CS-GY 6953 / ECE-GY 7123
Deep Learning

Homework #3

Artem Shlepchenko
as14836

Problem 1

a)

Parameter Efficiency: CNN models are typically more parameter efficient than RNN models for image classification tasks. This is because CNNs can share parameters across different regions of the input image, reducing the number of parameters needed to be learned. In contrast, RNNs have to learn separate parameters for each input element, which can lead to a much larger number of parameters.

Local Invariance: Another benefit of CNN models over RNN models for image classification is their ability to capture local invariances in the input image. CNNs are designed to recognise patterns in local regions of the image, allowing them to identify features regardless of their position or orientation. RNNs, on the other hand, are more suited for sequential data processing, such as natural language processing, where the order of the input elements matters.

b)

Sequential Processing: RNN models are designed to process sequential data and have a memory of previous inputs, making them well-suited for image classification tasks where the order of the input matters. For example, in tasks like image captioning, where the model needs to generate a sequence of words that describe the content of an image, RNNs can capture the sequential nature of language and generate more accurate and coherent descriptions.

Variable Length Inputs: Another benefit of RNN models over CNN models is their ability to handle inputs of variable length. In image classification tasks, inputs can have varying sizes and aspect ratios. While CNNs require fixed-size inputs, RNNs can handle inputs of different sizes and aspect ratios. This makes RNNs more flexible and adaptable to a wider range of input data, making them more suitable for certain types of image classification tasks.

Problem 2

$$y_t = sigmoid(1000\ h_t)$$

$$h_t = x_t - h_{t-1}$$

Since the input sequence is of even length, let us assume that the timesteps are numbered as 1, 2, 3, 4, ..., $2n$-2, $2n$-1, $2n$.
Then, we will evaluate h_{2n}

$$\begin{aligned} h_{2n} &= x_{2n} - h_{2n-1} \\ &= x_{2n} - (x_{2n-1} - h_{2n-2}) \\ &= x_{2n} - x_{2n-1} + (x_{2n-2} - h_{2n-3}) \\ &= x_{2n} - x_{2n-1} + x_{2n-2} - (x_{2n-3} - h_{2n-4}) \\ &= x_{2n} - x_{2n-1} + x_{2n-2} - x_{2n-3} + (x_{2n-4} - h_{2n-5}) \end{aligned}$$

If we evaluate this till the end, we get

$$= x_{2n} - x_{2n-1} + x_{2n-2} - x_{2n-3} + \ldots + x_4 - x_3 + x_2 - x_1 + h_0$$

Therefore, the value computed at the output unit is

$$y_{2n} = sigmoid(1000\ h_{2n})$$

where

$$h_{2n} = x_{2n} - x_{2n-1} + x_{2n-2} - x_{2n-3} + \ldots + x_4 - x_3 + x_2 - x_1 + h_0$$

Problem 3

We are given that all the queries (q), keys (k), and values (v) are the data points (x) themselves

$$x_i = q_i = k_i = v_i$$

Consider 4 orthogonal base a, b, c and d.

$$||a||_2 = ||b||_2 = ||c||_2 = ||d||_2 = \beta$$

From above 4 vectors, construct 3 tokens

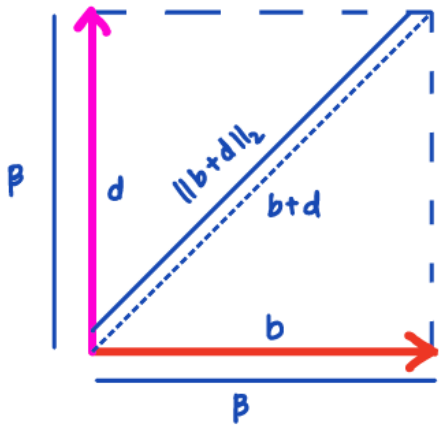
$$x_1 = b + d$$

$$x_2 = a$$

$$x_3 = c + b$$

a)
To calculate $||x_1||_2$, x_1 depends on b and d

$$||x_1||_2 = ||b + d||_2$$
$$\Rightarrow ||x_1||_2 = \sqrt{\beta^2 + \beta^2}$$
$$\Rightarrow ||x_1||_2 = \beta\sqrt{2}$$

