BioTranslator

BioTranslator team

Apr 27, 2023

CONTENTS

- Get started by browsing *Tutorial* or the main *API*.
- Follow updates in the release notes.

CHAPTER

ONE

NEWS

1.1 BioTranslator API has been published on PyPI 2022-08-26

(past news)

CHAPTER

TWO

LATEST ADDITIONS

CHAPTER

THREE

RELEASE NOTES

3.1 0.1.0 2022-08-26

- Released basic functions of BioTranslator.

3.2 0.1.1 2022-09-04

- BioTranslator allows Protein Sequence Prediction, Cell Type Classification, and Pathway Analysis.

3.2.1 API

Import methods from BioTranslator

```
from biotranslator import .
```

Setup a config:

- `setup_config`(config, data_type='seq')

Train text encoder:

- `train_text_encoder`(data_dir: str, save_path: str)

Train a BioTranslator Model:

- `train_biotranslator`(cfgs)

Test a BioTranslator Model:

- `test_biotranslator`(data_dir, anno_data, cfg, translator, task)

3.2.2 News

BioTranslator documentation has been released 2022-09-04

BioTranslator API has been published on PyPI 2022-08-26

3.2.3 Tutorial

Import BioTranslator as

```
import biotranslator as bt
```

Import methods from API

```
from biotranslator.biotranslator_function import setup_config, train_text_encoder, train_
↪biotranslator, \
    test_biotranslator
```

Train Text Encoder

```
model_path = './TextEncoder/model/encoder.pth'
graphine_repo = './TextEncoder/data/Graphine/dataset/'
train_text_encoder(graphine_repo, model_path)
```

Build Configs

Build config for protein sequence dataset:

```
seq_repo = f'./Protein'
seq_config = {
    'task': 'few_shot',
    'max_length': 2000,
    'data_repo': f'{seq_repo}/data/',
    'dataset': 'GOA_Human',
    'encoder_path': './Codebase/TextEncoder/Encoder/text_encoder.pth',
    'rst_dir': f'{seq_repo}/results/',
    'emb_dir': f'{seq_repo}/embeddings/',
    'ws_dir': f'{seq_repo}/working_space/',
    'hidden_dim': 1500,
    'features': 'seqs, description, network',
    'lr': 0.0003,
    'epoch': 30,
    'batch_size': 32,
```

(continues on next page)

(continued from previous page)

```
    'gpu_ids': '0',
}

seq_config = setup_config(seq_config, 'seq')
```

Build config for cell description vector dataset:

```
vec_repo = f'./SingleCell'
vec_config = {
    'task': 'cross_dataset',
    'eval_dataset': 'muris_facs',
    'vec_ontology_repo': f'{vec_repo}/data/Ontology_data/',
    'data_repo': f'{vec_repo}/data/sc_data/',
    'dataset': 'muris_droplet',
    'encoder_path': './Codebase/TextEncoder/Encoder/text_encoder.pth',
    'rst_dir': f'{vec_repo}/results/',
    'emb_dir': f'{vec_repo}/embeddings/',
    'ws_dir': f'{vec_repo}/working_space/',
    'hidden_dim': 30,
    'lr': 0.0001,
    'epoch': 15,
    'batch_size': 128,
    'gpu_ids': '0',
}

vec_config = setup_config(vec_config, 'vec')
```

Build config for pathway graph dataset:

```
graph_repo = f'./Pathway'
graph_config = {
    'max_length': 2000,
    'eval_dataset': 'KEGG',
    'graph_excludes': ['Reactome', 'KEGG', 'PharmGKB'],
    'data_repo': f'{graph_repo}/data/',
    'dataset': 'GOA_Human',
    'encoder_path': './Codebase/TextEncoder/Encoder/text_encoder.pth',
    'rst_dir': f'{graph_repo}/results/',
    'emb_dir': f'{graph_repo}/embeddings/',
    'ws_dir': f'{graph_repo}/working_space/',
    'hidden_dim': 1500,
    'features': 'seqs, description, network',
    'lr': 0.0003,
    'epoch': 30,
    'batch_size': 32,
    'gpu_ids': '0',
}

graph_config = setup_config(graph_config, 'graph')
```

Train BioTranlators

```
cfgs = [seq_config, vec_config, graph_config]
translators = train_biotranslator(cfgs)
```

