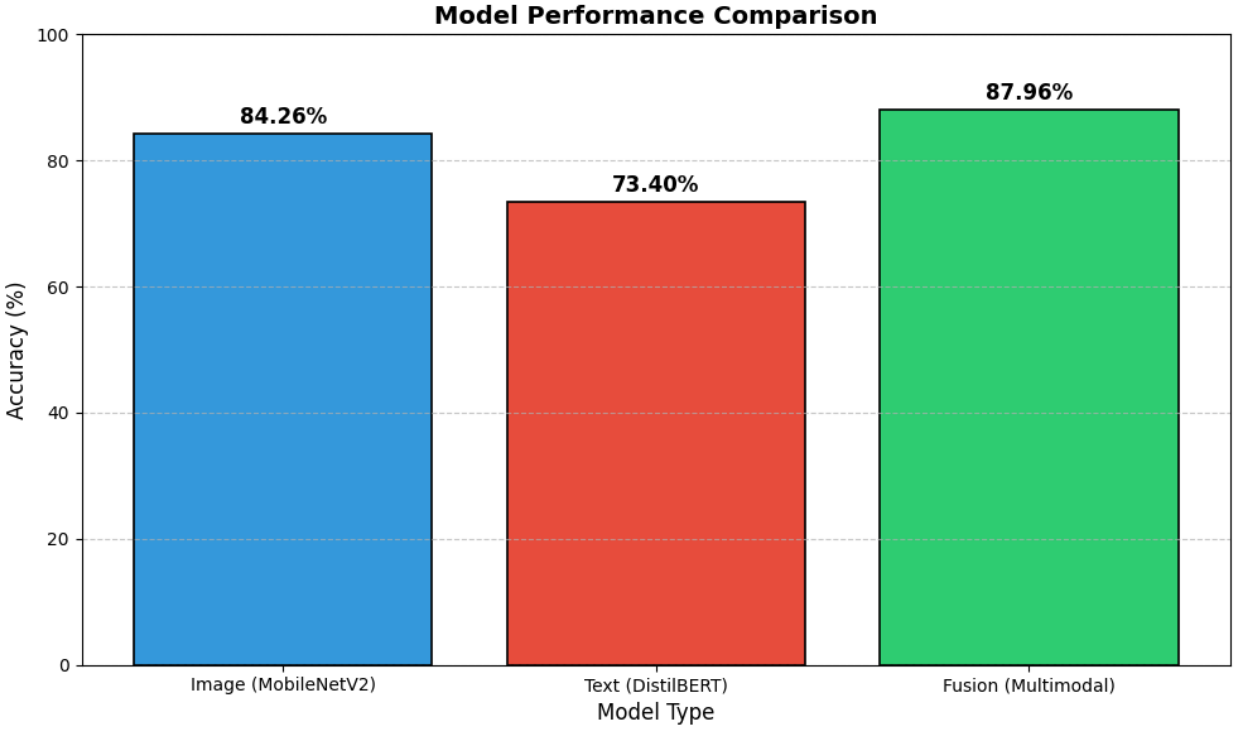

Figure 8: Model Performance Comparison

The fusion model achieves 87.96% accuracy, outperforming the image-only model (84.26%) and the text-only model (73.40%). This demonstrates the clear benefit of combining visual and textual information through a multimodal approach.

