HW: Sentece contrastive learning

This homework is about learning sentence representation and contrastive learning.

From previous homework, we used to build token/sequence classification task and learn it through only supervised method. In real-world scenario, **human annotation** requires a lot of cost and effort to do. Some annotation tasks might require domain experts such as medical domain, legal domain, etc. However, there are some **unsupervised** methods which are no need any annotations.

Contrastive learning is the popular one of unsupervised learning approach. It will learn the representation via similar and dissimilar examples.

For this homework, we will focus on **SimCSE** framework which is one of contrastive learning techniques. For SimCSE, it will learn sentence embedding by comparing between different views of the same sentence.

In this homework you will perform three main tasks.

1. Train a sentiment classification model using a pretrained model. This model uses freeze weights. That is it treats the pretrained model as a fixed feature extractor.
2. Train a sentiment classification model using a pretrained model. This model also performs weight updates on the base model's weights.
3. Perform SimCSE and use the sentence embedding to perform linear classification.

Install and import libraries

Install the `datasets` library under Huggingface and `Pytorch lightning` framework.

In [1]:
```
!pip install datasets pytorch-lightning scikit-learn
```

```
Requirement already satisfied: datasets in /usr/local/lib/python3.10/dist-package
s (3.2.0)
Requirement already satisfied: pytorch-lightning in /usr/local/lib/python3.10/dis
t-packages (2.5.0.post0)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-pac
kages (1.2.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-package
s (from datasets) (3.17.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-pack
ages (from datasets) (1.26.4)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.10/dist-
packages (from datasets) (19.0.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.10/di
st-packages (from datasets) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages
(from datasets) (2.2.3)
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.10/dist
-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.10/dist-pac
kages (from datasets) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.10/dist-packages
(from datasets) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in /usr/local/lib/python3.10/
dist-packages (from datasets) (0.70.16)
Requirement already satisfied: fsspec<=2024.9.0,>=2023.1.0 in /usr/local/lib/pyth
on3.10/dist-packages (from fsspec[http]<=2024.9.0,>=2023.1.0->datasets) (2024.9.
0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-packages
(from datasets) (3.11.11)
Requirement already satisfied: huggingface-hub>=0.23.0 in /usr/local/lib/python3.
10/dist-packages (from datasets) (0.28.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packag
es (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-pack
ages (from datasets) (6.0.2)
Requirement already satisfied: torch>=2.1.0 in /usr/local/lib/python3.10/dist-pac
kages (from pytorch-lightning) (2.5.1+cu121)
Requirement already satisfied: torchmetrics>=0.7.0 in /usr/local/lib/python3.10/d
ist-packages (from pytorch-lightning) (1.6.1)
Requirement already satisfied: typing-extensions>=4.4.0 in /usr/local/lib/python
3.10/dist-packages (from pytorch-lightning) (4.12.2)
Requirement already satisfied: lightning-utilities>=0.10.0 in /usr/local/lib/pyth
on3.10/dist-packages (from pytorch-lightning) (0.12.0)
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-pac
kages (from scikit-learn) (1.13.1)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-pa
