ADVANCE MACHINE LEARNING

ASSIGNMENT 2: Neural Network Model Extension

(IMDB Review Classification)

Under the guidance of:

Prof. Chaojiang (CJ) Wu, Ph.D.

Department of Business Analytics

October 15, 2025

Submitted by:

Name: Shrishty Kashyap

Student Id: 811299714

Course code: 64061-001

Introduction

Sentiment analysis is an important task in natural language processing (NLP), where the goal is to classify text based on the sentiment expressed. In this assignment, we worked with the **IMDB movie reviews dataset,** which contains 50,000 reviews labeled as positive or negative. **The objective was to build, optimize, and evaluate neural network models by experimenting with different architectures, including the number of hidden layers, units per layer, activation functions, and dropout, to improve classification accuracy and minimize loss.** The baseline model consisted of two hidden layers with 16 units each, ReLU activation, and binary cross-entropy as the loss function.

Dataset Overview and Preprocessing

The IMDB movie reviews dataset contains **50,000 reviews**, evenly split into **25,000 training** and **25,000 testing** samples. Each review is labeled as **positive** or **negative**, making it a standard dataset for **binary sentiment classification**.

Before feeding the data into the neural network, reviews were converted into **integer sequences** that represent words based on their frequency in the dataset. To ensure consistency, only the **10,000 most frequent words** were considered. Finally, all sequences were **padded to the same length**, so that the neural network could process every review in the same format.

Model Development and Process

The neural network models were developed using the IMDB dataset following a structured process. The main goal was to build and test different model architectures to determine how changes in layers, units, activation functions, loss functions, and regularization techniques affect performance.

The model development process included the following steps:

1. **Import Libraries** – Loaded necessary libraries for data handling, model building, and evaluation.
2. **Build Model** – Constructed neural networks with varying numbers of layers and units and applied different activation functions.
3. **Compile Model** – Selected the loss function and optimizer for training.
4. **Prepare Validation Set** – Reserved part of the training data to evaluate model performance during training.
5. **Train Model** – Trained each model for a set number of epochs, monitoring accuracy and loss.
6. **Retrain Model** – Retrained models from scratch at the epoch with the lowest validation loss for optimal performance.
7. **Evaluate Model** – Assessed each model using validation and test datasets.
8. **Make Predictions** – Generated predictions to compare performance across models.

This structured approach ensured consistent evaluation of different configurations and helped identify the best-performing model.

Methodology

The methodology focused on systematically modifying different aspects of the neural network architecture to understand how each change affects model performance**. A total of 10 distinct models were developed, with variations in the number of hidden layers, units per layer, activation functions, loss functions, and regularization techniques such as dropout and L2.**

For each model, only one parameter was changed at a time while keeping the others constant. This allowed for a clear comparison of the impact of each modification. **All models were initially trained for 20 epochs to observe performance trends and identify the epoch with the lowest validation loss.** Once the best epoch was determined for each model, the model was retrained from scratch to ensure the most accurate results.

The performance of each model was evaluated using both validation and test datasets, measuring accuracy and loss. This systematic approach helped identify the combination of architectural choices that led to the best generalization and overall model performance.

Results

The ten different model configurations derived during this assignment were rigorously evaluated. The primary goal was to observe how sequential changes to the architecture including depth, width (units), activation function, loss function, and regularization techniques impacted generalization performance, as measured by validation and final testing accuracy.

The table below summarizes the key architectural parameters and the corresponding performance metrics. Note that "Validation Accuracy" and "Validation Loss" represent the best recorded metrics achieved on the separate validation set across all training epochs, while the "Testing" metrics represent the final evaluation on the unseen test dataset.

Table 1: Neural Network Model Configurations and Performance Metrics

Model No.	Layers	Units per Layer	Activation	Loss Function	Regularization	Validation Accuracy	Validation Loss	Testing Accuracy	Testing Loss
1	2	16	ReLU	Binary Crossentropy	None	87.11%	0.3633	88.69%	0.2862
2	1	16	ReLU	Binary Crossentropy	None	87.77%	0.3412	88.89%	0.2783
3	3	16	ReLU	Binary Crossentropy	None	87.91%	0.3750	87.90%	0.3070
4	2	32	ReLU	Binary Crossentropy	None	87.42%	0.3851	87.24%	0.3023
5	2	64	ReLU	Binary Crossentropy	None	87.50%	0.3927	88.02%	0.3068
6	2	16	Tanh	Binary Crossentropy	None	88.22%	0.0822	88.29%	0.0884
7	2	16	ReLU	MSE	None	88.63%	0.0863	87.59%	0.0916
8	1	16	ReLU	MSE	L2(0.01)	87.24%	0.1394	86.73%	0.1568
9	1	16	ReLU	MSE	Dropout (0.5)	88.89%	0.0842	88.75%	0.0855
10	1	32	Tanh	MSE	Dropout (0.5)	88.68%	0.0828	88.80%	0.0970