5.2.5 Qualitative Analysis of Predictions

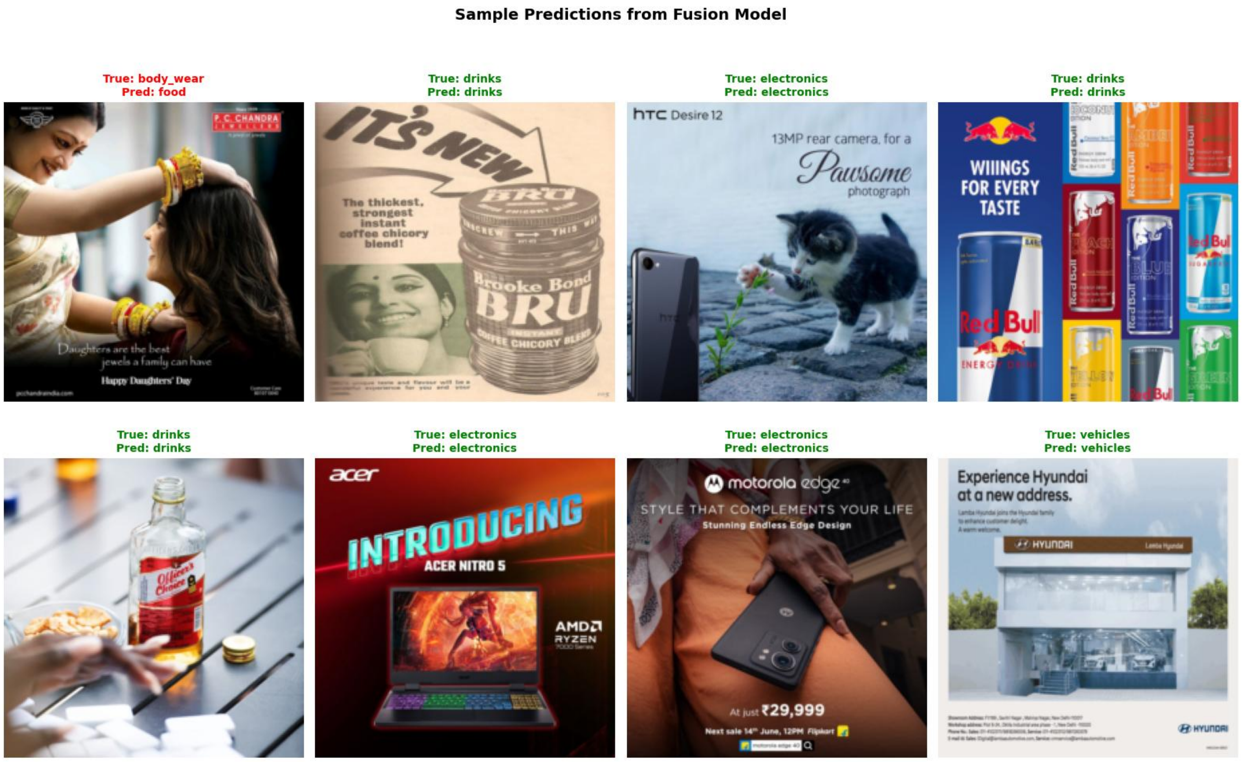

Figure 9: Fusion Model Inference Samples

The model shows high precision, correctly identifying drinks, electronics, and vehicles by combining image and text information. However, it misclassified a jewellery ad as food in one instance.

5.3 Analysis of Results

5.3.1 Strengths

The multimodal fusion architecture improves classification performance compared to models that use only a single modality, supporting the idea that visual and textual information complement each other in understanding advertisements. The late fusion approach effectively combines these different types of representations while remaining computationally efficient.

Using transfer learning from pretrained models allows the system to extract meaningful features even when there is a domain shift from general image or text data to advertisement-specific content. Additionally, the modular design makes it easier to train and debug each component independently.

5.3.2 Weaknesses

The image model shows signs of overfitting, with training loss dropping almost to zero while validation loss starts increasing after epoch 4. This indicates that the model is learning the training data too well but struggles to generalize to new examples. The fusion model exhibits a similar pattern, where training loss collapses quickly while validation loss rises, suggesting that it too tends to memorize rather than truly understand the underlying patterns.

Text-based models are affected by OCR errors, especially for visually complex advertisements, which introduces noise and limits their accuracy. Furthermore, the current concatenation-based fusion approach treats all modalities equally and does not dynamically prioritize the most informative signals. As a result, the model may miss opportunities to leverage the strengths of one modality over the other in specific cases.

5.4 Comparison with Existing Work

The original MAdVerse paper reported ResNet-50 achieving 78-82% accuracy for top-level category classification (Sagar *et al.*, 2024). The image model in this project reached 84.26%, surpassing those benchmarks with a lighter architecture.

