☰ TASK 4

Text regression

You are in the group 👥 Deepforgetting consisting of 👤 junterhol (junterhol@student.ethz.ch (mailto://junterhol@student.ethz.ch)), 👤 merklec (merklec@student.ethz.ch (mailto://merklec@student.ethz.ch)) and 👤 ribadov (ribadov@student.ethz.ch (mailto://ribadov@student.ethz.ch)).

📕 1. READ THE TASK DESCRIPTION

💻 2. SUBMIT SOLUTIONS

✉ 3. HAND IN FINAL SOLUTION

🔒 3. HANDIN CLOSED ON THURSDAY 23 MAY 2024 00:02

ℹ HOW TO OBTAIN POINTS

To obtain points for this task, you have to **individually** hand in the task as follows:
- You need to select one of your group's submissions for grading. You will only be graded on that submissions.
- You have to write a short report on the approach that you have used. **Each student has to individually write their own report and you are not allowed to share the report with your other group members.**

If you do not properly hand in the task, you will receive zero points for the task.

⚠ Please double check that your handin was successful by refreshing the page after pressing the hand in button!

CURRENT STATUS

✔ You have successfully handed in the task and it will be graded.

Submission selected for grading

result1.txt (test) with public score 1.1079904299308716

Report

The final task wanted us to predict the scores of product reviews from their titles and content. The "DistilBERT" model proved to be very effective for this task. We used "DistilBertTokenizer" from the Huggingface transformers library to tokenize titles and sentences. The tokenizer converts text-data into a format needed for the "DistilBERT" model, producing input IDs and attention masks. We implemented the "ReviewDataset"-class to return the tokenized inputs and available corresponding score for each data point, merging titles and the sentences into one array. We chose "DistilBERT" because it has shown to be performant yet computationally efficient. We implemented a custom "MyModel" class that served 2 purposes: 1. Set up the "DistilBERT" model and a fully connected layer. 2. Process input-IDs and attention masks through "DistilBERT" and apply a dropout layer and a fully connected layer to predict the scores. For training we set the batch size to 4. We couldn't get GPU to work and had to set "num_workers" to 0. Either the code ran out of memory or there didn't seem to be any progress after hours of computation. For each of the 5 training loop epochs we iterated over batches of data, performed forward and backward passes, computed MSE-loss and updated the model

NEW REPORT

Handin has closed on Thursday 23 May 2024 00:02. We cannot accept late handins.