Multimodal IMDB Analysis with TensorFlow and Keras

NAME: CYNTHIA CHINENYE UDOYE

STUDENT NO: 22029346

Introduction

In this project, two models were built and evaluated, a Convolutional Neural Network (CNN) and a Long Short-Term Memory network (LSTM), to classify film genres using multimodal data from the IMDB dataset. The dataset included film posters (images) and textual overviews. The CNN was trained to classify film genres based on posters, while the LSTM focused on classifying genres from textual overviews.

This report critically evaluates the performance of both models, identifies challenges encountered during training, and highlights areas for improvement.

Data Processing

1. Image Data Processing (Posters)

- Images were resized to 64x64 pixels to standardize input dimensions.
- The TensorFlow `tf.data` pipeline was used to efficiently preprocess the images.
- Image normalisation was performed to scale pixel values to the range [0,1], improving training stability.

2. Text Data Processing (Overviews)

- Textual overviews were tokenized using a text vectorisation layer, with sequences truncated or padded to a maximum length of 100 tokens.
- An embedding layer mapped tokens to a dense vector space of 256 dimensions.
- The pipeline ensured uniformity in input length and token representation.

Model Performance

1. CNN Results (Posters)

- **Architecture:** Six convolutional layers, dropout, and fully connected layers. Output: 25 genres.
- **Training:** Precision improved from 0.2386 to 0.6768; loss decreased from 0.4372 to 0.2103.
- **Validation:** Precision peaked at 0.6174; loss stabilized at 0.24199.
- **Key Observation:** Struggled with multi-genre predictions; often favored "Drama."

2. LSTM Results (Text)

- **Architecture:** Bidirectional LSTMs with dropout layers. Output: 25 genres.
- **Training:** Precision increased from 0.1142 to 0.5351; loss reduced from 0.6681 to 0.2491.
- **Validation:** Precision reached 0.6255; loss reached 0.2294.
- **Key Observation:** Better contextual understanding but struggled with rare genres like "Music."

Evaluation and Observations

The CNN model performed well in visually distinct genres with stable metrics but struggled with multi-genre classification and low recall. The LSTM model effectively captured textual

context and achieved better validation loss, though its recall was lower than precision, highlighting challenges with true positives. Both models were affected by class imbalance, favouring common genres like "Drama" and "Comedy," while underrepresented genres like "Music" led to misclassifications. Training and validation loss curves indicated consistent learning for both models, with steadily improving precision but relatively low recall. Below are the visualisations showcasing these observations.