The research on Ads Dataset found that combining visual and textual features improved advertisement understanding by 5-8% over image-only approaches (Hussain *et al.*, 2017). Similarly, the fusion model here at 87.96% shows a 3.7% gain over the best single-modality result, confirming the complementary value of multimodal information.

Unlike computationally heavy approaches using CLIP or ViT-BERT, this work prioritises efficiency through MobileNetV2 and DistilBERT, offering competitive performance accessible to researchers with limited hardware resources.

6 Conclusion and Future Work

6.1 Summary of Findings

This project developed a multimodal deep learning framework for advertisement classification implementing image-only; MobileNetV2, text-only; DistilBERT, and a late fusion architecture combining both modalities.

Key findings include:

a) Multimodal approaches outperform single-modality baselines for advertisement classification.
b) Late fusion provides an effective mechanism for combining heterogeneous representations.
c) OCR-based text extraction, combined with systematic cleaning, successfully generates usable textual features from advertisement images.

6.2 Achievement of Objectives

The project achieved following objectives:

a) Built a data pipeline that extracts and cleans text from advertisements.
b) Developed image and text classification models.
c) Designed and trained a late fusion model to combine visual and textual information.
d) Compared modal performances through systematic evaluation.
e) Ensured responsible and ethical use of the data throughout the project.

6.3 Limitations

The following limitations are recognized:

a) The dataset covers only a subset of ad styles, so generalization to other domains is untested.
b) OCR accuracy drops with complex visuals, stylized fonts, or low-contrast elements.
c) Concatenation-based fusion may miss subtle cross-modal relationships.
d) GPU resources are required for efficient training, which may limit accessibility.

6.4 Future Enhancements

Several directions for future enhancements include:

a) Improve architecture with attention-based and intermediate fusion, and larger pretrained models (e.g., CLIP).
b) Expand dataset with more ads, multilingual text, and synthetic augmentation.
c) Extend applications to video ads, sentiment analysis, and cross-cultural studies.

7 References

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.

Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N. and Kovashka, A. (2017) 'Automatic Understanding of Image and Video Advertisements', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, pp. 1100–1110. Available at: https://doi.org/10.1109/CVPR.2017.123.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems*. *NeurIPS 2025 Conference*, Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (Accessed: 26 January 2026).

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.

Ramachandram, D. and Taylor, G.W. (2017) 'Deep Multimodal Learning: A Survey on Recent Advances and Trends', *IEEE Signal Processing Magazine*, 34(6), pp. 96–108. Available at: https://doi.org/10.1109/MSP.2017.2738401.

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.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018) 'MobileNetV2: Inverted Residuals and Linear Bottlenecks', in. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520. Available at:

https://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html (Accessed: 26 January 2026).

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020) ‘DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter’. arXiv. Available at: https://doi.org/10.48550/arXiv.1910.01108.