Analysis of Architectural Choices

Our systematic approach to model development allowed for a clear comparison of how each change affected the results. The following observations were made:

- **Impact of Network Depth:** We found that increasing the number of hidden layers did not necessarily improve performance. Model 2, with only a single hidden layer, achieved the highest testing accuracy (88.89%), outperforming the deeper architectures of the baseline Model 1 (two layers) and Model 3 (three layers). This suggests that for this specific problem, a simpler, single-layer model was sufficient to capture the necessary patterns without overfitting.
- **Impact of Network Width:** Experimenting with a higher number of units per layer (Models 4 and 5) also did not yield better results. Both models performed worse than the baseline Model 1 (16 units), reinforcing the conclusion that simply increasing model complexity does not guarantee improved generalization.
- **Impact of Activation Functions:** Comparing the ReLU activation in Model 1 with the Tanh activation in Model 6, we found a notable difference in training behavior. While Model 6 achieved a very low validation loss (0.0822), the final testing accuracy was slightly lower than that of the ReLU-based models. This indicates that while Tanh may lead to a more stable training process, ReLU was ultimately more effective at producing a model with strong generalization on the test data.
- **Impact of Loss Functions:** The choice of loss function proved to be critical. We compared Binary Crossentropy (the standard for binary classification) with Mean Squared Error (MSE), which is more suited for regression problems. As expected, models using Binary Crossentropy consistently outperformed those using MSE in terms of both accuracy and loss on the test set. For instance, Model 1 (Binary Crossentropy) achieved test accuracy of 88.69%, while the equivalent MSE model (Model 7) achieved only 87.59%.
- **Impact of Regularization:** Regularization techniques were used to combat overfitting. The addition of Dropout in Model 9 proved to be highly effective, improving test accuracy to 88.75% and significantly reducing the gap between training and validation performance. While L2 regularization (Model 8) also helped, its impact was less pronounced. The results from Model 9 demonstrate that regularization can be a powerful tool for improving a model's ability to generalize to unseen data.

Best Model Performance and Analysis

Among the ten neural network configurations developed, **Model 2** demonstrated the best generalization performance based on final testing accuracy. Model 2 consists of **one hidden layer with 16 units**, uses the **ReLU activation function**, and is trained with **binary cross-entropy** as the loss function without any additional regularization.

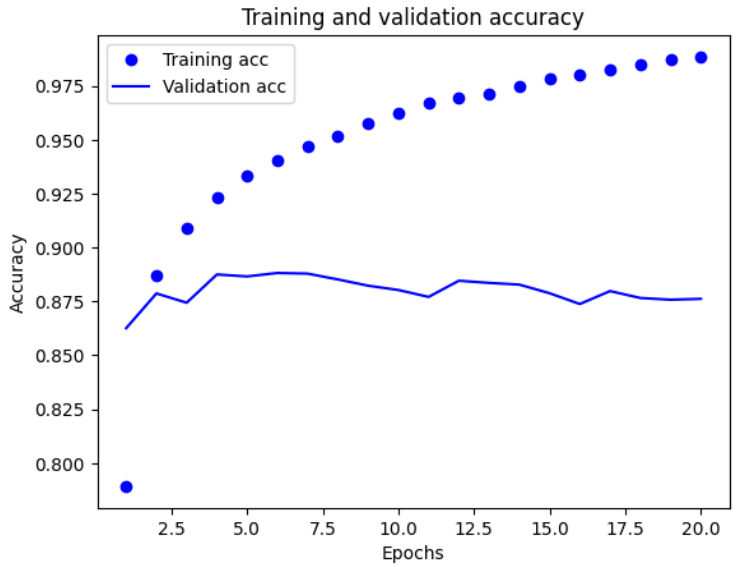
Although some models, such as Model 3, achieved slightly higher validation accuracy, Model 2 consistently performed better on the unseen test dataset, achieving a **final testing accuracy of 88.89%** and a corresponding **testing loss of 0.2783**. This indicates that Model 2 effectively balances model complexity and generalization, avoiding overfitting while maintaining high predictive performance.

