TEMPLATE

October 25, 2025

[9]:
```
%reload_ext autoreload
%autoreload 2
```

[10]:
```
from kret_studies import *
from kret_studies.notebook import *
from kret_studies.complex import *

logger = get_notebook_logger()
```

```
/Users/Akseldkw/coding/kretsinger/data/nb_log.log
```

[11]:
```
from uml_project import *

HF_DIR, HF_REGISTRY, DEVICE_TORCH, MODEL_DIR
```

[11]:
```
(PosixPath('/Users/Akseldkw/coding/Columbia/UML-Project/data/huggingface'),
 PosixPath('/Users/Akseldkw/coding/Columbia/UML-
Project/data/huggingface/REGISTRY.json'),
 device(type='mps'),
 PosixPath('/Users/Akseldkw/coding/Columbia/UML-Project/data/models'))
```

[12]:
```
IMDB_DIR = HF_DIR / "imdb"
```

[13]:
```
df_imdb_train = pd.read_parquet(IMDB_DIR / "train.parquet")
df_imdb_test = pd.read_parquet(IMDB_DIR / "test.parquet")
```

[14]:
```
df = df_imdb_train
test_df = df_imdb_test
```

[15]:
```
stsb_dict = load_dataset("glue", "stsb")
```

```
Generating train split:   0%|          | 0/5749 [00:00<?, ? examples/s]
```

```
Generating validation split:   0%|          | 0/1500 [00:00<?, ? examples/s]
```

```
Generating test split:   0%|          | 0/1379 [00:00<?, ? examples/s]
```

[16]:
```
df_stsb_train: pd.DataFrame = stsb_dict["train"].to_pandas()  # type: ignore
df_stsb_val: pd.DataFrame = stsb_dict["validation"].to_pandas()  # type: ignore
df_stsb_test: pd.DataFrame = stsb_dict["test"].to_pandas()  # type: ignore
```

[17]:
```
# df_stsb_train.sort_values("label", ascending=False)
# df_stsb_val.sort_values("label", ascending=False)
# df_stsb_test.sort_values("label", ascending=False)
```

[18]:
```
datasets = list(huggingface_hub.list_datasets(dataset_name="stsb"))
```

[19]:
```
word_emb = models.Transformer("bert-base-uncased")
pooling = models.Pooling(word_emb.get_word_embedding_dimension(),␣
 ↪pooling_mode_mean_tokens=True)
dense = models.Dense(
    in_features=word_emb.get_word_embedding_dimension(), out_features=128,␣
 ↪activation_function=torch.nn.Tanh()
)

model = SentenceTransformer(modules=[word_emb, pooling, dense])
```

[20]:
```
BASE_MODEL = "sentence-transformers/all-MiniLM-L6-v2"  # ABOBA: small, fast
TARGET_DIM = 64  # ABOBA desired embedding dimensionality (experiment with 32,
# 64, 128...)
BATCH_SIZE = 64
POOLER_LR = 2e-4
FINETUNE_LR = 2e-5
```

[21]:
```
EPOCHS_POOLER = 2  # step A epochs (pooler only)
EPOCHS_FINETUNE = 2  # step B epochs (unfreeze and train)
```

[22]:
```
nb_vars = uml_utils.NotebookVars(
    {
        "DEVICE": DEVICE_TORCH_STR,
        "BATCH_SIZE": BATCH_SIZE,
        "POOLER_LR": POOLER_LR,
        "FINETUNE_LR": FINETUNE_LR,
        "EPOCHS_POOLER": EPOCHS_POOLER,
        "EPOCHS_FINETUNE": EPOCHS_FINETUNE,
    }
)
```

[23]:
```
s_model = uml_sentence.build_model(BASE_MODEL, TARGET_DIM, DEVICE_TORCH_STR)
```

[24]:
```
s_model
```

[24]:
```
SentenceTransformer(
  (0): Transformer({'max_seq_length': 128, 'do_lower_case': False,
'architecture': 'BertModel'})
  (1): Pooling({'word_embedding_dimension': 384, 'pooling_mode_cls_token':
False, 'pooling_mode_mean_tokens': True, 'pooling_mode_max_tokens': False,
'pooling_mode_mean_sqrt_len_tokens': False, 'pooling_mode_weightedmean_tokens':
```

```
False, 'pooling_mode_lasttoken': False, 'include_prompt': True})
  (2): Dense({'in_features': 384, 'out_features': 64, 'bias': True,
'activation_function': 'torch.nn.modules.activation.Tanh'})
)
```

[ ]:

[27]:
```
stsb_vals = uml_utils.df_to_input_examples(df_stsb_val)
```

[28]:
```
FINETUNE = MODEL_DIR / "sentence_transformer_finetuned"
uml_sentence.train_pooler_then_finetune(s_model, df_stsb_train, stsb_vals,␣
 ↪out_dir=FINETUNE, notebook_vars=nb_vars)
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
File ~/micromamba/envs/kret_312/lib/python3.12/site-packages/pandas/core/indexe /
 ↪base.py:3812, in Index.get_loc(self, key)
   3811 try:
-> 3812     return self._engine.get_loc(casted_key)
   3813 except KeyError as err:

File pandas/_libs/index.pyx:167, in pandas._libs.index.IndexEngine.get_loc()

File pandas/_libs/index.pyx:196, in pandas._libs.index.IndexEngine.get_loc()

File pandas/_libs/hashtable_class_helper.pxi:7088, in pandas._libs.hashtable.
 ↪PyObjectHashTable.get_item()

File pandas/_libs/hashtable_class_helper.pxi:7096, in pandas._libs.hashtable.
 ↪PyObjectHashTable.get_item()

KeyError: 1890

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
Cell In[28], line 2
      1 FINETUNE = MODEL_DIR / "sentence_transformer_finetuned"
----> 2␣
 ↪uml_sentence.train_pooler_then_finetune(s_model, df_stsb_train, stsb_vals, ou _dir=FINETUN

File ~/coding/Columbia/UML-Project/uml_project/uml_models/sentence_trnsf.py:9,␣
 ↪in train_pooler_then_finetune(model, train_examples, val_examples, out_dir,␣
 ↪notebook_vars)
      0 <Error retrieving source code with stack_data see ipython/ipython#13598
```

```
File ~/micromamba/envs/kret_312/lib/python3.12/site-packages/
 ↪sentence_transformers/fit_mixin.py:274, in FitMixin.fit(self,␣
 ↪train_objectives, evaluator, epochs, steps_per_epoch, scheduler, warmup_steps ␣
 ↪optimizer_class, optimizer_params, weight_decay, evaluation_steps,␣
 ↪output_path, save_best_model, max_grad_norm, use_amp, callback,␣
 ↪show_progress_bar, checkpoint_path, checkpoint_save_steps,␣
 ↪checkpoint_save_total_limit, resume_from_checkpoint)
    272 texts = []
    273 labels = []
--> 274 for batch in data_loader:
    275␣
 ↪    batch_texts, batch_labels = zip(*[(example.texts, example.label) for exam le in batch]
    276     texts += batch_texts

File ~/micromamba/envs/kret_312/lib/python3.12/site-packages/torch/utils/data/
 ↪dataloader.py:733, in _BaseDataLoaderIter.__next__(self)
    730 if self._sampler_iter is None:
    731     # TODO(https://github.com/pytorch/pytorch/issues/76750)
    732     self._reset()  # type: ignore[call-arg]
--> 733 data = self._next_data()
    734 self._num_yielded += 1
    735 if (
    736     self._dataset_kind == _DatasetKind.Iterable
    737     and self._IterableDataset_len_called is not None
    738     and self._num_yielded > self._IterableDataset_len_called
    739 ):

File ~/micromamba/envs/kret_312/lib/python3.12/site-packages/torch/utils/data/
 ↪dataloader.py:789, in _SingleProcessDataLoaderIter._next_data(self)
    787 def _next_data(self):
    788     index = self._next_index()  # may raise StopIteration
--> 789     data = self._dataset_fetcher.fetch(index)  # may raise StopIteratio
    790     if self._pin_memory:
    791         data = _utils.pin_memory.pin_memory(data, self.
 ↪_pin_memory_device)

File ~/micromamba/envs/kret_312/lib/python3.12/site-packages/torch/utils/data/
 ↪_utils/fetch.py:52, in _MapDatasetFetcher.fetch(self, possibly_batched_index)
     50         data = self.dataset.__getitems__(possibly_batched_index)
     51     else:
---> 52         data = [self.dataset[idx] for idx in possibly_batched_index]
     53 else:
     54     data = self.dataset[possibly_batched_index]

File ~/micromamba/envs/kret_312/lib/python3.12/site-packages/pandas/core/frame.
 ↪py:4113, in DataFrame.__getitem__(self, key)
   4111 if self.columns.nlevels > 1:
   4112     return self._getitem_multilevel(key)
-> 4113 indexer = self.columns.get_loc(key)
```

```
   4114 if is_integer(indexer):
   4115     indexer = [indexer]

File ~/micromamba/envs/kret_312/lib/python3.12/site-packages/pandas/core/indexes/
 ↪base.py:3819, in Index.get_loc(self, key)
   3814     if isinstance(casted_key, slice) or (
   3815         isinstance(casted_key, abc.Iterable)
   3816         and any(isinstance(x, slice) for x in casted_key)
   3817     ):
   3818         raise InvalidIndexError(key)
-> 3819     raise KeyError(key) from err
   3820 except TypeError:
   3821     # If we have a listlike key, _check_indexing_error will raise
   3822     #  InvalidIndexError. Otherwise we fall through and re-raise
   3823     #  the TypeError.
   3824     self._check_indexing_error(key)

KeyError: 1890
```

[31]:
```
len(stsb_vals), df_stsb_train.shape
```

[31]:
```
(1500, (5749, 4))
```

[ ]:
```

```