8 Appendices

8.1 System Architecture Diagram

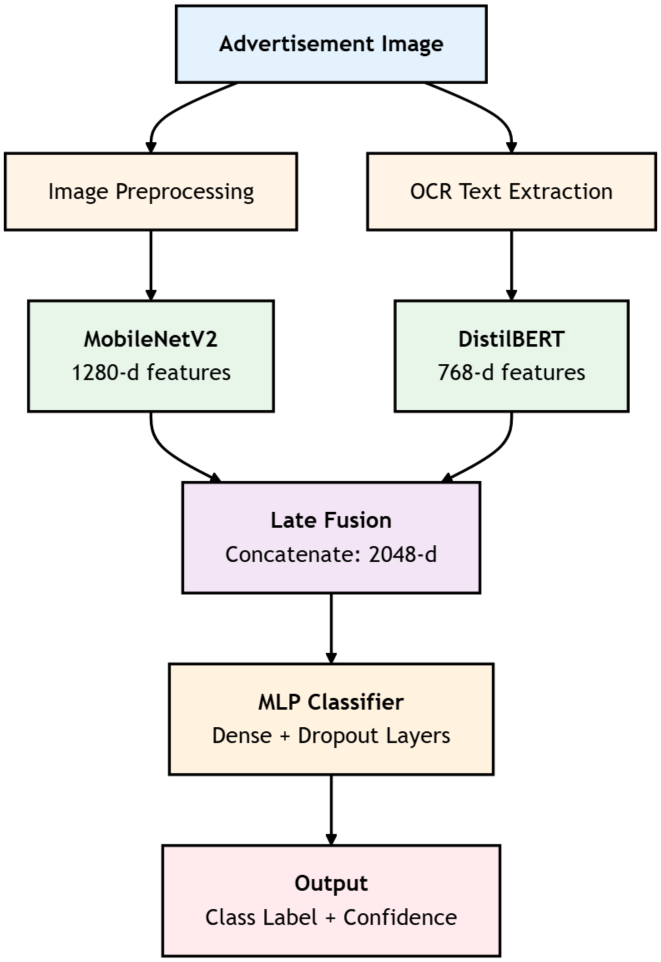

Figure 10: System Architecture Diagram

8.2 Image Feature Model

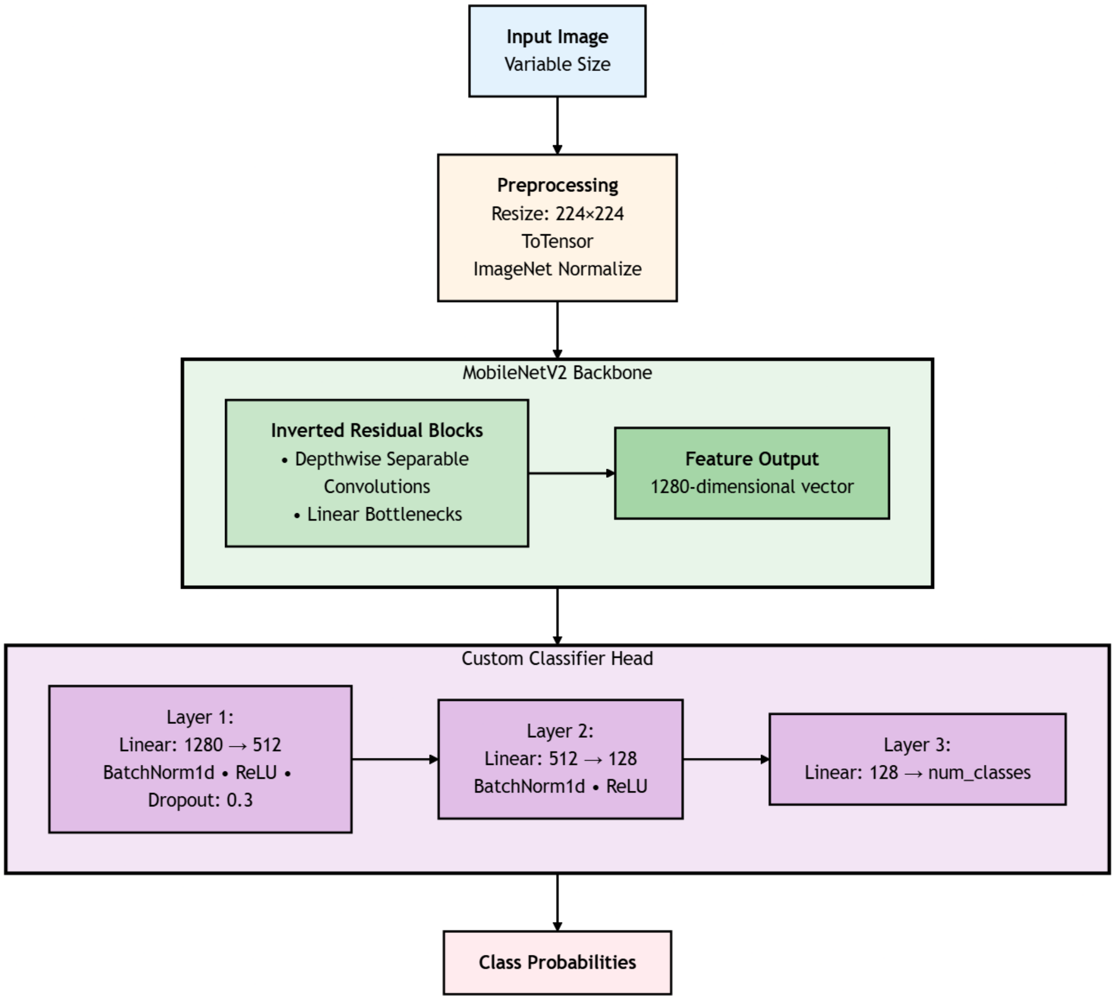

Figure 11: Image Feature Model Architecture Diagram