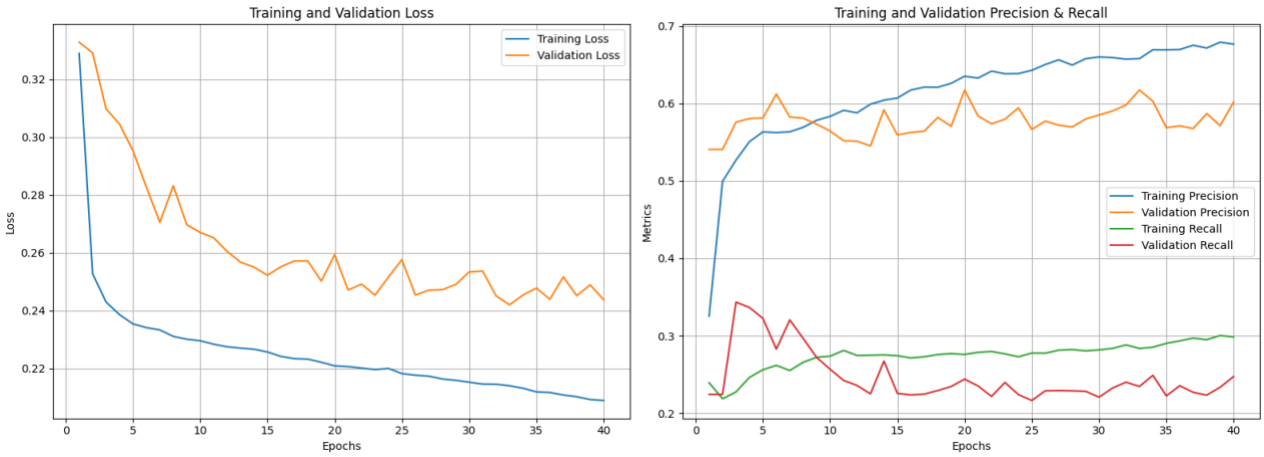


Figure 1: CNN Evaluation

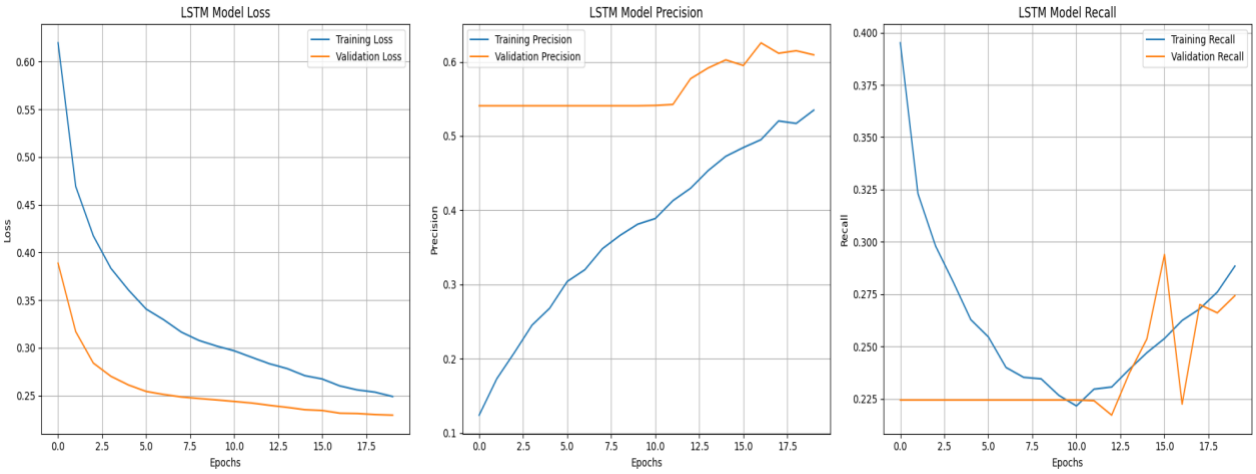


Figure 2: LSTM Evaluation

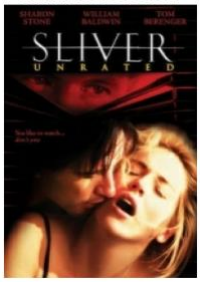
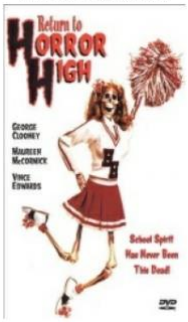
Recommendations for Improvement

1. **Address Class Imbalance:** Use oversampling or class-weight adjustments.
2. **Augmentation:** Apply data augmentation for poster images.
3. **Model Improvements:** Experiment with ResNet (CNN) or Transformer models (LSTM).
4. **Hyperparameter Tuning:** Optimize dropout rates, learning rates, and epochs.

Random Test Results

The table shows random test results, comparing CNN and LSTM predictions to ground truth genres with brief analyses:

Table 1: Comparison of Model Predictions with Ground Truth for Randomly Selected Films

Poster	Overview	LSTM, CNN & Ground Truth Genres	Analysis
Film ID: tt0108162 	A woman moves into an exclusive New York City apartment building, which she soon discovers houses tenants with all manner of shocking secrets.	**Ground Truth Genres:** Comedy, Romance **CNN Top 3 Predictions:** Drama (0.535177), Thriller (0.41146564), Crime (0.3190197) **LSTM Top 3 Predictions:** Drama (0.57104003), Comedy (0.44596103), Romance (0.20220152)	Both models identified Drama accurately. The CNN leaned toward Thriller and Crime, reflecting reliance on visual cues, while the LSTM captured Comedy and Romance, closer to the Ground Truth.
Film ID: tt0091019 	A Protestant World War II pilot and a Jewish girl fall in love in Jerusalem, even though their diverse backgrounds threaten to pull them apart.	**Ground Truth Genres:** Comedy, Crime, Drama **CNN Top 3 Predictions:** Drama (0.7104481), Romance (0.35832384), Thriller (0.28619272) **LSTM Top 3 Predictions:** Drama (0.7284702), Comedy (0.6159663), Romance (0.31404954)	Both models identified Drama accurately. The CNN leaned toward Romance and Thriller, missing Crime, while the LSTM aligned better with the Ground Truth by identifying Comedy and Drama effectively.
Film ID: tt0093854 	A few years ago, a mysterious serial-killer caused panic on Crippen High School. The killer was never caught. A movie company, Cosmic Pictures, has decided to make a feature movie about ...	**Ground Truth Genres:** Comedy, Drama, Romance **CNN Top 3 Predictions:** Drama (0.65049964), Comedy (0.48447198), Romance (0.42725855) **LSTM Top 3 Predictions:** Drama (0.58391047), Comedy (0.4360947), Romance (0.17776717)	Both models correctly identified Comedy and Drama as prominent genres. The CNN captured Romance more effectively than the LSTM, which struggled to assign high confidence to Romance.
Film ID: tt0092112 	A bullied teenage boy is devastated after the death of his heavy metal idol, Sammi Curr. But as Hallowe'en night approaches, he discovers that he may be the only one who can stop Sammi from making a Satanic comeback from beyond the grave.	**Ground Truth Genres:** Comedy, Drama, Music **CNN Top 3 Predictions:** Drama (0.7518147), Romance (0.3605625), Comedy (0.24803013) **LSTM Top 3 Predictions:** Drama (0.42050254), Comedy (0.31557244), Action (0.29273906)	Both models identified Drama and Comedy, but neither recognized Music as a genre. The CNN inaccurately included Romance, while the LSTM included Action, reflecting contextual limitations.

Conclusion

This project demonstrated the potential of multimodal data in film genre classification. While both models showed promising results, further optimisation and techniques to handle dataset imbalance are necessary for improved performance. The CNN and LSTM models complemented each other, with LSTM excelling in text analysis and CNN handling visual features effectively. Combining these modalities in a unified model could further enhance classification accuracy.