Universidad del Valle de Guatemala

Faculty of Engineering
CC3092 Deep Learning

Laboratory 2

Basic Neuronal Networks: Movies Genre Prediction

By: Samuel A. Chamalé
Email: cha21881@uvg.edu.gt

Submission Date: 8/4/2024

Contents

1 Notes and Resources

Data preprocessing was performed using the R programming language. The neural network modeling was conducted using Julia with the Flux library. The code and resources for this project are available at the following repository: `https://github.com/chamale-rac/movies-xd`.

2 Objective of the Network

The objective of the neural network in this laboratory is to **classify movie genres** based on a given dataset. The dataset initially comprises 4,380 records with various attributes related to movies, such as movie title, production date, genres, runtime in minutes, director information, and financial metrics.

2.1 Rationale for Using Neural Networks

Neural networks are particularly well-suited for this classification task due to their ability to model complex, non-linear relationships between the input features and the target variable. According to Zhang [Zha00], Neural networks excel in scenarios where traditional statistical models might fail due to the latter's underlying assumptions about data distributions. Neural networks can adaptively learn from the data without the need for explicit specification of functional forms, making them robust to the diverse and intricate patterns present in the selected movie attributes.

3 Pre-processing

3.1 Features Selection

In this specific case, we executed feature selection based on various scores such as chi-square, ANOVA, and simple correlation. Even though neural networks can potentially overperform using multiple features, we found that using all the variables was inconvenient and led to suboptimal results under the models we will explain further. By selecting a subset of relevant features, we aim to enhance the performance and accuracy of the neural network for genre classification.

For the purpose of this classification task, we selected a subset of features: *approval index*, *movie average rating*, *production budget*, *runtime minutes*, and *worldwide gross*.

3.2 Balanced Genre Sampling

Balanced Genre Sampling was employed to address the issue of imbalanced data in our dataset, particularly in the context of genre classification. There were several genres with poor representation, which could lead to biased predictions by the neural network. By selecting the top 7 genres based on their frequency and ensuring equal representation for each genre, we created a balanced dataset.

4 Composition and Results of Experimental Neural Network Models

This section describes three neural network models for genre classification, highlighting their architecture, performance, and key differences. The results for these neural networks were obtained using a batch size of 32, 20 epochs, and a learning rate of 0.001. The dataset used was from *movie_statistic_dataset.csv*. Notably, this dataset did not implement Balanced Genre Sampling.

4.1 Model 1: Simple Network with ReLU Activation and Dropout Regularization

Layer	Details
Layer 1	Dense (128 neurons, ReLU activation)
Layer 2	Dropout (rate = 0.3)
Layer 3	Dense (64 neurons, ReLU activation)
Layer 4	Dropout (rate = 0.3)
Output Layer	Dense (output size = number of genres, Softmax activation)

Figure 1: Model 1 Architecture

Metric	Value
Best Validation Accuracy	50.11% (epoch 13)
Final Train Accuracy	47.77%
Final Validation Accuracy	49.47%
Final Train Loss	5.92
Final Validation Loss	6.02

Figure 2: Model 1 Results

4.1.1 Discussion

Model 1 served as a baseline with ReLU activation and dropout regularization to mitigate overfitting. Despite these measures, its performance plateaued, suggesting limited model capacity. There was also noticeable overfitting in the later training epochs.

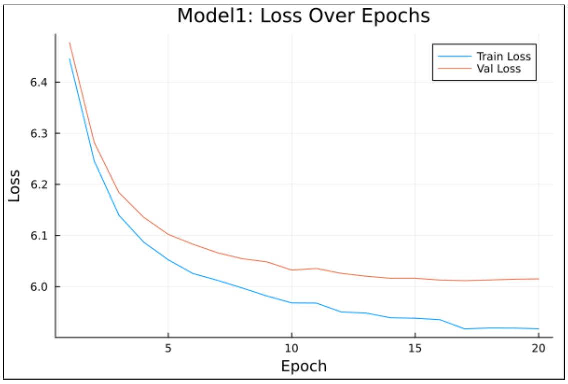

Figure 3: Model 1 Loss Over Epochs

4.2 Model 2: Deeper Network with ReLU Activation and L2 Regularization

Layer	Details
Layer 1	Dense (256 neurons, ReLU activation)
Layer 2	Dense (128 neurons, ReLU activation)
Layer 3	Dense (64 neurons, ReLU activation)
Output Layer	Dense (output size = number of genres, Softmax activation)

Figure 4: Model 2 Architecture

Metric	Value
Best Validation Accuracy	50.46% (epoch 20)
Final Train Accuracy	49.17%
Final Validation Accuracy	50.46%
Final Train Loss	24.40
Final Validation Loss	24.28

Figure 5: Model 2 Results

4.2.1 Discussion

Model 2, with increased depth and L2 regularization, showed higher training and validation losses compared to Model 1, suggesting overfitting despite the regularization. The increased complexity of Model 2 captured more patterns in the training data, but this did not translate to improved validation performance. Adjustments to the learning rate and regularization strength might be necessary to better control overfitting and improve generalization. Overall, while Model 2 has potential, it requires further tuning to balance complexity and regularization effectively.

