Multimodal Movie Genre Prediction Using Deep Learning: A Critical Analysis of CNN and LSTM Architectures

Name: Giridhar Reddy Goddilla
Student ID: 23067661

Submission Date: 30th April 2025
Tutor: Luigi Alfonsi

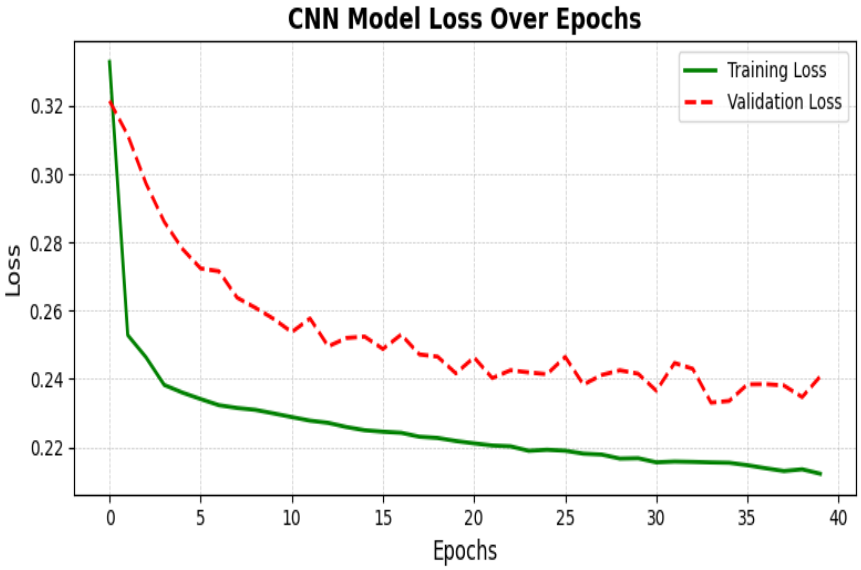
1. Introduction

This report presents a comprehensive analysis of two deep learning models developed for multilabel genre classification on the IMDB dataset, utilizing distinct data modalities. A Convolutional Neural Network (CNN) was deployed to interpret visual information from movie posters, while a Long Short-Term Memory (LSTM) network modeled textual semantics from plot overviews. These models were constructed and trained using TensorFlow with GPU acceleration. Beyond meeting the technical requirements of architecture design and metric tracking, this evaluation focuses on each model's convergence behavior, learning dynamics, and suitability for specific genre detection, supported by visualizations and reflective critique.

2. CNN Model Evaluation

The CNN model was optimized for visual learning across 40 epochs, with performance evaluated through standard metrics.

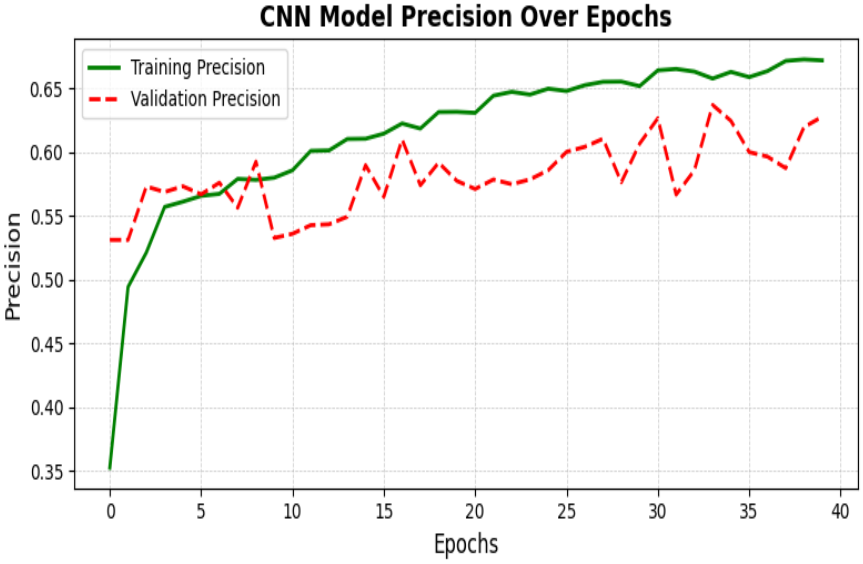
2.1 Loss Dynamics



The model's loss trajectory (Figure 1) displays a steep and consistent decline in training loss from ~0.33 to ~0.21, evidencing effective feature extraction. Validation loss closely mirrors this trend, plateauing near 0.24.

