Multimodal Advertisement Classification Using Image and Slogan with Deep Learning

Presenter: Avinna Maharjan | 20048839

Program: Deep Learning & Neural Networks| MSc. IT in AI

Agenda

Problem Statement

Dataset & Experimental Setup

Unimodal vs. Multimodal

Quantitative Results

Result Analysis & Interpretation

Limitations

Problem Statement & Research

Problem Statement

Advertisement classification is difficult due to varied visuals and promotional text

Single-modality models (image *or* text) miss complementary information

Ads combine branding imagery and slogans that jointly define category

Objective

Build a multimodal deep learning model using both image and text

Compare unimodal vs fusion performance to show improvement

Applications

Automated ad categorization, brand recognition, marketing analytics

Dataset & Experimental Setup

Dataset: MAdVerse

- **Source:** Zenodo
- **Classes:** 11 categories
- **Data Quality:** 99.42% complete

Preprocessing Pipeline:

- OCR Extraction
- Text Cleaning
- Train/Test Split

Training Configurations:

- **Epochs:** 15
- **Optimizer:** Adam
- **Loss:** CrossEntropyLoss

Unimodal vs. Multimodal Architecture

Model	Architecture	Input
Image Only	MobileNetV2 + MLP	224x224 images
Text Only	DistilBERT + MLP	OCR text
Multimodal	MobileNetV2 + DistilBERT (Late Fusion)	Image + text

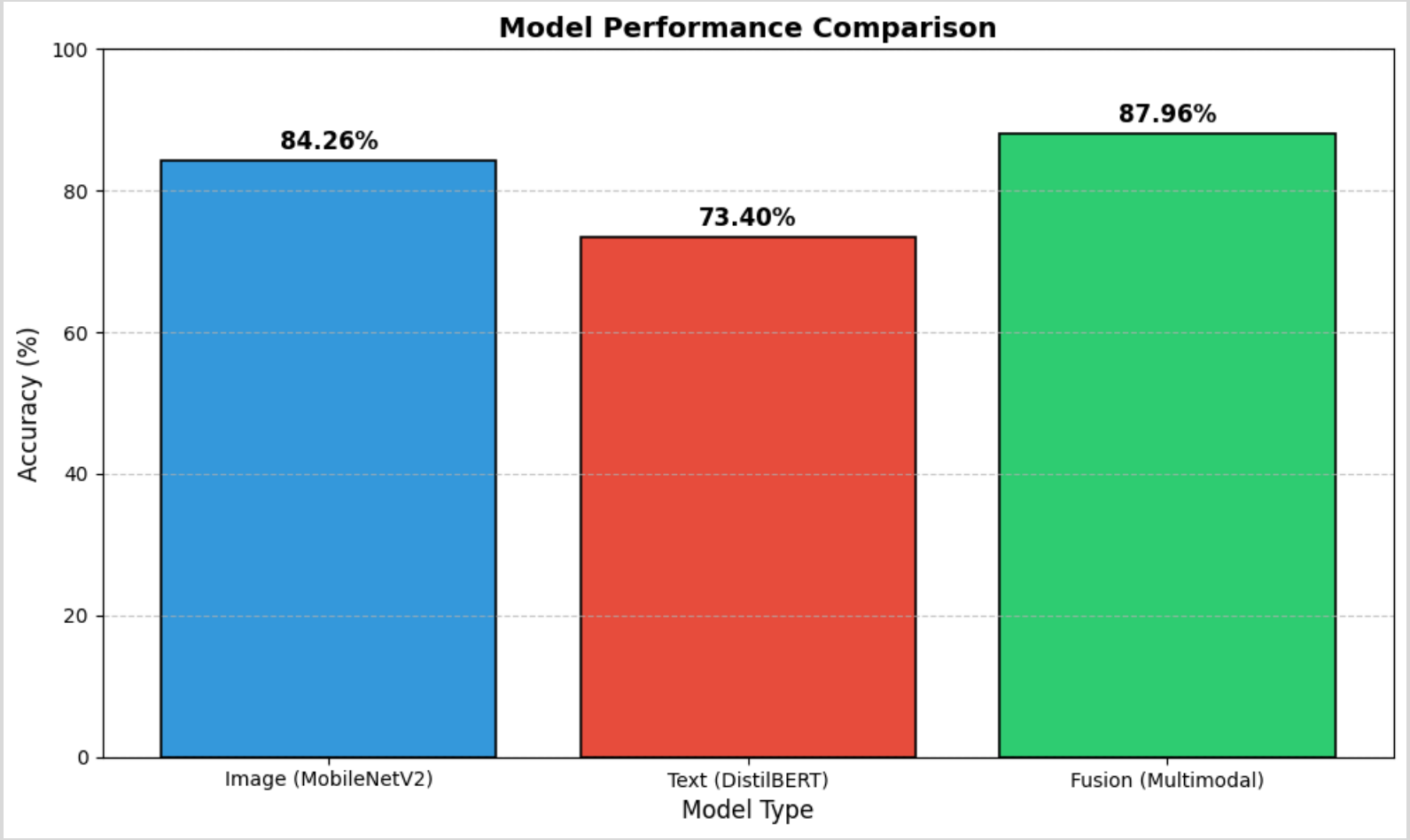
Fusion Strategy:

- Late Fusion approach: Train unimodal separately, freeze encoders
- Concatenate image and text features

Why Late Fusion?

- Leverages pretrained representations
- Allows independent feature learning per modality
- Simple yet effective concatenation

Quantitative Results



Result Analysis & Interpretation

Key Findings

- Image features outperform text features
- Multimodal fusion provides consistent improvement
- Total Misclassifications: 537 / 4459 (12.04%)

True: body_wear
Pred: food

True: drinks
Pred: drinks

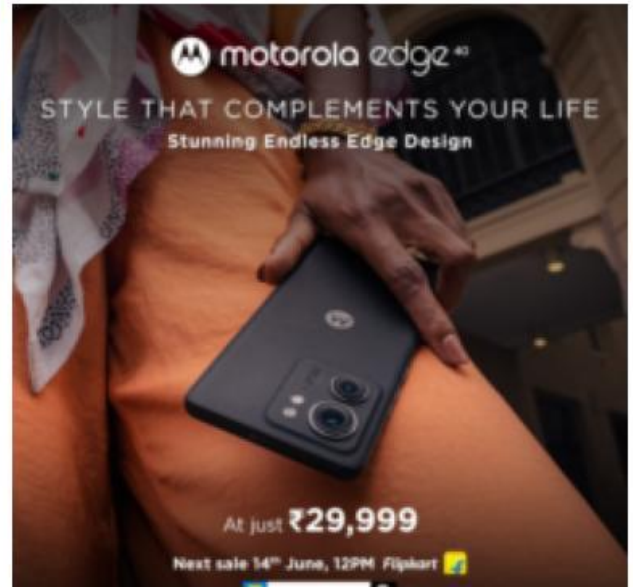
True: electronics
Pred: electronics

True: drinks
Pred: drinks

True: drinks
Pred: drinks

True: electronics
Pred: electronics

True: electronics
Pred: electronics



True: vehicles
Pred: vehicles

Limitations

- **Dataset Constraints**
- **OCR Dependency**
- **Computational Cost**
- **Limited Model Architecture**

Conclusion

- Built an effective multimodal ad classification framework
- Late fusion achieved highest accuracy
- Results confirmed multimodal learning significantly outperforms single-modality models

Thank You