ckages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/
dist-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.
10/dist-packages (from aiohttp->datasets) (2.4.4)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist
-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: async-timeout<6.0,>=4.0 in /usr/local/lib/python3.
10/dist-packages (from aiohttp->datasets) (5.0.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-pa
ckages (from aiohttp->datasets) (25.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dis
t-packages (from aiohttp->datasets) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.10/d
```

```
ist-packages (from aiohttp->datasets) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.10/dist
-packages (from aiohttp->datasets) (0.2.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.10/dis
t-packages (from aiohttp->datasets) (1.18.3)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packa
ges (from lightning-utilities>=0.10.0->pytorch-lightning) (75.1.0)
Requirement already satisfied: mkl_fft in /usr/local/lib/python3.10/dist-packages
(from numpy>=1.17->datasets) (1.3.8)
Requirement already satisfied: mkl_random in /usr/local/lib/python3.10/dist-packa
ges (from numpy>=1.17->datasets) (1.2.4)
Requirement already satisfied: mkl_umath in /usr/local/lib/python3.10/dist-packag
es (from numpy>=1.17->datasets) (0.1.1)
Requirement already satisfied: mkl in /usr/local/lib/python3.10/dist-packages (fr
om numpy>=1.17->datasets) (2025.0.1)
Requirement already satisfied: tbb4py in /usr/local/lib/python3.10/dist-packages
(from numpy>=1.17->datasets) (2022.0.0)
Requirement already satisfied: mkl-service in /usr/local/lib/python3.10/dist-pack
ages (from numpy>=1.17->datasets) (2.4.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python
3.10/dist-packages (from requests>=2.32.2->datasets) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-pac
kages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/di
st-packages (from requests>=2.32.2->datasets) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/di
st-packages (from requests>=2.32.2->datasets) (2025.1.31)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-package
s (from torch>=2.1.0->pytorch-lightning) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
(from torch>=2.1.0->pytorch-lightning) (3.1.4)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.10/dist-pa
ckages (from torch>=2.1.0->pytorch-lightning) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.10/di
st-packages (from sympy==1.13.1->torch>=2.1.0->pytorch-lightning) (1.3.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.1
0/dist-packages (from pandas->datasets) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-pac
kages (from pandas->datasets) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-p
ackages (from pandas->datasets) (2025.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-package
s (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-
packages (from jinja2->torch>=2.1.0->pytorch-lightning) (3.0.2)
Requirement already satisfied: intel-openmp>=2024 in /usr/local/lib/python3.10/di
st-packages (from mkl->numpy>=1.17->datasets) (2024.2.0)
Requirement already satisfied: tbb==2022.* in /usr/local/lib/python3.10/dist-pack
ages (from mkl->numpy>=1.17->datasets) (2022.0.0)
Requirement already satisfied: tcmlib==1.* in /usr/local/lib/python3.10/dist-pack
ages (from tbb==2022.*->mkl->numpy>=1.17->datasets) (1.2.0)
Requirement already satisfied: intel-cmplr-lib-rt in /usr/local/lib/python3.10/di
st-packages (from mkl_umath->numpy>=1.17->datasets) (2024.2.0)
Requirement already satisfied: intel-cmplr-lib-ur==2024.2.0 in /usr/local/lib/pyt
hon3.10/dist-packages (from intel-openmp>=2024->mkl->numpy>=1.17->datasets) (202
4.2.0)
```