Image Feature Module Description

8.3 Text Feature Model

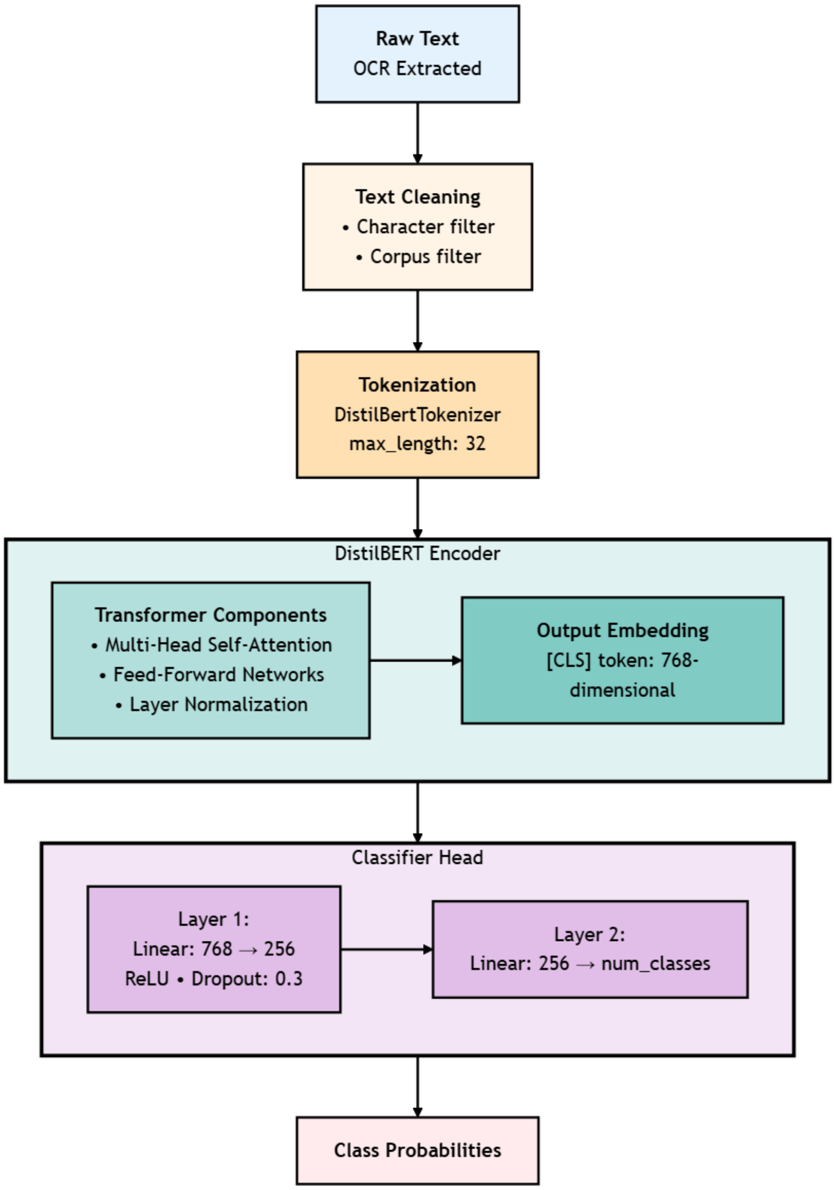

Figure 12: Text Feature Model Architecture Diagram

Text Feature Module Description

8.4 Multimodal Fusion Model

