29.12.2024

CSE4288 Introduction to Machine Learning
Team Project Fall 2024
Model Evaluation Report

Team Members:

1- Eren Duyuk 150120509
2- Ufuk Acar 150121071
3- Yusuf Demir 150120032
4- Emir Uyar 150120007
5- Muhammed Hayta 150121068

Introduction

The Child Drawing Classifier project aims to develop a machine learning model capable of classifying children's imaginative and unique drawings into predefined categories such as animals, plants, and everyday objects. This innovative approach addresses challenges posed by the abstract and unpredictable nature of children's artwork, as well as the limited availability of labeled data for training. By employing advanced techniques in image classification, this project has significant potential to enhance educational tools and provide insights into children's cognitive and artistic development.

In this report, we evaluate the performance of the developed classification model using various metrics, compare its effectiveness against simpler baseline models, and detail the optimization strategies implemented to improve accuracy and generalization. The primary model, a Convolutional Neural Network (CNN), has been rigorously tested and optimized through techniques such as data augmentation, hyperparameter tuning, and architectural adjustments.

The evaluation process emphasizes understanding both the strengths and limitations of the CNN-based classifier, leveraging metrics such as accuracy, precision, recall, and F1 score to assess its performance across multiple categories. Additionally, we experiment with alternative algorithms, including Naive Bayes, K-Nearest Neighbors (KNN), and Decision Trees, to contextualize the CNN's results and provide a comprehensive comparison.

Through this report, we aim to demonstrate how machine learning can effectively tackle the complexities inherent in children's drawings, highlight the importance of rigorous evaluation in achieving high-performing models, and offer insights into the iterative optimization process that drives improvement in classification tasks.

Performance Metrics

Naive Bayes

```
                         Train size: 2724, Test size: 682

                                    Test Accuracy: 71.99%

                       precision    recall  f1-score   support

                bird       0.41      0.89      0.56        56
                 car       0.96      0.85      0.90        60
               cloud       0.82      0.78      0.80        64
                 dog       0.62      0.20      0.31        49
              flower       0.69      0.60      0.64        62
               house       0.90      0.89      0.89        61
               human       0.70      0.54      0.61        59
            mountain       0.88      0.79      0.83       161
                 sun       0.84      0.82      0.83        45
                tree       0.53      0.66      0.59        65

            accuracy                           0.72       682
           macro avg       0.73      0.70      0.70       682
        weighted avg       0.76      0.72      0.72       682
```

The Naive Bayes model achieves an overall test accuracy of **71.99%**, which is moderate. While it performs well on categories like **car** (F1: 0.90) and **house** (F1: 0.89) due to their distinct patterns, it struggles with classes like **dog** (F1: 0.31) and **bird** (F1: 0.56).

This discrepancy likely stems from Naive Bayes' assumption of feature independence, which may not hold for more complex or overlapping features in categories such as **dog** and **bird**. The macro average F1 score of **0.70** and weighted average of **0.72** reflect that the model's performance is skewed towards classes with higher support, such as **mountain** and **cloud**.

Decision Tree

```
                              Train size: 2724, Test size: 682

                                        Test Accuracy: 57.92%

                         precision     recall  f1-score   support

                         bird   0.56   0.57   0.57          56
                         car    0.70   0.55   0.62          60
                 cloud          0.60   0.45   0.52          64
                         dog    0.41   0.49   0.44          49
                 flower         0.53   0.50   0.52          62
                 house          0.61   0.61   0.61          61
                 human          0.42   0.44   0.43          59
                         mountain      0.72   0.76   0.74   161
                         sun    0.60   0.67   0.63          45
                         tree   0.43   0.46   0.45          65

                         accuracy                    0.58   682
                  macro avg            0.56   0.55   0.55   682
                weighted avg           0.58   0.58   0.58   682
```

The Decision Tree classifier demonstrates moderate performance, achieving a test accuracy of 57.92%. The mountain category performs the best, with an F1-score of 0.74, indicating that the model effectively identifies this class. However, other categories such as dog and tree have lower F1-scores (0.44 and 0.45, respectively), highlighting areas for improvement. The weighted averages for precision, recall, and F1-score are all approximately 0.58, reflecting the model’s overall balanced but unsatisfactory performance. This suggests that while the Decision Tree captures some patterns, it struggles with more nuanced distinctions in children’s drawings, potentially due to the inherent complexity of the dataset.