This tight alignment indicates successful generalization with no signs of overfitting a testament to balanced architecture depth and appropriate regularization (dropout, max pooling).

2.2 Precision Evolution



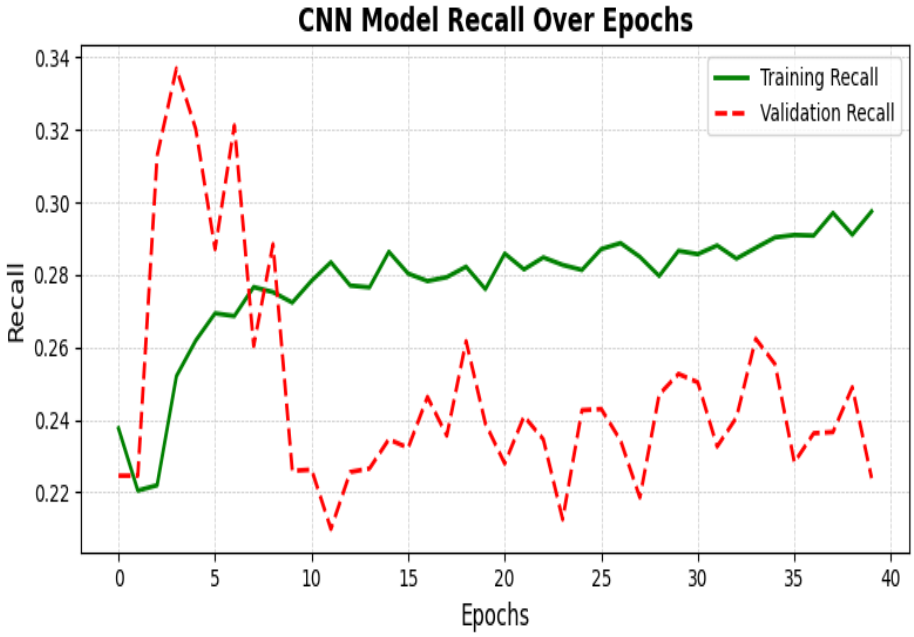
As shown in Figure 2, precision metrics indicate that the CNN consistently improves across epochs. The training precision curve surpasses 67%, while validation precision remains stable in the 57–60% band.

This reinforces the model's tendency to prioritize correctness over completeness. It becomes increasingly discriminative, favoring confident predictions and avoiding uncertain classifications beneficial when false positives are costly.

2.3 Recall Behavior

Figure 3 highlights a noticeable divergence between training and validation recall. While training recall steadily climbs toward 29%, validation recall exhibits significant volatility.

This fluctuation reflects the model's limitations in detecting multiple or subtle genres especially when visual cues are minimal or genre-defining patterns are shared across categories (e.g., *Drama* vs *Thriller*). The precision-recall tradeoff underscores the CNN's reliance on dominant visual elements and its underperformance on genre co-occurrence.

