National Tsing Hua University

1130IEEM 513600

Deep Learning and Industrial Applications

Homework 2

113034506 李家欣

1. (20 pts) Select 2 hyper-parameters of the artificial neural network used in Lab 2 and set 3 different values for each. Perform experiments to compare the effects of varying these hyper-parameters on the loss and accuracy metrics across the training, validation, and test datasets. Present your findings with appropriate tables.

A. Selected Hyper-parameter:

(1) learning rate:0.01, 0.001, 0.0001

(2) batch size: 16, 32, 64

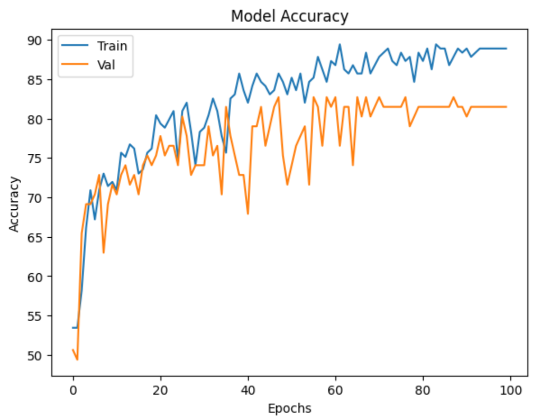
B. Experiment Setup

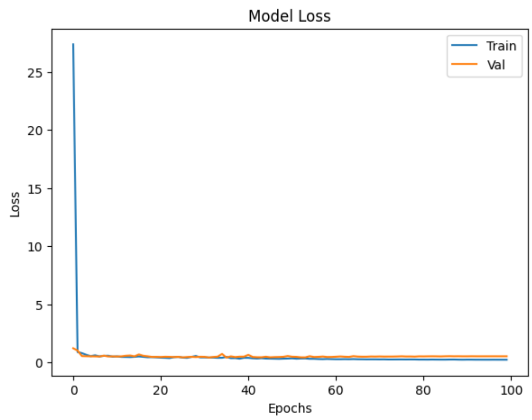
(1) Each experiment is conducted three times

(2) The average is used for comparison: Train loss, Train Accuracy, Test loss, Test Accuracy, Valid Accuracy