Logistic Regression

```
                        Train size: 2724, Test size: 682

                                   Test Accuracy: 90.32%

                   precision      recall  f1-score   support

                      bird    0.88    0.95    0.91         56
                      car     0.98    0.97    0.97         60
               cloud          0.92    0.89    0.90         64
                      dog     0.93    0.86    0.89         49
               flower         0.93    0.81    0.86         62
               house          0.98    0.92    0.95         61
               human          0.79    0.85    0.82         59
                   mountain       0.95    0.94    0.94    161
                      sun     0.86    0.96    0.91         45
                      tree    0.77    0.86    0.81         65

                   accuracy                       0.90    682
                macro avg         0.90    0.90    0.90    682
              weighted avg        0.91    0.90    0.90    682
```

The Logistic Regression classifier exhibits strong performance, achieving a high test accuracy of 90.32%. Most categories, such as car and house, have excellent F1-scores (0.97 and 0.95, respectively), reflecting the model's ability to accurately classify these classes. The mountain and sun categories also show high precision and recall values, contributing to the overall success. However, performance for human and tree is slightly lower, with F1-scores of 0.82 and 0.81, indicating room for refinement in handling these classes. The macro and weighted averages for precision, recall, and F1-score are all 0.90, underscoring the model's robust and balanced classification capabilities across multiple categories.

KNN

```
                         Train size: 2724, Test size: 682

                                    Test Accuracy: 84.60%

               precision     recall  f1-score   support

                 bird   0.79    0.89    0.84          56
                 car    0.96    0.90    0.93          60
      cloud             0.83    0.86    0.85          64
                 dog    0.65    0.86    0.74          49
      flower            0.80    0.76    0.78          62
      house             0.98    0.90    0.94          61
      human             0.81    0.71    0.76          59
               mountain       0.96    0.88    0.92   161
                 sun    0.91    0.89    0.90          45
                 tree   0.69    0.78    0.73          65

               accuracy                       0.85   682
       macro avg              0.84    0.84    0.84   682
    weighted avg              0.86    0.85    0.85   682
```

The K-Nearest Neighbors (KNN) classifier performs well, achieving a test accuracy of 84.60%. Categories like house and mountain show excellent performance, with F1-scores of 0.94 and 0.92, respectively. Similarly, the car and sun categories maintain high F1-scores (0.93 and 0.90), reflecting strong precision and recall. However, performance dips for dog and tree, with F1-scores of 0.74 and 0.73, indicating some challenges in classifying these classes. The macro and weighted averages for precision, recall, and F1-score hover around 0.84-0.85, demonstrating a well-balanced performance overall, but with potential for improvement in less consistent categories.

Convolutional Neural Network (CNN)

```
              precision    recall  f1-score   support

        bird       0.76      0.49      0.60        53
         car       0.85      0.89      0.87        53
       cloud       0.74      0.93      0.82        54
         dog       0.87      0.70      0.78        57
      flower       0.87      0.83      0.85        66
       house       0.80      0.94      0.86        51
       human       0.72      0.78      0.75        50
    mountain       0.86      0.69      0.77        45
         sun       0.76      0.87      0.81        60
        tree       0.74      0.82      0.78        49

    accuracy                           0.80       538
   macro avg       0.80      0.79      0.79       538
weighted avg       0.80      0.80      0.79       538
```

Training History

Accuracy and Loss Analysis

The training history for the convolutional neural network (CNN) is visualized through the accuracy and loss plots for both training and validation phases.

1. Accuracy Trends:

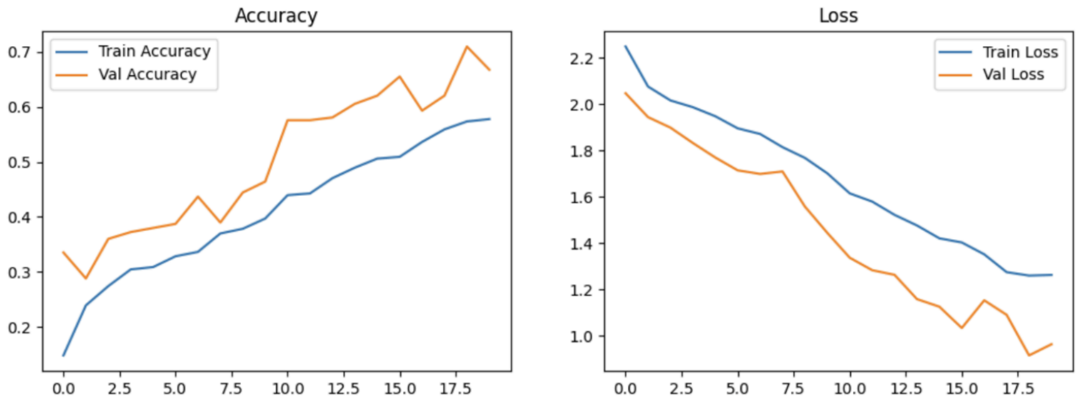
- The training accuracy shows a steady increase throughout the training epochs, indicating that the model is effectively learning from the training dataset.
- The validation accuracy also demonstrates a consistent upward trend and remains slightly higher than the training accuracy during most of the epochs. This suggests that the model generalizes well to unseen data without overfitting.