In [2]:

```
import torch
from torch import nn
import torch.nn.functional as F
```

```python
from transformers import (
    AutoTokenizer, AutoModelForSequenceClassification, AutoModel
)
from datasets import load_dataset
import pytorch_lightning as pl
from pytorch_lightning import LightningModule, Trainer
from torch.utils.data import DataLoader
from torchmetrics import Accuracy

import numpy as np
import matplotlib.pyplot as plt
from sklearn.manifold import TSNE
```

Setup

The dataset we use for this homework is **Wisesight-Sentiment** (huggingface, github) dataset. It is a Thai social media dataset which are labeled as **4 classes** e.g. positive, negative, neutral, and question. Furthermore, It contains both Thai, English, Emoji, and etc. That is why we choose the distilled version of multilingual BERT (mBERT) DistilledBERT paper to be a base model.

In [3]:
```python
model_name = 'distilbert-base-multilingual-cased'
dataset = load_dataset('pythainlp/wisesight_sentiment')

# Load tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name) # Or a Thai-specific token
```

```
README.md:   0%|          | 0.00/12.1k [00:00<?, ?B/s]
train-00000-of-00001.parquet:   0%|          | 0.00/2.58M [00:00<?, ?B/s]
validation-00000-of-00001.parquet:   0%|          | 0.00/286k [00:00<?, ?B/s]
test-00000-of-00001.parquet:   0%|          | 0.00/327k [00:00<?, ?B/s]
Generating train split:   0%|          | 0/21628 [00:00<?, ? examples/s]
Generating validation split:   0%|          | 0/2404 [00:00<?, ? examples/s]
Generating test split:   0%|          | 0/2671 [00:00<?, ? examples/s]
tokenizer_config.json:   0%|          | 0.00/49.0 [00:00<?, ?B/s]
config.json:   0%|          | 0.00/466 [00:00<?, ?B/s]
vocab.txt:   0%|          | 0.00/996k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/1.96M [00:00<?, ?B/s]
```

Loading Dataset and DataLoader

Preprocessing step

In [4]:
```python
# Preprocessing function
def preprocess_function(examples):
    return tokenizer(examples['texts'], padding='max_length', truncation=True)

# Apply preprocessing
encoded_dataset = dataset.map(preprocess_function, batched=True)
print(encoded_dataset)
# Change `category` key to `labels`
encoded_dataset = encoded_dataset.map(lambda examples: {'labels': [label for lab
encoded_dataset
```

```
Map:   0%|          | 0/21628 [00:00<?, ? examples/s]
```

```
Map:   0%|          | 0/2404 [00:00<?, ? examples/s]
Map:   0%|          | 0/2671 [00:00<?, ? examples/s]
DatasetDict({
    train: Dataset({
        features: ['texts', 'category', 'input_ids', 'attention_mask'],
        num_rows: 21628
    })
    validation: Dataset({
        features: ['texts', 'category', 'input_ids', 'attention_mask'],
        num_rows: 2404
    })
    test: Dataset({
        features: ['texts', 'category', 'input_ids', 'attention_mask'],
        num_rows: 2671
    })
})
Map:   0%|          | 0/21628 [00:00<?, ? examples/s]
Map:   0%|          | 0/2404 [00:00<?, ? examples/s]
Map:   0%|          | 0/2671 [00:00<?, ? examples/s]
```

Out[4]:

```
DatasetDict({
    train: Dataset({
        features: ['texts', 'category', 'input_ids', 'attention_mask', 'label
s'],
        num_rows: 21628
    })
    validation: Dataset({
        features: ['texts', 'category', 'input_ids', 'attention_mask', 'label
s'],
        num_rows: 2404
    })
    test: Dataset({
        features: ['texts', 'category', 'input_ids', 'attention_mask', 'label
s'],
        num_rows: 2671
    })
})
```

Define Dataset class

In [5]:

```python
# Create PyTorch Dataset
class SentimentDataset(torch.utils.data.Dataset):
    def __init__(self, encodings, labels):
        self.encodings = encodings
        self.labels = labels

    def __getitem__(self, idx):
        item = {
            key: torch.tensor(val) for key, val in self.encodings[idx].items()
            if key in ['input_ids', 'attention_mask']
        }
        item['labels'] = torch.tensor(self.labels[idx])
        return item

    def __len__(self):
        return len(self.labels)
```

Declare Dataset and DataLoader

In [6]:
```python
# Create Dataset object from DataFrame
train_dataset = SentimentDataset(encoded_dataset['train'], encoded_dataset['trai
val_dataset = SentimentDataset(encoded_dataset['validation'], encoded_dataset['v
test_dataset = SentimentDataset(encoded_dataset['test'], encoded_dataset['test']

# Create dataloaders
train_loader = DataLoader(train_dataset, batch_size=32, shuffle=True)
val_loader = DataLoader(val_dataset, batch_size=32)
test_loader = DataLoader(test_dataset, batch_size=32)
```

Define base model classes

Here we define model classes which will be used in the next sections.

Base Model class

`BaseModel` is a parent class for building other models e.g.