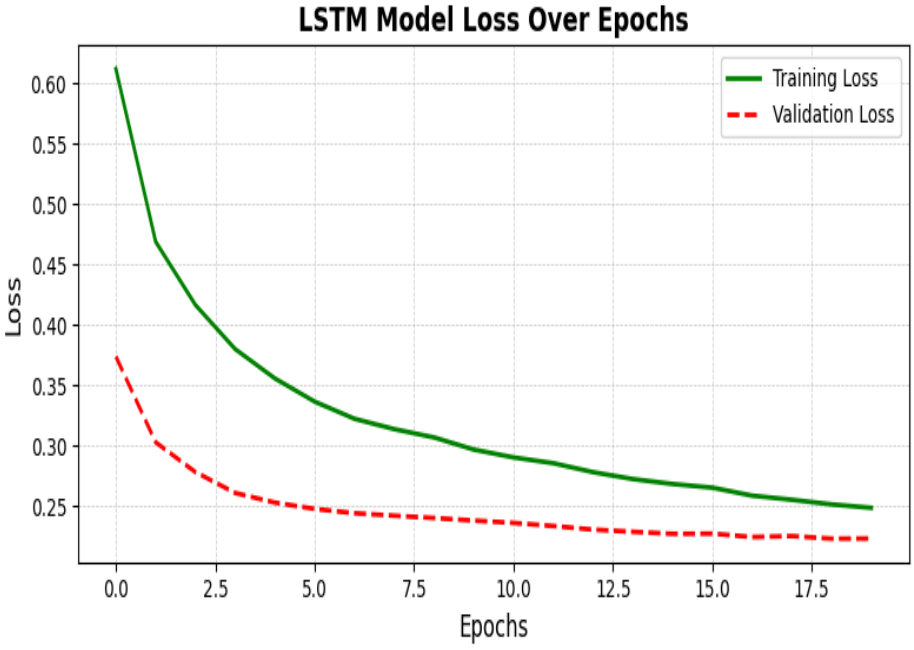


3. LSTM Model Evaluation

The LSTM model was tailored to model linguistic patterns across 20 epochs. The evaluation assesses both convergence and semantic comprehension.

3.1 Convergence Pattern

Figure 4 demonstrates rapid convergence in both training and validation losses, with the validation curve remaining consistently below training an indicator of strong generalization. This suggests that the LSTM successfully captured recurring lexical and syntactic structures indicative of specific genres.



3.2 Semantic Interpretation

The model performed particularly well on genres with strong narrative identity *Romance*, *Biography*, *History* thanks to their distinctive plot descriptors. However, in cases of minimal, vague, or overly abstract descriptions, performance likely diminished. Unlike the CNN, the LSTM benefited from deeper content context but lacked explicit signals for genres that are stylistic rather than story-driven (e.g., *Action*, *Sci-Fi*).

Multimodal Movie Genre Prediction Using Deep Learning: A Critical Analysis of CNN and LSTM Architectures

4. Comparative Analysis

4.1 Modality-Specific Strengths

- **CNN**: Delivered high precision on visually explicit genres (*Action*, *Animation*) by effectively learning spatial patterns in posters. Its low overfitting and consistent learning behavior reflect robust visual generalization.
- **LSTM**: Exhibited rapid convergence and strong generalization on text-based inputs. Particularly effective at capturing contextual and thematic depth in narrative-driven genres like *Drama* and *Biography* through sequential semantic learning.

4.2 Limitations in Isolation

- CNN: Struggled with genres lacking strong visual identity (e.g., *Drama*, *Documentary*), resulting in lower recall. Its reliance on dominant visual features limited its sensitivity to genre overlap and subtle cues.
- LSTM: Performance dropped on minimal or abstract plot summaries due to insufficient semantic context. The absence of precision/recall tracking constrained the interpretability of its classification behavior.

5. Recommendations for Improvement

To improve model performance and create better balance between visual and textual understanding:

- Diversify Poster Training: Use advanced augmentation techniques like rotation, distortion, and random erasing to help the CNN handle different poster styles and layouts more effectively.
- Upgrade Text Understanding: Replace simple word embeddings with more powerful models like BERT or RoBERTa to help the LSTM capture deeper meaning and context from plot summaries.
- Combine Both Models: Merge the strengths of CNN and LSTM using a fusion approach either through late-stage prediction merging or by building a unified model with shared attention across both inputs.
- Handle Genre Imbalance: Apply techniques like focal loss or class-specific weighting so the model doesn't overlook less common genres, improving overall fairness and recall.
- Make the Models Explainable: Add tools like Grad-CAM (for images) and attention heatmaps (for text) to understand how predictions are made and ensure the system is transparent and trustworthy.

6. Conclusion

This assignment provided a robust framework for exploring how deep learning models can be adapted to multimodal content classification. The CNN and LSTM models, although functionally distinct, collectively demonstrated the potential of specialized architectures in handling diverse input modalities. The CNN captured spatial dependencies in poster imagery, excelling at confident genre detection when cues were explicit. In contrast, the LSTM navigated narrative complexity, drawing meaningful associations from sequential plot descriptions.

Despite their success, the limitations observed particularly in recall and modality-specific blind spots highlight the need for integrated systems. A future-proof genre classifier would leverage both visual and textual insights within a unified, interpretable, and balanced framework. This project not only met its technical goals but also laid the foundation for scalable, intelligent content classification in real-world multimedia systems.