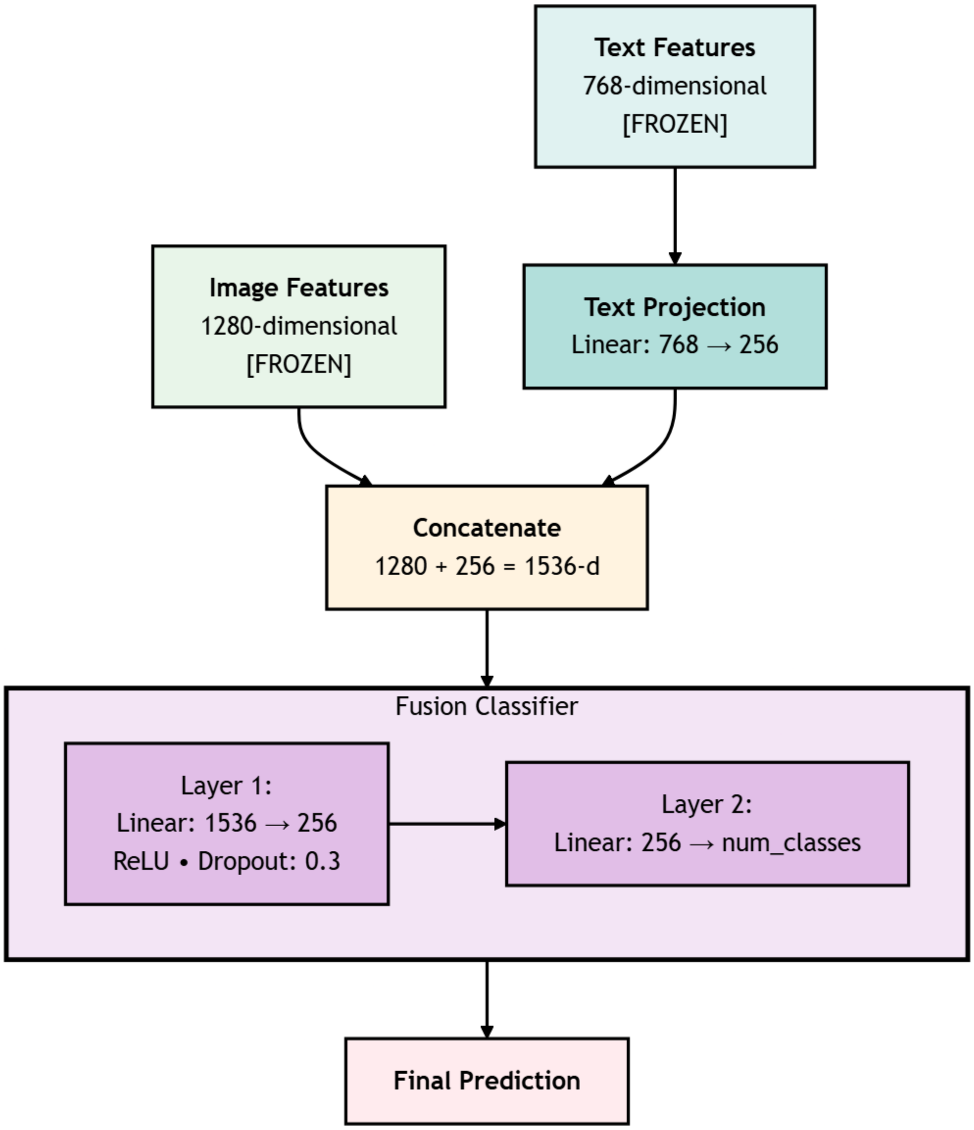

Figure 13: Late Fusion Model Architecture Diagram

Multimodal Fusion Module Description

8.5 GitHub Repository

The codebase for the project can be found on: Coursework Repo