AKANKSHA NADUKULA

anadukul@kent.edu

Assignment 2 - Neural Networks

Report

Data Collection:

The IMDb review dataset contains 50,000 movie reviews, with 25,000 reviews labeled as "positive" or "good" and 25,000 reviews labeled as "negative". The sentiment labels were assigned based on the overall tone of the reviews and reflect the subjective opinion of the reviewer about the movie. This dataset has been widely used for sentiment analysis and other natural language processing tasks in research and industry.

Data Pre-Processing:

For the IMDb review dataset, we performed several preprocessing steps to convert the raw text data into a format suitable for training a neural network. Firstly, we only considered the top 10,000 most frequent words in the dataset, as including all the words would result in a very high-dimensional input space. Next, we converted the text reviews into integer representations using a dictionary mapping the words to their corresponding indices in the top 10,000-word list.

However, neural networks cannot take integers as input, so we needed to convert the integer representations to tensors. To achieve this, we ensured that all reviews were of the same length by padding shorter reviews with zeros and truncating longer reviews. This resulted in a fixed-length vector representation of each review, where each element corresponded to the index of a particular word in the dictionary.

Finally, we performed one-hot encoding on the integer representations to convert them into binary values. This resulted in a binary matrix representation of the data, with each row corresponding to a review and each column corresponding to a particular word in the dictionary.

To evaluate the performance of our neural network model, we split the dataset into training and testing sets. We randomly selected 80% of the data for training and the remaining 20% for testing, ensuring that the sentiment distribution was roughly the same in both sets. This allowed us to train the model on a subset of the data and evaluate its performance on unseen data.

Approaches:

To determine the most effective neural network architecture for the IMDb review dataset, we experimented with different network configurations and analyzed their performance on the training and validation sets. Specifically, we trained and evaluated three different network architectures: a three-layer network with 60 neurons in each layer, a two-layer network with 60 neurons in each layer, and a four-layer network with 60 neurons in each layer.

For each network architecture, we tracked the accuracy and validation loss over epochs to observe how the model's performance improved over time. We then compared the results of the different approaches to determine which architecture provided the best results.

The use of different network architectures allowed us to explore the impact of network depth and width on model performance. By testing multiple configurations, we gained insights into how changes in architecture affect the learning process and the model's ability to generalize to new data. Overall, this approach enabled us to identify the most effective neural network architecture for the IMDb review dataset.

Continuing from the previous answer, we also experimented with different loss functions and activation functions to see how they impacted the performance of the different network architectures. Initially, we used binary_crossentropy as the loss function and relu as the activation function for all three network architectures.

After training and evaluating these initial network configurations, we then tested the same three architectures with a different loss function and activation function combination: mse as the loss function and tanh as the activation function.

By comparing the performance of the different network configurations with these alternate loss and activation functions, we gained further insights into how the choice of loss and activation functions impacts the learning process and the model's ability to generalize to new data. This approach allowed us to make more informed decisions about the selection of loss and activation functions for neural network models trained on similar datasets in the future. Overall, this comprehensive experimentation and analysis enabled us to identify the most effective neural network architecture, loss function, and activation function combination for the IMDb review dataset.

Results:

This table summarizes the accuracy achieved by different neural network architectures trained on the IMDb review dataset, with different loss and activation function combinations.

The first column lists the number of layers in the network, while the second column indicates the loss function used in training. The third column indicates the activation function used in training. The fourth column lists the accuracy achieved by the network on the test set.

Based on the results in the table, it can be observed that the 2-layer network with mse as the loss function and tanh as the activation function achieved the highest accuracy of **88.2%.** Meanwhile, the 3-layer and 4-layer networks also achieved high accuracy, with the best performance obtained with the mse loss function and tanh activation function. Overall, these results highlight the importance of experimenting with different network configurations, loss functions, and activation functions to identify the best model for a particular task.

Network Architecture	Loss Function	Activation Function	Accuracy
3-layer	binary_crossentropy	relu	87.5%
3-layer	mse	tanh	88.1%
2-layer	binary_crossentropy	relu	88.0%
2-layer	mse	tanh	88.2%
4-layer	binary_crossentropy	relu	88.2%
4-layer	mse	tanh	88.1%