- `Pretrained LM with a linear classifier`
- `Fine-tuned LM with a linear classifier`
- `Contrastive learning based (SimCSE) LM with a linear classifier`

In [7]:
```python
class BaseModel(LightningModule):
    def __init__(
            self,
            model_name: str = 'distilbert-base-multilingual-cased',
            learning_rate: float = 2e-5
    ):
        super().__init__()
        self.save_hyperparameters()

        self.encoder = AutoModel.from_pretrained(model_name)
        self.learning_rate = learning_rate

    def get_embeddings(self, input_ids, attention_mask):
        # TODO 1: get CLS token embedding to represent as a sentence embedding
        encoded = self.encoder(input_ids=input_ids, attention_mask=attention_mas
        cls_embedding = encoded.last_hidden_state[:, 0, :]
        return cls_embedding

    def configure_optimizers(self):
        optimizer = torch.optim.AdamW(self.parameters(), lr=self.learning_rate)
        return optimizer

    def forward(self, input_ids, attention_mask):
        return self.get_embeddings(input_ids, attention_mask)
```

LMWithLinearClassifier class

`LMWithLinearClassifier` class is designed to update both LM's parameters in the supervised approach and a linear layer's parameters.

`LMWithLinearClassfier` consists of

1. `ckpt_path` (checkpoint path) refers to the best checkpoint after training SimCSE method. We will load the encoder's weights from the checkpoint into the local encoder. This parameter will be in the section of training a linear classifier after SimCSE training part.
2. `freeze_weights` function is to convert the training status of encoder's weights to non-trainable. This function will be used in the linear classifier training part under both `Pretrained LM with a linear classifier` and `SimCSE with a linear classifier`.
3. `freeze_encoder_weights` is defined to choose whether freeze or unfreeze encoder's weights.

In [8]:
```
class LMWithLinearClassfier(BaseModel):
    def __init__(
          self,
          model_name: str = 'distilbert-base-multilingual-cased',
          ckpt_path: str = None,
          learning_rate: float = 2e-5,
          freeze_encoder_weights: bool = False
    ):
        super().__init__(
            model_name,
            learning_rate
        )
        self.save_hyperparameters()

        # TODO 2: Load encoder's weights from Pytorch Lightning's checkpoint
        if ckpt_path:
            ckpt = torch.load(ckpt_path)
            ckpt['state_dict'] = {k.replace("encoder.", ""): v for k, v in ckpt[
            self.encoder.load_state_dict(ckpt['state_dict'])

        # TODO 3: define a linear classifier which will output the 4 classes
        self.classifier = nn.Linear(self.encoder.config.hidden_size, 4)


        if freeze_encoder_weights:
          self.freeze_weights(self.encoder)  # Freeze model

        self.accuracy = Accuracy(task='multiclass', num_classes=4)

    # TODO 4: implement `freeze_weights` function which will set requires_grad
    # in the model.parameters() so that no gradient update will be done on the
    # base model. Only the linear_layer will be updated.
    def freeze_weights(self, model):
        for param in model.parameters():
            param.requires_grad = False

    # TODO 5: get logits from the classifier
    def forward(self, input_ids, attention_mask):
        embeddings = self.get_embeddings(input_ids, attention_mask)
        logits = self.classifier(embeddings)
        return logits

    def training_step(self, batch, batch_idx):
        # TODO 6.1: implement cross entropy loss for text classification
        # and log loss and acc
```

```
        input_ids = batch['input_ids']
        attention_mask = batch['attention_mask']
        labels = batch['labels']
        logits = self.forward(input_ids, attention_mask)
        loss = F.cross_entropy(logits, labels)
        acc = self.accuracy(logits, labels)

        self.log('train_loss', loss)
        self.log('train_acc', acc, prog_bar=True)
        return loss

    def validation_step(self, batch, batch_idx):
        # TODO 6.2: implement same as `training_step`
        input_ids = batch['input_ids']
        attention_mask = batch['attention_mask']
        labels = batch['labels']
        logits = self.forward(input_ids, attention_mask)
        loss = F.cross_entropy(logits, labels)
        acc = self.accuracy(logits, labels)

        self.log('val_loss', loss, prog_bar=True)
        self.log('val_acc', acc, prog_bar=True)

    def test_step(self, batch, batch_idx):
        # TODO 6.3: implement same as `training_step`
        input_ids = batch['input_ids']
        attention_mask = batch['attention_mask']
        labels = batch['labels']
        logits = self.forward(input_ids, attention_mask)
        loss = F.cross_entropy(logits, labels)
        acc = self.accuracy(logits, labels)

        self.log('test_loss', loss)
        self.log('test_acc', acc, prog_bar=True)
```