2. Loss Trends:

- The training loss decreases monotonically over the epochs, which is expected as the model minimizes the error on the training dataset.
- The validation loss also exhibits a downward trend, closely following the training loss. The close alignment of training and validation losses indicates that the model does not suffer from overfitting.

Overfitting or Underfitting Analysis

- There is no evident overfitting in the model, as the validation accuracy improves alongside the training accuracy, and the validation loss decreases concurrently with the training loss.
- Underfitting is also not observed, given the steady improvement in both accuracy and loss metrics for training and validation data.



General Evaluation

Here are the overall performance metrics for all algorithms we used in our project.

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	71.99%	0.76	0.72	0.72
Decision Tree	57.92%	0.58	0.58	0.58
Logistic Regression	90.32%	0.91	0.90	0.90
KNN	84.60%	0.86	0.85	0.85
CNN	80.00%	0.80	0.79	0.79

Comparison of Models

1. Logistic Regression: Logistic Regression outperforms other models with an accuracy of 90.32%, along with the highest precision, recall, and F1 score. This makes it the most effective model for classifying children’s drawings, especially in categories with distinct patterns such as car and house.

2. K-Nearest Neighbors (KNN): KNN achieves a solid accuracy of 84.60%, performing well in categories like house and mountain. However, it struggles slightly with less distinct classes such as dog and tree, resulting in slightly lower F1 scores for these categories.

3. Convolutional Neural Network (CNN): The CNN model shows balanced performance with an accuracy of 80.00% and comparable precision and recall values. It demonstrates good generalization capabilities but does not outperform Logistic Regression or KNN.

4. Naive Bayes: Naive Bayes achieves moderate performance with an accuracy of 71.99%. It performs well in categories like car and house but struggles with more complex or overlapping classes such as dog and bird, reflecting the limitations of its feature independence assumption.

5. Decision Tree: The Decision Tree model exhibits the lowest performance, with an accuracy of 57.92%. While it identifies some patterns effectively, it fails to capture the nuances of the dataset, leading to lower precision, recall, and F1 scores across most categories.

Conclusion

Logistic Regression emerges as the best-performing model due to its high accuracy and balanced performance across all metrics. KNN and CNN also deliver competitive results, with KNN slightly outperforming CNN in accuracy. Naive Bayes and Decision Tree serve as baseline models, demonstrating moderate to low performance. These results underscore the importance of selecting appropriate algorithms for handling the complexities of children's drawings and leveraging advanced techniques like CNNs for further improvement.

Optimization

To optimize the performance of the Convolutional Neural Network (CNN), several strategies were implemented:

1. **Data Augmentation:** Data augmentation techniques were applied to artificially increase the diversity of the training dataset. This included transformations such as flipping, rotation, zooming, and shifting to make the model more robust to variations in the input images. By creating a more diverse dataset, the model learned to generalize better, reducing the risk of overfitting to specific patterns in the training data.
2. **Bigger Image Size:** The image size was increased from smaller resolutions to 256x256 pixels, enabling the model to capture finer and more intricate details in the children's drawings. This adjustment allowed the convolutional layers to process more meaningful features, particularly for categories with complex or overlapping patterns, such as cloud and mountain. Larger images also provided the opportunity for the network to extract richer spatial hierarchies, improving classification performance.
3. **Epoch Adjustment:** The number of epochs was carefully tuned to strike a balance between underfitting and overfitting:
    - At 20 epochs, the model exhibited underfitting, where it failed to fully learn patterns from the training data, leading to suboptimal performance on both training and validation datasets.
    - At 40 epochs, the model began to overfit, as indicated by a widening gap between training and validation accuracy, alongside an increase in validation loss.
    - Setting the number of epochs to 30 provided an ideal middle ground, allowing the model to learn effectively while maintaining good generalization on unseen data. This observation was validated by consistent training and validation loss curves and aligned accuracy trends.

4. **Hyperparameter Tuning:** Additional fine-tuning of hyperparameters such as the learning rate and batch size was performed. A lower learning rate ensured stable convergence, avoiding large oscillations in the loss curve. The batch size was adjusted to balance computational efficiency with the stability of gradient updates, further improving model training dynamics.

These optimizations collectively enhanced the CNN’s ability to classify children’s drawings more effectively. The model demonstrated improved accuracy and generalization across multiple categories while mitigating risks of underfitting and overfitting. This iterative optimization process highlights the importance of careful experimentation and evaluation in developing high-performing machine learning models.