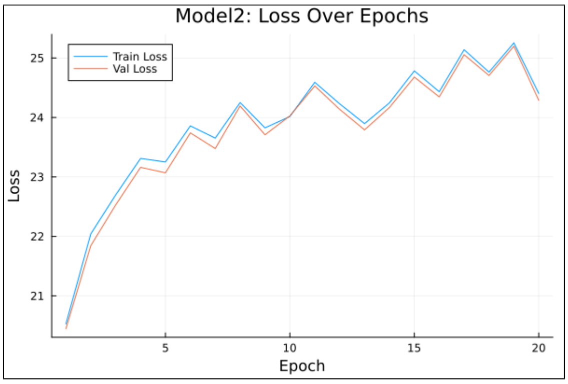

Figure 6: Model 2 Loss Over Epochs

4.3 Model 3: Wider Network with Tanh Activation and Batch Normalization

Layer	Details
Layer 1	Dense (512 neurons, Tanh activation), Batch Normalization
Layer 2	Dense (256 neurons, Tanh activation), Batch Normalization
Output Layer	Dense (output size = number of genres, Softmax activation)

Figure 7: Model 3 Architecture

Metric	Value
Best Validation Accuracy	51.14% (epoch 4)
Final Train Accuracy	47.77%
Final Validation Accuracy	46.35%
Final Train Loss	5.71
Final Validation Loss	6.06

Figure 8: Model 3 Results

4.3.1 Discussion

Model 3 incorporated wider layers, Tanh activation, and batch normalization. This setup achieved the highest early validation accuracy, indicating its effectiveness in capturing complex patterns at the initial epochs. However, performance plateaued, suggesting a need for further tuning. Additionally we see an xtreme overfitting.

Figure 9: Model 3 Loss Over Epochs

4.4 Summary

The three models showed varied success in genre classification. Model 1 served as a baseline with simple architecture and dropout regularization. Model 2's deeper architecture with L2 regularization improved performance slightly. Model 3's wider architecture with Tanh activation and batch normalization achieved the highest early validation accuracy, highlighting the importance of architectural choices and regularization techniques.

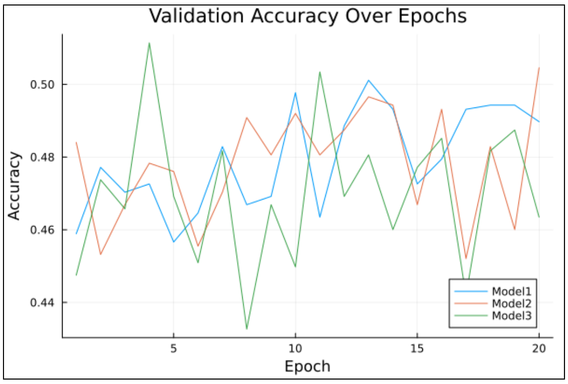

Figure 10: Models Accuracy Over Epochs

5 Optimal Neuronal Network

Layer	Details
Layer 1	Dense (512 neurons, ReLU activation), Dropout (0.5), Batch Normalization
Layer 2	Dense (256 neurons, ReLU activation), Dropout (0.5), Batch Normalization
Layer 3	Dense (128 neurons, ReLU activation)
Output Layer	Dense (output size = number of genres, Softmax activation)

Figure 11: Model 4 Architecture

Metric	Value
Best Validation Accuracy	53.52% (epoch 26)
Final Train Accuracy	48.30%
Final Validation Accuracy	47.01%
Final Train Loss	2.36
Final Validation Loss	2.36

Figure 12: Model 4 Results

Model 4 achieved the highest accuracy in the lab, reaching 53.2%, and demonstrated low training and validation loss, averaging around 2.3%. This model effectively addressed the overfitting issues observed in previous models. By incorporating dropout and batch normalization, Model 4 maintained a balanced complexity that allowed for better generalization. Additionally, using a balanced genre sampling approach mitigated the impact of initial genre frequency imbalances in the dataset. This fourth model was developed based on insights from the previous models, and several key changes were implemented: a batch size of 256 instead of 32, an adjusted learning rate using Flux.Optimise.ADAM(0.02) compared to the previous 0.001, and an extended training period of 50 epochs instead of 20. These adjustments collectively contributed to the enhanced performance and stability of Model 4. The dataset used was from *final_data_top_7_genres.csv*. Notably, this dataset did not implement Balanced Genre Sampling.

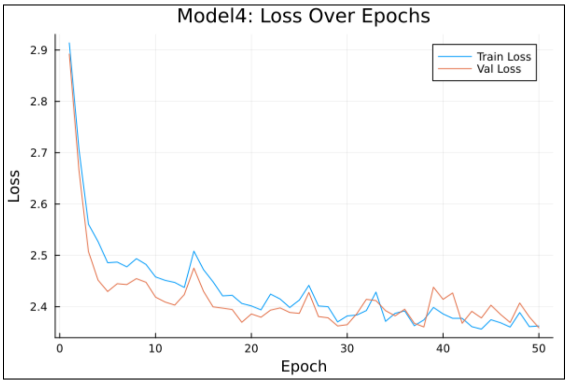

Figure 13: Model 4 (Optimal) Loss Over Epochs

References

[Zha00] G.P. Zhang. "Neural networks for classification: a survey". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 30.4 (2000), pp. 451–462. DOI: 10.1109/5326.897072.

List of Figures