Grubert - All - READMe

-----------------------

Description: This folder contains various Google Colab notebook dedicated to the Grubert model

-----------------------

Content:

- Grubert - Training.ipynb: Train the Grubert model

- Grubert Evaluation on Testset for Kaggle submission.ipyn: Generates intermediary prediction on testset set, which is then used in the ensemble, or can be uploaded to Kaggle

- Grubert Evaluation on Trainingset for Ensemble.ipyn: Generates intermediary prediction on the training set, which is then used in the least squares ensemble

- Grubert Evaluation on Validationset for Ensemble.ipyn: Generates intermediary prediction on validation set, which is then used in the ensemble

-----------------------

Prerequisites:

- Logged in with a Google account, with access to Google Colab, Google Colab's CPUs and GPUs (subscription needed to have prioritised access GPU's for a significant longer time) (for these notebooks, it is recommended to have access to GPUs)

- Granting permission to run each of our Google Colab notebooks and python files.

- Mount drive at /content/drive/ [This is done when running the notebook, see below]

- The folders/files regarding Preprocessed training and validation datasets are at the described location (see READMe for folder "data").

- Install transformers via:

command: pip install transformers

or alternativly via command: pip3 install transformers

-----------------------

Which notebook has to run first:

Grubert - Training.ipynb

-----------------------

-----------------------

Grubert - Training.ipynb

-----------------------

How to run: Grubert - Training.ipynb

Before running, the user can specify

- which preprocessed datasets they want to work with

In the first section, there is a variable

PREPROCESSING\_CHOICE defined by PREPROCESSING\_OPTIONS[INDEX], where INDEX is an integer [0;8]. The user can choose INDEX from that range, which corresponds to the preprocessed dataset.

After selecting INDEX:

It is highly recommended to have access to GPUs. In tab "Runtime", click on "change runtime type", choose GPU as hardware accelerator and save this setting.

Run all sections, top to bottom.

-----------------------

Output:

In the directory /content/drive/CIL22022, the trained models are stored as file (for each epoch) and the performance of model is logged during training.

-----------------------

-----------------------

Grubert Evaluation on Validationset for Ensemble.ipyn:

-----------------------

How to run: Grubert Evaluation on Validationset for Ensemble.ipyn

Before running, the user can specify

- which preprocessed datasets they want to work with

In the first section, there is a variable

PREPROCESSING\_CHOICE defined by PREPROCESSING\_OPTIONS[INDEX], where INDEX is an integer [0;8]. The user can choose INDEX from that range, which corresponds to the preprocessed dataset.

In the last section, the user must specify the path (variable "path\_to\_model") to the model they want to use to generate an intermediary prediction on the validation set.

After speciyfing the variables in question:

It is highly recommended to have access to GPUs. In tab "Runtime", click on "change runtime type", choose GPU as hardware accelerator and save this setting.

Run all sections, top to bottom.

-----------------------

Output:

In the directory /content/drive/CIL22022/data/test data/, the intermediary prediction is stored as csv file.

-----------------------

-----------------------

Grubert Evaluation on Testset for Kaggle submission.ipyn:

- Same instructions as for the above notebook "Grubert Evaluation on Validationset for Ensemble.ipyn:"

-----------------------

-----------------------

Grubert Evaluation on Trainingset for Ensemble.ipyn:

- Same instructions as for the above notebook "Grubert Evaluation on Validationset for Ensemble.ipyn:"