AndriyMulyar / **bert_document_classification**

architectures and pre-trained models for long document classification.

29 commits	2 branches	0 packages	1 release	2 contributors

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

AndriyMulyar Merge pull request #11 from sjmielke/patch-1 ··· Latest commit `e9d9cd4` on 28 Apr

bert_document_classification	Add freezing methods to linear classifier to unify interface	last month
examples	Removed additional freezing references in examples.	2 months ago
.gitignore	began code for replication of actual paper experiments	8 months ago
MANIFEST.in	Initial package set-up no testing done yet	8 months ago

README.md	Update README.md	7 months ago
setup.py	Initial package set-up no testing done yet	8 months ago

README.md

📖 BERT Long Document Classification 📖

an easy-to-use interface to fully trained BERT based models for multi-class and multi-label long document classification.

pre-trained models are currently available for two clinical note (EHR) phenotyping tasks: smoker identification and obesity detection.

To sustain future development and improvements, we interface pytorch-transformers for all language model components of our architectures. Additionally, their is a blog post describing the architecture.

Model	Dataset	# Labels	Evaluation F1
n2c2_2006_smoker_lstm	I2B2 2006: Smoker Identification	4	0.981
n2c2_2008_obesity_lstm	I2B2 2008: Obesity and Co-morbidities Identification	15	0.997

Installation

Install with pip:

```
pip install bert_document_classification
```

or directly:

```
pip install git+https://github.com/AndriyMulyar/bert_document_classification
```

Use

Maps text documents of arbitrary length to binary vectors indicating labels.

```
from bert_document_classification.models import SmokerPhenotypingBert
from bert_document_classification.models import ObesityPhenotypingBert

smoking_classifier = SmokerPhenotypingBert(device='cuda', batch_size=10) #defaults to GPU prediction

obesity_classifier = ObesityPhenotypingBert(device='cpu', batch_size=10) #or CPU if you would like.

smoking_classifier.predict(["I'm a document! Make me long and the model can still perform well!"])
```

More examples.

Replication

Go to the directory /examples/ml4health_2019_replication. This README will give instructions on how to appropriately insert data from DBMI to replicate the results in the paper.

Notes

- For training you will need a GPU.

- For bulk inference where speed is not of concern lots of available memory and CPU cores will likely work.
- Model downloads are cached in `~/.cache/torch/bert_document_classification/`. Try clearing this folder if you have issues.

Acknowledgement

If you found this project useful, consider citing our extended abstract.

```
@misc{mulyar2019phenotyping,
    title={Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized
Neural Language Models},
    author={Andriy Mulyar and Elliot Schumacher and Masoud Rouhizadeh and Mark Dredze},
    year={2019},
    eprint={1910.13664},
    archivePrefix={arXiv},
    primaryClass={cs.CL}
}
```

Implementation, development and training in this project were supported by funding from the Mark Dredze Lab at Johns Hopkins University.