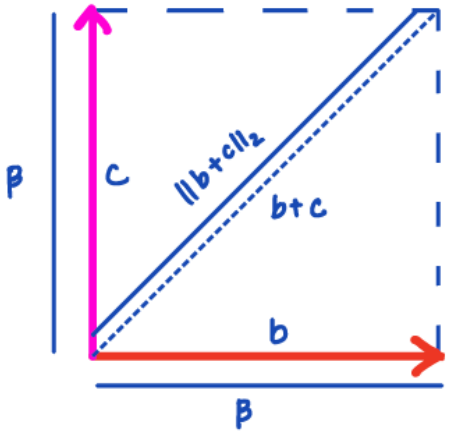


To calculate $||x_2||_2$, x_2 depends on a

$$||x_2||_2 = ||a||_2$$
$$\Rightarrow ||x_2||_2 = \beta$$

To calculate $||x_3||_2$, x_3 depends on c and b

$$||x_3||_2 = ||c + b||_2$$
$$\Rightarrow ||x_3||_2 = \sqrt{\beta^2 + \beta^2}$$
$$\Rightarrow ||x_3||_2 = \beta\sqrt{2}$$



b)
Compute (y_1, y_2, y_3) = Self-attention (x_1, x_2, x_3). We know that

$$y_i = (q_i {k_1}^T) v_i$$

or, in matrix form:

$$Y = (Q\ K)\ V$$

We have given that

$$Q = K = V = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \end{bmatrix}_{3 \times P}$$

when P is an arbitrary dimension.

$$Y = \left[\begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \end{bmatrix} \times \begin{bmatrix} - & - & - \\ x_1 & x_2 & x_3 \\ - & - & - \end{bmatrix}\right] \times \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \end{bmatrix}$$

$$= \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & x_1^T x_3 \\ x_2^T x_1 & x_2^T x_2 & x_2^T x_3 \\ x_3^T x_1 & x_3^T x_2 & x_3^T x_3 \end{bmatrix} \times \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \end{bmatrix}$$

$$= \begin{bmatrix} (b+d)^T(b+d) & (b+d)^T a & (b+d)^T(c+b) \\ a^T(b+d) & a^T a & a^T(c+b) \\ (c+b)^T(b+d) & (c+b)^T x_2 & (c+b)^T(c+b) \end{bmatrix} \times \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \end{bmatrix}$$

$$= \begin{bmatrix} b^T b + \cancel{b^T d} + \cancel{d^T b} + d^T d & \cancel{b^T a + d^T a} & \cancel{b^T c} + b^T b + \cancel{d^T c} + \cancel{d^T b} \\ \cancel{a^T b + a^T d} & a^T a & \cancel{a^T c + a^T b} \\ \cancel{c^T b} + \cancel{c^T d} + b^T b + \cancel{b^T d} & \cancel{c^T a + b^T a} & c^T c + \cancel{c^T b} + \cancel{b^T c} + b^T b \end{bmatrix} \times \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \end{bmatrix}$$

$$= \begin{bmatrix} 2\beta^2 & 0 & \beta^2 \\ 0 & \beta^2 & 0 \\ \beta^2 & 0 & 2\beta^2 \end{bmatrix} \times \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \end{bmatrix}$$

$$\Rightarrow y_1 = 2\beta^2 x_1 + \beta^2 x_3$$

$$\Rightarrow y_2 = \beta^2 x_2$$

$$\Rightarrow y_3 = \beta^2 x_1 + 2\beta^2 x_3$$

The linear combination of x_1, x_3 is approximated by y_1 and y_3, x_2 is approximated by y_2.

c)

If we look at y_1, y_2 and y_3

$$\Rightarrow y_1 = 2\beta^2 x_1 + \beta^2 x_3$$

$$\Rightarrow y_2 = \beta^2 x_2$$

$$\Rightarrow y_3 = \beta^2 x_1 + 2\beta^2 x_3$$

y_1 depends on linear combination of x_1 and x_3, it means network copies the input (x_1 and x_3) to compute y_1.

y_2 depends only x_2, it means network copies the x_2 to compute y_2.

y_3 depends on linear combination of x_1 and x_3, it means network copies the input (x_1 and x_3) to compute y_3.