Pretrained LM with a linear classifier

To benchmark models, we need to have some baselines to compare how good the models' perfomance are.

The simplest baseline to measure the contrastive learning-based method is the pretrained LM which just fine-tunes only the last linear classifier head to predict sentiments (positive/negative/neutral/questions).

Define model

In [9]:
```
pretrained_lm_w_linear_model = LMWithLinearClassfier(
    model_name,
    ckpt_path=None,
    freeze_encoder_weights=True
)
```

```
model.safetensors:   0%|          | 0.00/542M [00:00<?, ?B/s]
```

Train a linear classifier

In [10]:
```python
# Create a ModelCheckpoint callback (recommended way):
pretrained_lm_w_linear_checkpoint_callback = pl.callbacks.ModelCheckpoint(
    monitor="val_acc",  # Metric to monitor
    mode="max",  # "min" for loss, "max" for accuracy
    save_top_k=1,  # Save only the best model(s)
    save_weights_only=True, # Saves only weights, not the entire model
    dirpath="./checkpoints/", # Path where the checkpoints will be saved
    filename="best_pretrained_w_linear_model-{epoch}-{val_acc:.2f}", # Customize
    verbose=True,
)

# Initialize trainer
pretrained_lm_w_linear_trainer = Trainer(
    max_epochs=3,
    accelerator='auto',
    callbacks=[pretrained_lm_w_linear_checkpoint_callback], # Add the ModelCheck
    gradient_clip_val=1.0,
    precision=16, # Mixed precision training
    devices=1,
)

# Train the model
pretrained_lm_w_linear_trainer.fit(pretrained_lm_w_linear_model, train_loader, v
```

```
/usr/local/lib/python3.10/dist-packages/lightning_fabric/connector.py:572: `preci
sion=16` is supported for historical reasons but its usage is discouraged. Please
set your precision to 16-mixed instead!
Sanity Checking: |          | 0/? [00:00<?, ?it/s]
/usr/local/lib/python3.10/dist-packages/pytorch_lightning/trainer/connectors/data
_connector.py:425: The 'val_dataloader' does not have many workers which may be a
bottleneck. Consider increasing the value of the `num_workers` argument` to `num_
workers=3` in the `DataLoader` to improve performance.
/usr/local/lib/python3.10/dist-packages/pytorch_lightning/trainer/connectors/data
_connector.py:425: The 'train_dataloader' does not have many workers which may be
a bottleneck. Consider increasing the value of the `num_workers` argument` to `nu
m_workers=3` in the `DataLoader` to improve performance.
Training: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
```

Evaluate

In [11]:
```python
pretrained_lm_w_linear_result = pretrained_lm_w_linear_trainer.test(pretrained_l
pretrained_lm_w_linear_result
```

```
/usr/local/lib/python3.10/dist-packages/pytorch_lightning/trainer/connectors/data
_connector.py:425: The 'test_dataloader' does not have many workers which may be
a bottleneck. Consider increasing the value of the `num_workers` argument` to `nu
m_workers=3` in the `DataLoader` to improve performance.
Testing: |          | 0/? [00:00<?, ?it/s]
```

Test metric	DataLoader 0
test_acc	0.5439910292625427
test_loss	1.0286043882369995

Out[11]: [{'test_loss': 1.0286043882369995, 'test_acc': 0.5439910292625427}]

2) Fine-tuned LM

This is the same as part 1, but you will also gradient update on the base model weights.