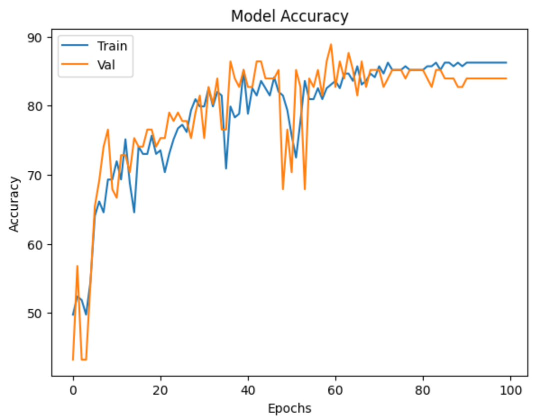
C. Result

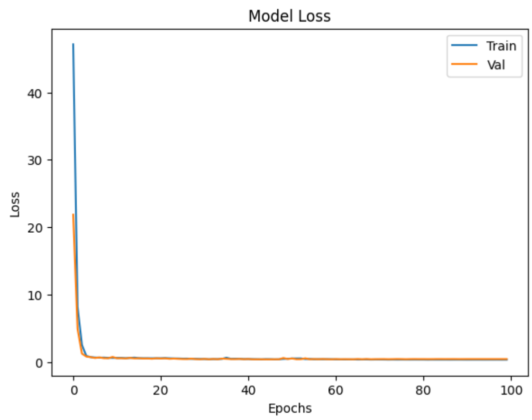
(1) learning rate: 0.01, batch size: 16



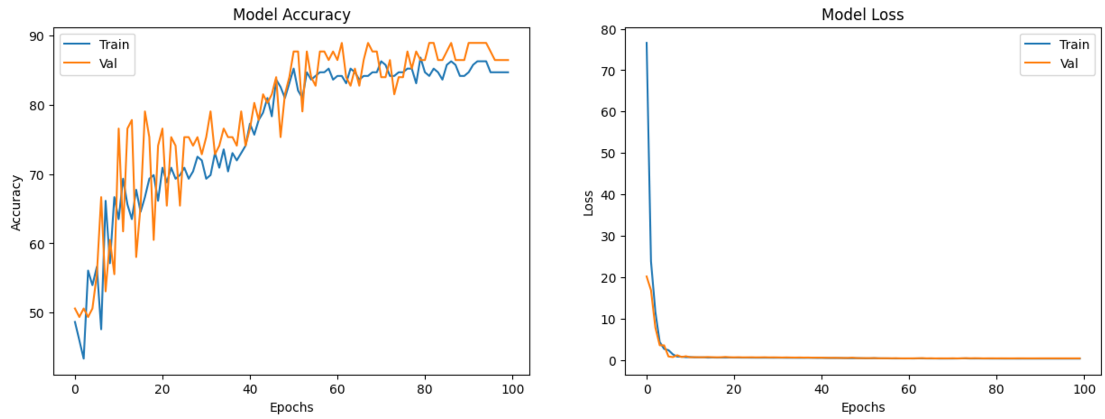


(2) learning rate: 0.01, batch size: 32

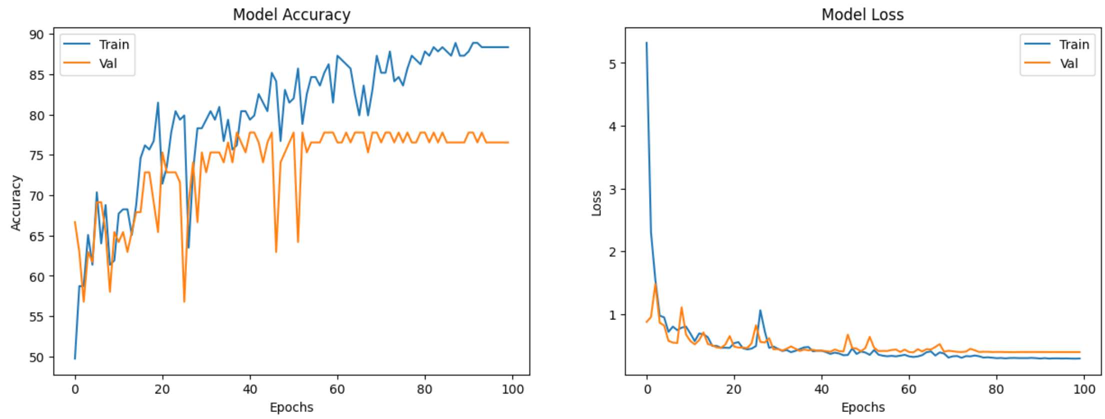




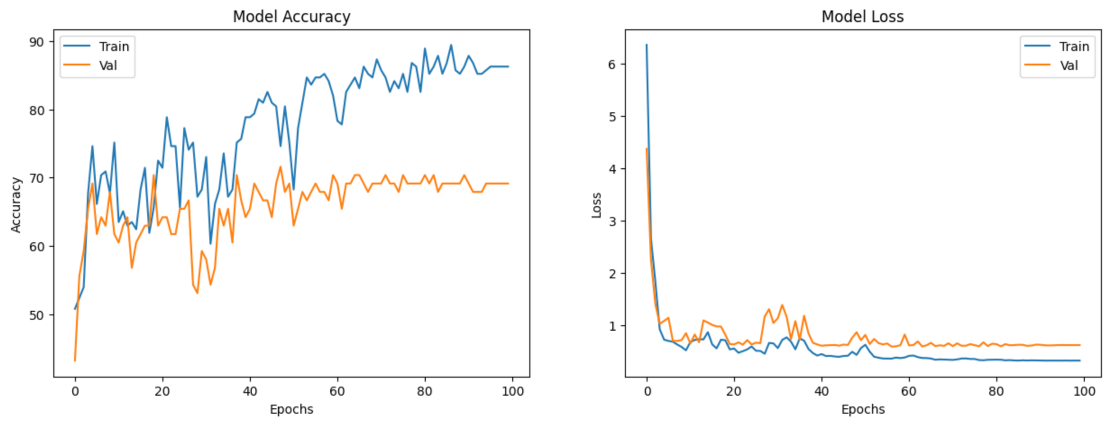