Problem 4

Analyzing movie reviews using transformers

This problem asks you to train a sentiment analysis model using the BERT (Bidirectional Encoder Representations from Transformers) model, introduced here. Specifically, we will parse movie reviews and classify their sentiment (according to whether they are positive or negative.)
We will use the Huggingface transformers library to load a pre-trained BERT model to compute text embeddings, and append this with an RNN model to perform sentiment classification.

Data preparation

Before delving into the model training, let's first do some basic data processing. The first challenge in NLP is to encode text into vector-style representations. This is done by a process called *tokenization*.

```
%pip install -U torch==1.8.0 torchtext==0.9.0

# Reload environment
exit()
```

```
import torch
import random
import numpy as np

SEED = 1234

random.seed(SEED)
np.random.seed(SEED)
torch.manual_seed(SEED)
torch.backends.cudnn.deterministic = True
```

Let us load the transformers library first.

```
%pip install transformers
```

Each transformer model is associated with a particular approach of tokenizing the input text. We will use the `bert-base-uncased` model below, so let's examine its corresponding tokenizer.

```
from transformers import BertTokenizer

tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
```

The `tokenizer` has a `vocab` attribute which contains the actual vocabulary we will be using. First, let us discover how many tokens are in this language model by checking its length.

```
# Q1a: Print the size of the vocabulary of the above tokenizer.
print(f"The size of the vocabulary is: {len(tokenizer)}")
```

```
The size of the vocabulary is: 30522
```

Using the tokenizer is as simple as calling `tokenizer.tokenize` on a string. This will tokenize and lower case the data in a way that is consistent with the pre-trained transformer model.

```
tokens = tokenizer.tokenize('Hello WORLD how ARE yoU?')

print(tokens)
```

```
['hello', 'world', 'how', 'are', 'you', '?']
```

We can numericalize tokens using our vocabulary using `tokenizer.convert_tokens_to_ids`.

```
indexes = tokenizer.convert_tokens_to_ids(tokens)

print(indexes)
```

```
[7592, 2088, 2129, 2024, 2017, 1029]
```

The transformer was also trained with special tokens to mark the beginning and end of the sentence, as well as a standard padding and unknown token.

Let us declare them.

```
init_token = tokenizer.cls_token
eos_token = tokenizer.sep_token
pad_token = tokenizer.pad_token
unk_token = tokenizer.unk_token

print(init_token, eos_token, pad_token, unk_token)
```

```
[CLS] [SEP] [PAD] [UNK]
```

We can call a function to find the indices of the special tokens.

```
init_token_idx = tokenizer.convert_tokens_to_ids(init_token)
eos_token_idx = tokenizer.convert_tokens_to_ids(eos_token)
pad_token_idx = tokenizer.convert_tokens_to_ids(pad_token)
unk_token_idx = tokenizer.convert_tokens_to_ids(unk_token)

print(init_token_idx, eos_token_idx, pad_token_idx, unk_token_idx)
```

```
101 102 0 100
```

We can also find the maximum length of these input sizes by checking the `max_model_input_sizes` attribute (for this model, it is 512 tokens).

```
max_input_length = tokenizer.max_model_input_sizes['bert-base-uncased']
```

Let us now define a function to tokenize any sentence, and cut length down to 510 tokens (we need one special `start` and `end` token for each sentence).

```
def tokenize_and_cut(sentence):
    tokens = tokenizer.tokenize(sentence)
    tokens = tokens[:max_input_length-2]
    return tokens
```