Define model

In [12]:
```
finetuned_lm_w_linear_model = LMWithLinearClassfier(
    model_name,
    ckpt_path=None,
    freeze_encoder_weights=False
)
```

Train both LM and a linear classifier

In [13]:
```
# Create a ModelCheckpoint callback (recommended way):
finetuned_lm_w_linear_checkpoint_callback = pl.callbacks.ModelCheckpoint(
    monitor="val_acc",  # Metric to monitor
    mode="max",  # "min" for loss, "max" for accuracy
    save_top_k=1,  # Save only the best model(s)
    save_weights_only=True, # Saves only weights, not the entire model
    dirpath="./checkpoints/", # Path where the checkpoints will be saved
    filename="best_finetuned_w_linear_model-{epoch}-{val_acc:.2f}", # Customized
    verbose=True,
)

# Initialize trainer
finetuned_lm_w_linear_trainer = Trainer(
    max_epochs=3,
    accelerator='auto',
    callbacks=[finetuned_lm_w_linear_checkpoint_callback], # Add the ModelCheckp
    gradient_clip_val=1.0,
    precision=16, # Mixed precision training
    devices=1,
)

# Train the model
finetuned_lm_w_linear_trainer.fit(finetuned_lm_w_linear_model, train_loader, val
```

```
/usr/local/lib/python3.10/dist-packages/pytorch_lightning/callbacks/model_checkpo
int.py:654: Checkpoint directory /kaggle/working/checkpoints exists and is not em
pty.
Sanity Checking: |          | 0/? [00:00<?, ?it/s]
Training: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
```

Evaluate

In [14]:
```
finetuned_lm_w_linear_result = finetuned_lm_w_linear_trainer.test(finetuned_lm_w
finetuned_lm_w_linear_result
```

```
Testing: |          | 0/? [00:00<?, ?it/s]
```

Test metric	DataLoader 0
test_acc	0.6911269426345825
test_loss	0.7557462453842163

Out[14]:
```
[{'test_loss': 0.7557462453842163, 'test_acc': 0.6911269426345825}]
```

Contrastive-based model (SimCSE) with a linear classifier

SimCSE (Simple Contrastive Learning of Sentence Embeddings) is a self-supervised learning method that learns high-quality sentence embeddings without relying on any labeled data. It leverages contrastive learning, a technique where similar examples are encouraged to have similar representations, while dissimilar examples are pushed apart in representation space.

Here's the core idea in a nutshell:

- **Data Augmentation**: SimCSE starts with a batch of sentences. For each sentence, it creates two slightly different "views" of the same sentence. These views are created through simple augmentations, like dropout (randomly masking some words) or other minor perturbations. These augmented sentences are semantically similar to the original.

- **Contrastive Objective**: The core of SimCSE is a contrastive loss function. It treats the two different views of the same sentence as a positive pair – the model should learn to make their embeddings similar. All other sentences in the batch (including their augmented versions) are treated as negative pairs – their embeddings should be dissimilar.

- **Learning**: The model is trained to minimize this contrastive loss. This forces the model to learn sentence embeddings that are robust to the augmentations and capture the underlying semantic meaning of the sentences. Sentences with similar meanings will have embeddings close together, while sentences with different meanings will have embeddings far apart.

Paper: https://arxiv.org/pdf/2104.08821.pdf

Unsupervised SimCSE is the foundation of the SimCSE method. It's a way to learn sentence embeddings without any labeled data.

Core idea of its concept

- **Dropout as Augmentation**: The key idea in unsupervised SimCSE is to use dropout (randomly masking some words during training) as a form of minimal data augmentation.

- **Two Views**: When you feed the same sentence through your transformer model twice, with dropout turned on, you get two slightly different representations (embeddings) of that sentence. These are like two "views" of the same sentence.
- **Contrastive Learning**: The two embeddings of the same sentence (the "views") are treated as a positive pair. The model is trained to make these embeddings similar to each other. The embeddings of different sentences in the batch are treated as negative pairs. The model is trained to make these embeddings dissimilar to each other.

