ECE180 Final project

Name

May 24, 2024

Abstract

Have a one-paragraph abstract of your project here. Let us know the task you chose and a brief summary of your project. Be concise and clear about the description.

1 Instruction

You have five tasks to choose from for the final projects. Three of them are computer vision tasks, and the other two are NLP tasks. You will decide which task to choose, and work on the project from the beginning to the end. Given the dataset in which we provide the link, (1) download the dataset, (2) apply proper preprocessing, (3) design your model, (4) train, and (5) evaluate.

You should select proper data preprocessing and model architecture. Also, many hyperparameters related to model training can affect your model performance.

Try to apply all that you learned from the lecture, and demonstrate the rationale for choosing the final data preprocessing and model design (e.g., if you don't add batch normalization, give us the reason).

When using a Colab Notebook, selecting a GPU can significantly speed up your training process. To do this, change the runtime type and choose GPU. Here's how you can do it:

1. Go to the menu and click on `Runtime`.

2. Select `Change runtime type`.

3. In the pop-up window, under `Hardware accelerator`, choose `GPU`.

Proceed with your project using **only PyTorch**. Additionally, ensure to retrieve the training history using PyTorch.

2 Tasks

Your job is to find the best data preprocessing strategy and model design to solve one of these tasks. Be careful not to use the validation and test set during training. Validation should be used for hyperparameter selections (including data preprocessing and model design choice, but not in the training loop), and the test is for the last report only.

2.1 100 Sports Image Classification

This image classification task[1] has 13,493 images with (3, 224, 224) resolution from 100 sports.

2.2 Butterfly & Moths Image Classification 100 Species

This dataset[2] contains 12,594 images with (3, 224, 224) resolution from 100 species of butterfly and moths.

---

[1] https://www.kaggle.com/datasets/gpiosenka/sports-classification/data
[2] https://www.kaggle.com/datasets/gpiosenka/butterfly-images40-species/data

2.3   30 Types of Balls Image Classification

This image classification task[3] has $35,595$ images with $(3, 224, 224)$ resolution from different 30 types of balls.

2.4   Google stock prediction

The Google stock prediction[4] dataset contains 14 columns and $1,257$ Columns. Each column is allocated to a quality, and lines contain the values for that property.

2.5   Twitter Sentiment Analysis

This dataset[5] focuses on sentiment analysis regarding issues faced by major U.S. airlines. You will initially classify tweets into positive, negative, or neutral categories. Subsequently, you identify specific reasons for negative sentiments, such as "late flight" or "rude service."

3   Organization of the report

Organize your report into sections, including an introduction, methods (data preprocessing, model design, and training strategy), and results with analysis. Your analysis of the dataset, comparison of various model hyperparameters, and a description of the final model choice are required for this report. Be as specific and clear as possible. **Lastly, have a section at the end of the report to summarize the contributions of each team member.**

4   Evalutaion

We will evaluate your project based on data preprocessing, model design, and the final performance. We want you to find the best hyperparameters (data preprocessing, model design) based on the validation set. Report the performance in comparison (that would be in the method) based on the validation set, and please provide both the validation and test performance for your final model. Performance will be evaluated based on that validation set, not the test set. Best performance is not the only metric we consider, and you should use a proper data preprocessing strategy based on the analysis of the dataset since it influences the overall performance. Please be specific and clear in your explanation about data preprocessing and model choices. Use tables and/or plots to demonstrate your results.

Rubrics:

- 30% : Quality of report
- 30% : Data Preprocessing
- 30% : Model Selection
- 10% : Reasonable Performance

5   Submission

Submit this report and your ipynb file. The ipynb file should not contain a history of your hyperparameter search, but detailed codes about your data preprocessing and model design are necessary. Performance from hyperparameter search should be contained in this report.

References

[3] https://www.kaggle.com/datasets/gpiosenka/balls-image-classification/data

[4] https://www.kaggle.com/datasets/shreenidhihipparagi/google-stock-prediction/data

[5] https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment/data