Finally, we are ready to load our dataset. We will use the IMDB Moview Reviews dataset. Let us also split the train dataset to form a small validation set (to keep track of the best model).

```
from torchtext.legacy import data

TEXT = data.Field(batch_first = True,
                  use_vocab = False,
                  tokenize = tokenize_and_cut,
                  preprocessing = tokenizer.convert_tokens_to_ids,
                  init_token = init_token_idx,
                  eos_token = eos_token_idx,
                  pad_token = pad_token_idx,
                  unk_token = unk_token_idx)

LABEL = data.LabelField(dtype = torch.float)
```

```
from torchtext.legacy import datasets
```

```
train_data, test_data = datasets.IMDB.splits(TEXT, LABEL)

train_data, valid_data = train_data.split(random_state = random.seed(SEED))
```

```
downloading aclImdb_v1.tar.gz
aclImdb_v1.tar.gz: 100%|| 84.1M/84.1M [00:03<00:00, 27.7MB/s]
```

Let us examine the size of the train, validation, and test dataset.

```
# Q1b. Print the number of data points in the train, test, and validation sets.
print(f"The number of data points in the train set:\t {len(train_data)}")
print(f"The number of data points in the test set:\t {len(test_data)}")
print(f"The number of data points in the validation set: {len(valid_data)}")
```

```
The number of data points in the train set:         17500
The number of data points in the test set:          25000
The number of data points in the validation set:    7500
```

We will build a vocabulary for the labels using the `vocab.stoi` mapping.

```
LABEL.build_vocab(train_data)
```

```
print(LABEL.vocab.stoi)
```

```
defaultdict(None, {'neg': 0, 'pos': 1})
```

Finally, we will set up the data-loader using a (large) batch size of 128. For text processing, we use the `BucketIterator` class.

```
BATCH_SIZE = 128

device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')

train_iterator, valid_iterator, test_iterator = data.BucketIterator.splits(
    (train_data, valid_data, test_data),
    batch_size = BATCH_SIZE,
    device = device)
```

Model preparation

We will now load our pretrained BERT model. (Keep in mind that we should use the same model as the tokenizer that we chose above).

```
from transformers import BertTokenizer, BertModel

bert = BertModel.from_pretrained('bert-base-uncased')
```

As mentioned above, we will append the BERT model with a bidirectional GRU to perform the classification.

```
import torch.nn as nn

class BERTGRUSentiment(nn.Module):
    def __init__(self,bert,hidden_dim,output_dim,n_layers,bidirectional,dropout):

        super().__init__()

        self.bert = bert

        embedding_dim = bert.config.to_dict()['hidden_size']

        self.rnn = nn.GRU(embedding_dim,
                          hidden_dim,
                          num_layers = n_layers,
                          bidirectional = bidirectional,
                          batch_first = True,
                          dropout = 0 if n_layers < 2 else dropout)

        self.out = nn.Linear(hidden_dim * 2 if bidirectional else hidden_dim, output_dim)

        self.dropout = nn.Dropout(dropout)

    def forward(self, text):

        #text = [batch size, sent len]

        with torch.no_grad():
            embedded = self.bert(text)[0]

        #embedded = [batch size, sent len, emb dim]

        _, hidden = self.rnn(embedded)
```

```
        #hidden = [n layers * n directions, batch size, emb dim]

        if self.rnn.bidirectional:
            hidden = self.dropout(torch.cat((hidden[-2,:,:], hidden[-1,:,:]), dim = 1))
        else:
            hidden = self.dropout(hidden[-1,:,:])

        #hidden = [batch size, hid dim]

        output = self.out(hidden)

        #output = [batch size, out dim]

        return output
```

Next, we'll define our actual model.