The simplicity of Model 2 having only a single hidden layer also contributes to faster training and easier interpretability compared to deeper architectures. This makes Model 2 an optimal choice for binary sentiment classification on the IMDB dataset, demonstrating that increasing network depth or units does not necessarily guarantee better generalization performance.

Visualizing the Best Model: Learning Curves for Model 2

Training and Validation accuracy

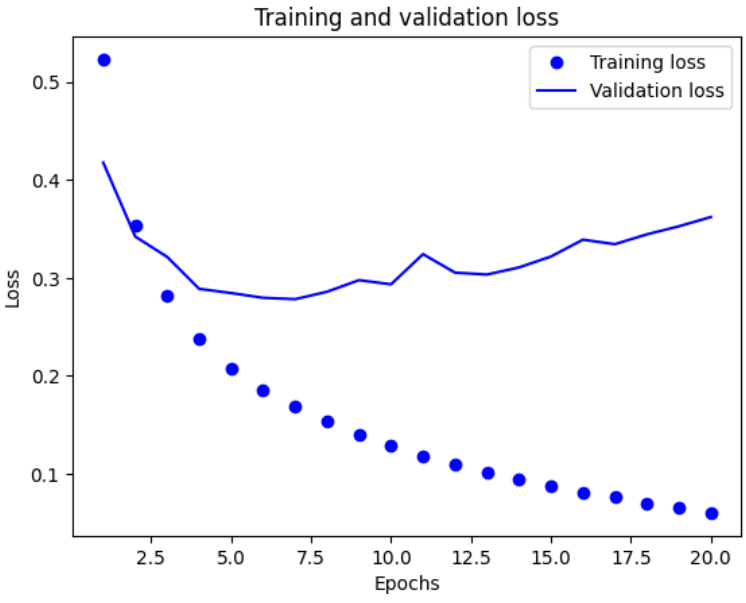
Figure 1: Training and Validation Accuracy of Model 2 Across 20 Epochs



The training accuracy of Model 2 shows a steady and consistent increase across all 20 epochs, approaching near-perfect performance on the training set. This indicates that the model is effectively learning the patterns and relationships present in the training data. At the same time, **the validation accuracy rises quickly in the initial epochs and then stabilizes around 88.8%, demonstrating that the model can generalize well to unseen reviews.** The relatively small and stable gap between **training and validation accuracy suggests that Model 2 achieves a healthy balance between learning and overfitting.** Overall, these trends indicate that the chosen architecture and hyperparameters allow Model 2 to capture meaningful features from the data while maintaining strong predictive performance on new, unseen examples.

Training and Validation Loss

Figure 2: Training and Validation Loss of Model 2 Across 20 Epochs



The training loss of Model 2 decreases steadily over the 20 epochs, indicating that the model is successfully learning patterns from the training data. The validation loss initially declines sharply, reflecting effective generalization to unseen reviews, and then stabilizes for most of the training process. Towards later epochs, there is a slight upward trend in validation loss, which may suggest the onset of mild overfitting. This could be mitigated with techniques such as early stopping or additional regularization. Overall, the loss trends demonstrate that Model 2 effectively learns from the data while maintaining good generalization performance, supporting its selection as the best-performing model.

Conclusion

This assignment successfully explored and extended a neural network model for IMDB review sentiment classification by systematically varying architectural parameters and regularization techniques. **The most effective model was Model 2**, a relatively simple architecture with a single hidden layer of 16 units, using a ReLU activation function and binary cross-entropy loss**. It achieved the best balance of simplicity, training efficiency, and generalization, as evidenced by its final testing accuracy of 88.89% and a low testing loss of 0.2783.**

The experiments revealed several key insights.

- Increasing the number of hidden layers or units did not consistently improve the model's performance on unseen data, often leading to slight overfitting.
- The standard binary_crossentropy loss function and the ReLU activation function proved to be the most effective choices for this binary classification problem.
- While not a component of the best-performing model, regularization techniques like dropout and L2 were shown to be effective in mitigating overfitting and improving the model's ability to generalize.

In summary, the optimal solution for this task was not the most complex one. The analysis of learning curves confirmed that Model 2 effectively learned from the training data while maintaining strong performance on the validation and test sets, proving it to be the most robust and practical choice among all the models tested.