(3) learning rate: 0.01, batch size: 64



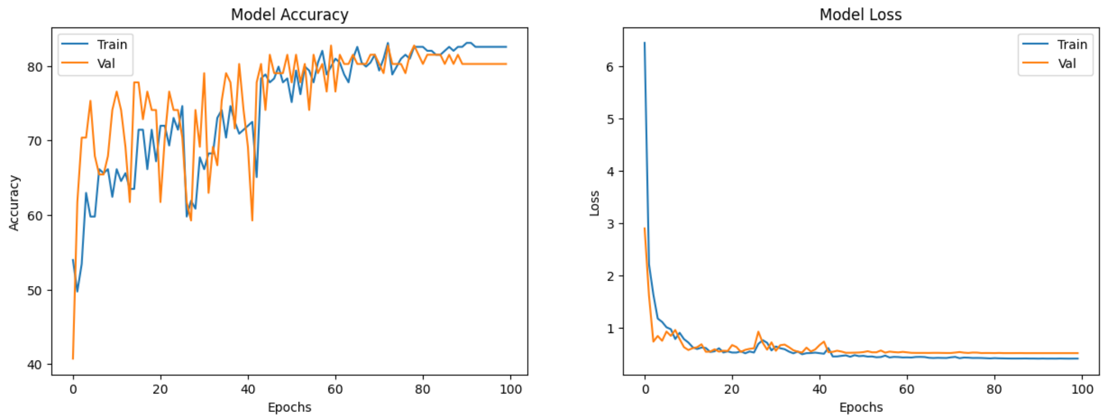
(4) learning rate: 0.001, batch size: 16



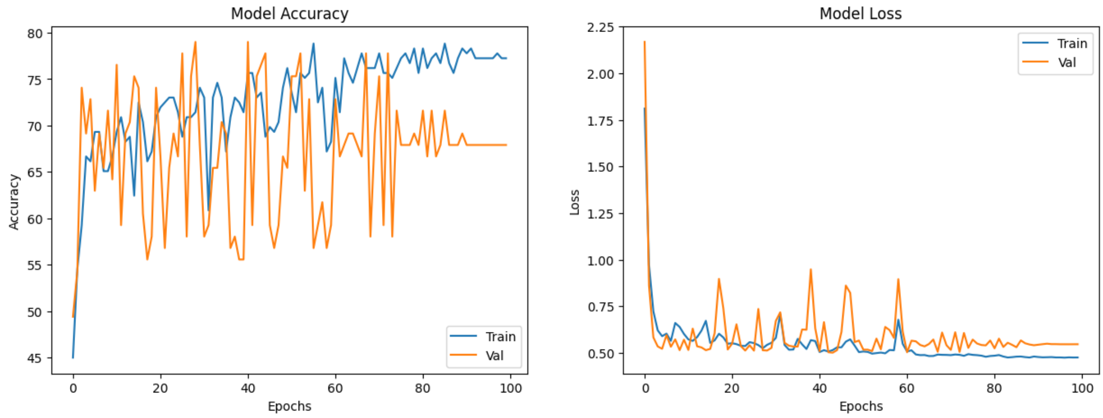
(5) learning rate: 0.001, batch size: 32



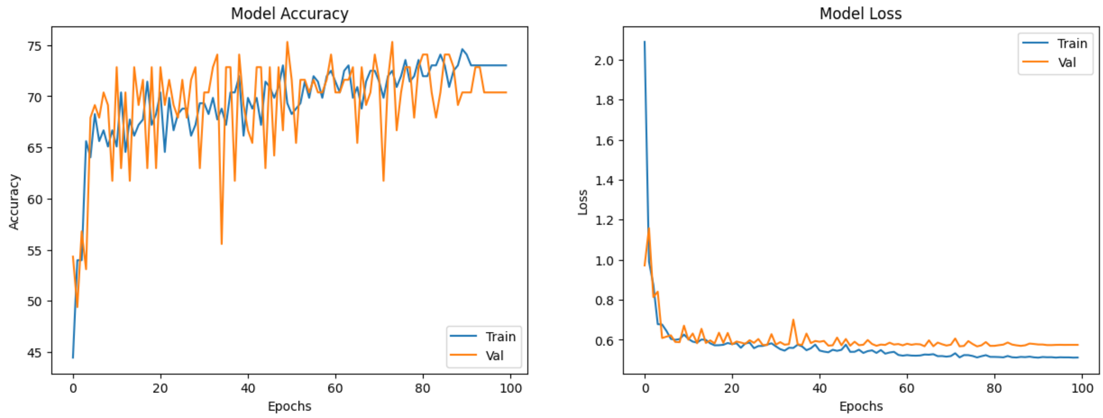
(6) learning rate: 0.001, batch size: 64