Test BioTranslators

```
tasks = dict(
        seq=['prot_func_pred'],
        vec=['cell_type_cls'],
        graph=['node_cls', 'edge_pred'])
    vec_files = BioLoader(vec_config)
    anno_data = dict(
        seq=[pd.read_pickle(f'{seq_config.data_repo}{seq_config.dataset}/validation_data_
↪fold_0.pkl')],
        vec=[vec_files.test_data],
        graph=[pd.read_pickle(f'{graph_config.data_repo}{graph_config.eval_dataset}/
↪pathway_dataset.pkl'),
               pd.read_pickle(f'{graph_config.data_repo}{graph_config.eval_dataset}/
↪pathway_dataset.pkl')],
    )
    for tp_idx, tp in enumerate(list(tasks.keys())):
        for task_idx in range(len(tasks[tp])):
            cfg = cfgs[tp_idx]
            encoder = translators[tp_idx]
            annos = test_biotranslator(cfg.data_repo, anno_data[tp][task_idx], cfg,␣
↪encoder, tasks[tp][task_idx])
    print(annos)
```

3.2.4 Installation

PyPI Version

Install latest BioTranslator from PyPI (consider using `pip3` to access Python 3):

```
pip install biotranslator
```

Development Version

To work with the latest version on GitHub: clone the repository and `cd` into its root directory.

Install with HTTPS:

```
https://github.com/ywzhao2002/biotranslator.git
cd biotranslator
```

Install with Github CLI:

```
gh repo clone ywzhao2002/biotranslator
cd biotranslator
```

3.2.5 Setup Dataset

Processed datasets including *CAFA3*, *GOA_Human*, *GOA_Mouse*, *GOA_Yeast*, *KEGG*, *PharmGKB*, *Reactome*, and *Swissprot* are available at https://figshare.com/articles/dataset/Protein_Pathway_data_tar/20120447

Processed datasets including *Tabula_Microcebus* and *Tabula_Sapiens* can be found at: https://figshare.com/ndownloader/files/31777475 and https://figshare.com/ndownloader/files/28846647 . Remaining datasets can be found from OnClass package.

Graphine dataset used for training text encoder can be downloaded from https://zenodo.org/record/5320310/files/Graphine.zip?download=1.

Example dataset structure for protein sequence prediction and pathway analysis tasks

```
├── data
│   ├── CAFA3
│   ├── GOA_Human
│   ├── GOA_Mouse
│   ├── GOA_Yeast
│   ├── KEGG
│   ├── PharmGKB
│   ├── Reactome
│   └── SwissProt
```

Example dataset structure for single cell classification task

```
├── data
│   ├── ont_data
│   │   ├── allen.ontology
│   │   ├── cl.obo
│   │   ├── cl.ontology
│   │   └── cl.ontology.nlp.emb
│   ├── sc_data
│   │   ├── 26-datasets
│   │   │   ├── 293t_jurkat
│   │   │   ├── brain
│   │   │   ├── hsc
│   │   │   ├── macrophage
```

(continues on next page)

(continued from previous page)

```
│   │   │   ├── pancreas
│   │   │   └── pbmc
│   │   ├── Allen_Brain
│   │   │   ├── features.pkl
│   │   │   ├── genes.pkl
│   │   │   └── labels.pkl
│   │   ├── gene_marker_expert_curated.txt
│   │   ├── HLCA
│   │   │   ├── 10x_features.pkl
│   │   │   ├── 10x_genes.pkl
│   │   │   └── 10x_labels.pkl
│   │   ├── Lemur
│   │   │   ├── microcebusAntoine.h5ad
│   │   │   ├── microcebusBernard.h5ad
│   │   │   ├── microcebusMartine.h5ad
│   │   │   └── microcebusStumpy.h5ad
│   │   ├── Tabula_Microcebus
│   │   │   └── LCA_complete_wRaw_toPublish.h5ad
│   │   ├── Tabula_Muris_Senis
│   │   │   ├── tabula-muris-senis-droplet-official-raw-obj.h5ad
│   │   │   └── tabula-muris-senis-facs-official-raw-obj.h5ad
│   │   └── Tabula_Sapiens
│   │   │   └── TabulaSapiens.h5ad
```

Example dataset structure for text encoder

```
├── data
│   ├── Graphine
│   │   └── dataset
```

3.2.6 Release Notes

0.1.0 2022-08-26

- Released basic functions of BioTranslator.

0.1.1 2022-09-04

- BioTranslator allows Protein Sequence Prediction, Cell Type Classification, and Pathway Analysis.