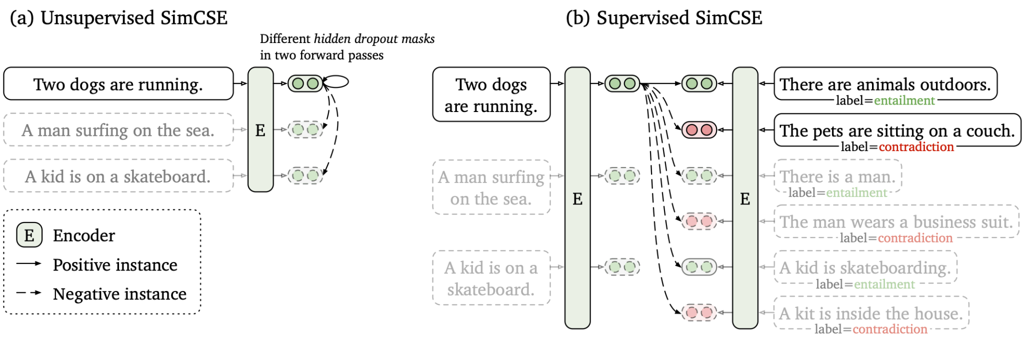

Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch instances as negatives.

Defined Unsupervised SimCSE model and InfoNCE loss

$$L_{UnsupervisedInfoNCE} = -\log \frac{e^{cos(z_i,z_j)/\tau}}{e^{cos(z_i,z_j)/\tau} . \sum_{k=0}^{N} (e^{cos(z_i,z_k)/\tau})}$$

Notation

z_i indicates the anchor representation (the representation that we are focusing on). The anchor sentence is the initial sentence which its representation is augmented by the dropout masking layer.

z_j indicates the positive representation (the representation that has the same semantic direction). The positive sentence is the same sentence as the anchor one but the positive representation is augmented in different way by the same dropout masking layer.

z_k indicates the negative representation (the representation that has the opposite semantic direction). The negative sentence are the other sentences sampled besides the anchor/positive sentence.

$cos(\cdot,\cdot)$ is cosine similairity function

N is the number of negative examples

Hint

For loss calculation section, I suggest you to use `F.crossentropy` function and the idea of in-batch negative sampling.

In [19]:

```
class UnsupervisedSimCSE(BaseModel):
    def __init__(
          self,
          model_name: str = 'distilbert-base-multilingual-cased',
          learning_rate: float = 2e-6,
          temperature: float = 0.05,
    ):
        super().__init__(
            model_name,
            learning_rate
        )
        self.save_hyperparameters()
        self.temperature = temperature

        # TODO 7: enable dropout masking in transformer layers to do data augmen
        # Dropout layers behave differently during training and inference
        # https://discuss.pytorch.org/t/if-my-model-has-dropout-do-i-have-to-alt
        self.encoder.train()

    def forward(self, input_ids, attention_mask):
        # TODO 8: get sentence embeddings
        outputs = self.encoder(input_ids=input_ids, attention_mask=attention_mas
        return outputs.last_hidden_state[:, 0]

    def info_nce_loss(self, embeddings):
        batch_size = embeddings.shape[0] // 2
        similarity_matrix = F.cosine_similarity(embeddings.unsqueeze(1), embeddi
        labels = torch.arange(batch_size, device=embeddings.device)
        logits = similarity_matrix[:batch_size, batch_size:] / self.temperature
        return F.cross_entropy(logits, labels)

    def training_step(self, batch, batch_idx):
        # TODO 9.1: implement unsupervised InfoNCE loss
        input_ids = batch['input_ids']
        attention_mask = batch['attention_mask']

        # First forward pass
        embeddings1 = self(input_ids, attention_mask)

        # Second forward pass with different dropout
        embeddings2 = self(input_ids, attention_mask)

        ## Combine embeddings
        embeddings = torch.cat([embeddings1, embeddings2], dim=0)
        ## Calculate loss
        loss = self.info_nce_loss(embeddings)
        ## Log loss
        self.log("train_loss", loss, prog_bar=True, logger=True)
        return loss

    def validation_step(self, batch, batch_idx):
        # TODO 9.2: implement the same as `training_step`
        return self.training_step(batch, batch_idx)

    def test_step(self, batch, batch_idx):
        # TODO 9.3: implement the same as `training_step`
        return self.training_step(batch, batch_idx)
```