Our model will consist of

- the BERT embedding (whose weights are frozen)
- a bidirectional GRU with 2 layers, with hidden dim 256 and dropout=0.25.
- a linear layer on top which does binary sentiment classification.

Let us create an instance of this model.

```
# Q2a: Instantiate the above model by setting the right hyperparameters.

# insert code here
HIDDEN_DIM = 256
OUTPUT_DIM = 1
N_LAYERS = 2
BIDIRECTIONAL = True
DROPOUT = 0.25

model = BERTGRUSentiment(bert,
                         HIDDEN_DIM,
                         OUTPUT_DIM,
                         N_LAYERS,
                         BIDIRECTIONAL,
                         DROPOUT)
```

We can check how many parameters the model has.

```
# Q2b: Print the number of trainable parameters in this model.

# insert code here.
def count_parameters(model):
  return sum(param.numel() for param in model.parameters() if param.requires_grad)

print(f"The number of trainable parameters in this model is {count_parameters(model)}")
```

```
The number of trainable parameters in this model is 112241409
```

Oh no~ if you did this correctly, youy should see that this contains *112 million* parameters. Standard machines (or Colab) cannot handle such large models.
However, the majority of these parameters are from the BERT embedding, which we are not going to (re)train. In order to freeze certain parameters we can set their `requires_grad` attribute to `False`. To do this, we simply loop through all of the `named_parameters` in our model and if they're a part of the `bert` transformer model, we set `requires_grad = False`.

```
for name, param in model.named_parameters():
    if name.startswith('bert'):
        param.requires_grad = False
```

```
# Q2c: After freezing the BERT weights/biases, print the number of remaining
# trainable parameters.

print(f"The number of trainable parameters in this model is {count_parameters(model)}")
```

```
The number of trainable parameters in this model is 2759169
```

We should now see that our model has under 3M trainable parameters. Still not trivial but manageable.

Train the Model

All this is now largely standard.
We will use:

- the Binary Cross Entropy loss function: `nn.BCEWithLogitsLoss()`
- the Adam optimizer

and run it for 2 epochs (that should be enough to start getting meaningful results).

```
import torch.optim as optim

optimizer = optim.Adam(model.parameters())
```

```
criterion = nn.BCEWithLogitsLoss()
```

```
model = model.to(device)
criterion = criterion.to(device)
```

Also, define functions for:

- calculating accuracy.
- training for a single epoch, and reporting loss/accuracy.
- performing an evaluation epoch, and reporting loss/accuracy.
- calculating running times.

```
def binary_accuracy(preds, y):

    # Q3a. Compute accuracy (as a number between 0 and 1)
    rounded_preds = torch.round(torch.sigmoid(preds))
    corrections = (rounded_preds == y).float()
    acc = corrections.sum() / len(corrections)
    # ...

    return acc
```

```
def train(model, iterator, optimizer, criterion):

    # Q3b. Set up the training function
    epoch_loss = 0
    epoch_acc = 0

    model.train()

    for batch in iterator:

        optimizer.zero_grad()
```

```
        predictions = model(batch.text).squeeze(1)

        loss = criterion(predictions, batch.label)

        acc = binary_accuracy(predictions, batch.label)

        loss.backward()

        optimizer.step()

        epoch_loss += loss.item()
        epoch_acc += acc.item()
    # ...

    return epoch_loss / len(iterator), epoch_acc / len(iterator)
```

```
def evaluate(model, iterator, criterion):

    # Q3c. Set up the evaluation function.
    epoch_loss = 0
    epoch_acc = 0

    model.eval()

    with torch.no_grad():

        for batch in iterator:

            predictions = model(batch.text).squeeze(1)

            loss = criterion(predictions, batch.label)

            acc = binary_accuracy(predictions, batch.label)

            epoch_loss += loss.item()
            epoch_acc += acc.item()
    # ...

    return epoch_loss / len(iterator), epoch_acc / len(iterator)
```

Problem 4 continued on next page. . .

```
import time

def epoch_time(start_time, end_time):
    elapsed_time = end_time - start_time
    elapsed_mins = int(elapsed_time / 60)
    elapsed_secs = int(elapsed_time - (elapsed_mins * 60))
    return elapsed_mins, elapsed_secs
```

We are now ready to train our model.

