EE-559 Deep learning – Mini-Project, Students' questions

A question is denoted by **Q**; the corresponding answer is denoted by **A**.

Q: What do we use for the project in terms of computational resources?
A: For the project, you will be given access to high performance computing (HPC) cluster at EPFL. In a nutshell, you will be sending a *job* which will contain the information about your *code*. This *job* will be held in a high-priority *queue* to be run on a supercomputer. More details about how to access the HPC cluster, and how to submit your first job, will be provided later in the course.

Q: Which modality/modalities is the best modality to be used for the project?
A: You can choose based on your interests. Come and talk to us during the lab slots if you'd like to discuss this choice.

Q: Can we use datasets that are not from the initial list?
A: Yes, here is a link to additional datasets (also listed at the end of the initial list), but you can also search for other options. Just make sure that the license associated with any dataset you choose permits its use in your project.

Q: Can we use the "Hateful Memes Dataset" or "Russia-Ukraine war multimodal" dataset?
A: This dataset was available for the Hateful Memes Challenge, however, it is now not available via the official repository. While someone uploaded the dataset to Kaggle, the license there states "unknown". By default, anything that is uploaded or shared online is protected by copyright, which means that we cannot use or redistribute that data.

In the case of the latter dataset, note that while the initial repository is released under the MIT license, it does not contain the dataset. This repository contains a link to another repository, which does not have any license information, which prevents us from using this dataset. In this case, you could consider reaching out to the authors to request that they add license information to their GitHub repository. However, keep in mind that there is no guarantee they will respond.

Q: Can we use models that rely on methods that are not covered in the course (such as LSTMs) ?
A: Yes, that's allowed.

Q: Will we have the deadline for the project in the middle of the term?
A: No. Please, refer to details about the project deadlines in Week 2 lecture slides, p.28.

Q: Can I train BERT from scratch for classification?
A: You should assess whether you have sufficient data to effectively train the model. BERT was originally pretrained on a 3.3 billion word corpus. Training the BERT model from scratch also requires significant computational resources and time. You might want to consider smaller versions of BERT, which require less data and computational resources for training. For examples, see DistilBERT, BERT miniatures, and ALBERT.

To train BERT from scratch for classification you can initialize a model with random weights with HuggingFace using `config` instead of `.from_pretrained`:

```
from transformers import BertConfig, BertForSequenceClassification
# either load pre-trained config
config = BertConfig.from_pretrained("bert-base-cased")
# or instantiate yourself
config = BertConfig(
    vocab_size=2048,
    max_position_embeddings=768,
    intermediate_size=2048,
    hidden_size=512,
    num_attention_heads=8,
    num_hidden_layers=6,
    type_vocab_size=5,
    hidden_dropout_prob=0.1,
    attention_probs_dropout_prob=0.1,
    num_labels=3,
)

# pass the config to model constructor instead of from_pretrained
# this creates the model as per the params in config
# but with weights randomly initialized
model = BertForSequenceClassification(config)
```