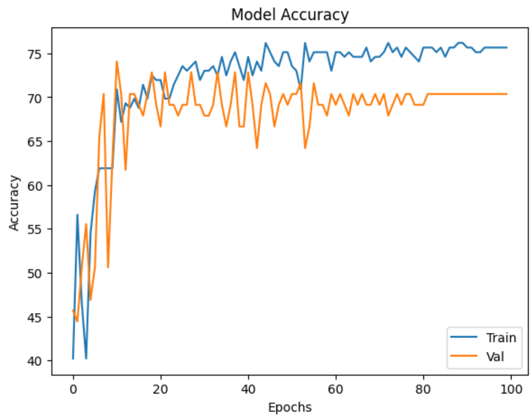
(7) learning rate: 0.0001, batch size: 16

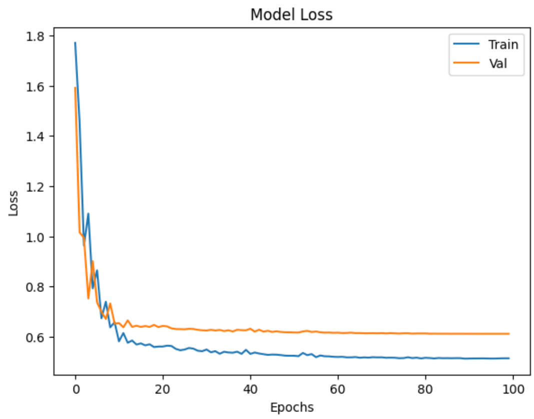


(8) learning rate: 0.0001, batch size: 32



(9) learning rate: 0.0001, batch size: 64



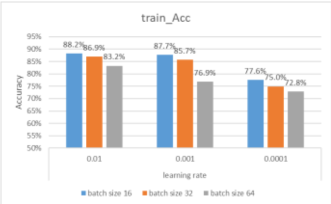


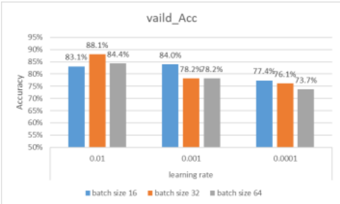
Findings:

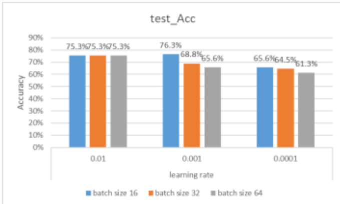
- Effect of Batch Size:
  Under the same learning rate, comparing different batch sizes (Figures 7, 8, and 9) reveals that a smaller batch size (16) results in more fluctuations in loss and requires a longer time to converge. As batch size increases, the loss curve becomes smoother, indicating stability.
- Effect of Learning Rate:
  The learning rate controls the step size of parameter updates. By comparing Figures (8) and (2), it can be observed that with a smaller learning rate, Train Accuracy increases slowly, while with a larger learning rate, training is faster but less stable.
- Validation Accuracy Stability:
  From Figure (7), it can be seen that the fluctuations in validation accuracy are large, indicating that a smaller batch size leads to greater randomness in gradient updates, causing the model to fluctuate significantly between different epochs.
- Overfitting:
  Figures (4), (5), and (7) show a significant difference between Train Accuracy and Validation Accuracy, indicating the occurrence of overfitting.

2. (20 pts) Based on your experiments in Question 1, analyze the outcomes. What differences do you observe with the changes in hyper-parameters? Discuss whether these adjustments contributed to improvements in model performance, you can use plots to support your points. (Approximately 100 words.)

(1) A smaller learning rate (0.0001) results in slower but steady improvement in accuracy, while a larger learning rate (0.01) accelerates training but introduces instability.
(2) smaller batch sizes (16) exhibit greater variance in loss curves, implying unstable gradient updates, whereas larger batch sizes (64) lead to smoother convergence but may limit generalization.
(3) After conduct each experiment three times and then calculate the average of accuracy, learning rate of 0.01 and batch size of 32 achieved the best overall performance across training, validation, and test datasets.