Train LM through SimCSE approach

In [22]:
```
# Initialize model
model = UnsupervisedSimCSE()

# Initialize trainer
simcse_trainer = Trainer(
    max_epochs=3,
    accelerator='auto',
    devices=1,
    gradient_clip_val=1.0,
    precision=16  # Mixed precision training
)

# Train the model
simcse_trainer.fit(model, train_loader)

# Save the latest checkpoint
simcse_trainer.save_checkpoint('/content/latest_simcse_checkpoint.ckpt')
```

```
Training: |          | 0/? [00:00<?, ?it/s]
```

Define SimCSE with a linear classifier model

After training SimCSE on the data, we proceed to train a linear classifier on top of the trained model. Be sure to freeze the encoder weights.

In [23]:
```
latest_simcse_ckpt_path = '/content/latest_simcse_checkpoint.ckpt'

simcse_lm_w_linear_model = LMWithLinearClassfier(
    model_name,
    ckpt_path=latest_simcse_ckpt_path,
    freeze_encoder_weights=True
)
```

```
<ipython-input-8-2ddb6839cf3f>:17: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for more details). In a future release, the default value for `weights_only` will be flipped to `True`. This limits the functions that could be executed during unpickling. Arbitrary objects will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted by the user via `torch.serialization.add_safe_globals`. We recommend you start setting `weights_only=True` for any use case where you don't have full control of the loaded file. Please open an issue on GitHub for any issues related to this experimental feature.
  ckpt = torch.load(ckpt_path)
```

Train a linear classifier

In [24]:
```
# Create a ModelCheckpoint callback (recommended way):
simcse_lm_w_linear_checkpoint_callback = pl.callbacks.ModelCheckpoint(
    monitor="val_acc",  # Metric to monitor
    mode="max",  # "min" for loss, "max" for accuracy
    save_top_k=1,  # Save only the best model(s)
```

```python
    save_weights_only=True, # Saves only weights, not the entire model
    dirpath="./checkpoints/", # Path where the checkpoints will be saved
    filename="best_simcse_linear_model-{epoch}-{val_acc:.2f}", # Customized name
    verbose=True,
)

# Initialize trainer
simcse_lm_w_linear_trainer = Trainer(
    max_epochs=3,
    accelerator='auto',
    callbacks=[simcse_lm_w_linear_checkpoint_callback], # Add the ModelCheckpoin
    gradient_clip_val=1.0,
    precision=16, # Mixed precision training
    devices=1,
)

# Train the model
simcse_lm_w_linear_trainer.fit(simcse_lm_w_linear_model, train_loader, val_loade
```

```
Sanity Checking: |          | 0/? [00:00<?, ?it/s]
Training: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
Validation: |          | 0/? [00:00<?, ?it/s]
```

Evaluate

In [25]:

```python
simcse_lm_w_linear_result = simcse_lm_w_linear_trainer.test(simcse_lm_w_linear_m
simcse_lm_w_linear_result
```

```
Testing: |          | 0/? [00:00<?, ?it/s]
```

Test metric	DataLoader 0
test_acc	0.5844253301620483
test_loss	1.0677490234375

Out[25]:

```
[{'test_loss': 1.0677490234375, 'test_acc': 0.5844253301620483}]
```