3. (20 pts) In Lab 2, you may have noticed a discrepancy in accuracy between the training and test datasets. What do you think causes this occurrence? Discuss potential reasons for the gap in accuracy. (Approximately 100 words.)

The gap in accuracy indicates that the model might have poor real-world performance because it is poor at generalizing to unseen data.
Potential reasons:
(1) Overfitting: The model learns too much details and noises in training, and thus fails to generalize
(2) Training and test sets come from different distributions
(3) Insufficient training samples may fail to capture generalizable patterns
(4) The test set may contain harder or noisier samples
Among these, overfitting is the most common issue, which can be detected by comparing the loss curves between training and test (if training loss keeps decreasing while test loss increases, overfitting is likely occurring).

4. (20 pts) Discuss methodologies for selecting relevant features in a tabular dataset for machine learning models. Highlight the importance of feature selection and how it can impact model performance. You are encouraged to consult external resources to support your arguments. Please cite any sources you refer to. (Approximately 100 words, , excluding reference.)

Feature selection aims to identify the optimal subset of features. The importance and the benefits are:

(1) Improve model efficiency and performance by eliminating irrelevant or redundant ones,

(2) Decreases computational cost by reducing feature size

The table provided is an experiment predicting dementia family caregiver chronic stress levels using three regression models with different feature selection methods (Filter, Wrapper, Embedded). We can figure that selecting suitable feature selection can improve prediction accuracy compared to using regression models alone.

表 12 預測模型的性能比較:因素 1

因素 1: 家庭照顧者因睡眠問題導致的日常干擾

模型及特徵選擇方法	評估指標		
	MSE	MAE	R 平方
多元線性回歸*	0.77	0.68	19%
多元線性回歸+皮爾森相關係數	0.67	0.6	30%
多元線性回歸+向前特徵選擇	1.16	0.83	5%
多元線性回歸+RFE	0.68	0.61	29%
多元線性回歸+Lasso 特徵	0.72	0.66	18%
隨機森林*	0.78	0.65	20%
隨機森林+皮爾森相關係數	0.73	0.62	25%
隨機森林+向前特徵選擇	0.58	0.55	32%
隨機森林+RFE	0.79	0.65	19%
隨機森林+內建	0.54	0.60	34%
Lasso 回歸*	0.66	0.61	33%
Lasso+皮爾森相關係數	0.67	0.60	30%
Lasso+向前特徵選擇	**0.55**	**0.54**	**38%**
Lasso+RFE	0.66	0.62	32%
Lasso+內建	0.63	0.71	19%

*未進行任何特徵選擇

Reference:

Yeh, Wei-Chang, Chiu, Yi-Chen (2024). Exploring predictive models of dementia family caregiver chronic stress: Multivariate research based on machine learning

5. (20 pts) While artificial neural networks (ANNs) are versatile, they may not always be the most efficient choice for handling tabular data. Identify and describe an alternative deep learning model that is better suited for tabular datasets. Explain the rationale behind its design specifically for tabular data, including its key features and advantages. Ensure you to reference any external sources you consult. (Approximately 150 words, excluding reference.)

Traditional deep learning models worse at tabular data, categorical features, and low-dimensional feature spaces. They need large datasets to perform better than decision tree-based models. Therefore, Google introduced TabNet in 2019.

A. Key Structure

(1) Apply DNN in a way with the idea of decision trees, allowing structured decision-making.

(2) It directly takes raw tabular data without preprocessing and is optimized using gradient descent, making it easy to integrate into end-to-end learning.

(3) Sequential Attention: at each decision step, it selectively focuses on the most relevant features, improving efficiency and interpretability.

B. Advantages

(1) Handles Numerical and Categorical Features: Avoids one-hot encoding, preventing dimensional explosion and making it suitable for datasets with many categorical variables.

(2) Superior Performance: The paper demonstrates that TabNet show better performance than traditional machine learning models like XGBoost on real-world datasets.

(3) Built-in Interpretability: Uses attention weights to analyze feature importance, making model decisions more transparent.

Reference:

Sercan O. Arik, Tomas Pfister (2019). TabNet: Attentive Interpretable Tabular Learning