Statutory warning: Training such models will take a very long time since this model is considerably larger than anything we have trained before. Even though we are not training any of the BERT parameters, we still have to make a forward pass. This will take time; each epoch may take upwards of 30 minutes on Colab.

Let us train for 2 epochs and print train loss/accuracy and validation loss/accuracy for each epoch. Let us also measure running time.

Saving intermediate model checkpoints using

`torch.save(model.state_dict(),'model.pt')`

may be helpful with such large models.

```
N_EPOCHS = 2

best_valid_loss = float('inf')

for epoch in range(N_EPOCHS):

    # Q3d. Perform training/valudation by using the functions you defined earlier.

    start_time = time.time()

    train_loss, train_acc = train(model, train_iterator, optimizer, criterion)
    valid_loss, valid_acc = evaluate(model, valid_iterator, criterion)

    end_time = time.time()

    epoch_mins, epoch_secs = epoch_time(start_time, end_time)

    if valid_loss < best_valid_loss:
```

```
        best_valid_loss = valid_loss
        torch.save(model.state_dict(), 'model.pt')

    print(f'Epoch: {epoch+1:02} | Epoch Time: {epoch_mins}m {epoch_secs}s')
    print(f'\tTrain Loss: {train_loss:.3f} | Train Acc: {train_acc*100:.2f}%')
    print(f'\t Val. Loss: {valid_loss:.3f} |  Val. Acc: {valid_acc*100:.2f}%')
```

```
Epoch: 01 | Epoch Time: 150m 40s
Train Loss: 0.520 | Train Acc: 73.74%
 Val. Loss: 0.315 |  Val. Acc: 87.54%
Epoch: 02 | Epoch Time: 133m 13s
Train Loss: 0.314 | Train Acc: 86.86%
 Val. Loss: 0.238 |  Val. Acc: 90.54%
```

Load the best model parameters (measured in terms of validation loss) and evaluate the loss/accuracy on the test set.

```
model.load_state_dict(torch.load('model.pt'))

test_loss, test_acc = evaluate(model, test_iterator, criterion)

print(f'Test Loss: {test_loss:.3f} | Test Acc: {test_acc*100:.2f}%')
```

```
Test Loss: 0.220 | Test Acc: 91.11%
```

Inference

We'll then use the model to test the sentiment of some fake movie reviews. We tokenize the input sentence, trim it down to length=510, add the special start and end tokens to either side, convert it to a `LongTensor`, add a fake batch dimension using `unsqueeze`, and perform inference using our model.

```
def predict_sentiment(model, tokenizer, sentence):
    model.eval()
    tokens = tokenizer.tokenize(sentence)
    tokens = tokens[:max_input_length-2]
    indexed = [init_token_idx] + tokenizer.convert_tokens_to_ids(tokens) + [eos_token_idx]
    tensor = torch.LongTensor(indexed).to(device)
    tensor = tensor.unsqueeze(0)
    prediction = torch.sigmoid(model(tensor))
    return prediction.item()
```

```
# Q4a. Perform sentiment analysis on the following two sentences.

predict_sentiment(model, tokenizer, "Justice League is terrible. I hated it.")
```

```
0.015669671818614006
```

```
predict_sentiment(model, tokenizer, "Avengers was great!!")
```

```
0.7836444973945618
```

Great! Try playing around with two other movie reviews (you can grab some off the internet or make up text yourselves), and see whether your sentiment classifier is correctly capturing the mood of the review.

```
# Q4b. Perform sentiment analysis on two other movie review fragments of your choice.
predict_sentiment(model,
                  tokenizer,
                  "One of the greatest family-oriented, fantasy-adventure movies ever.")
```

```
0.9650945663452148
```

```
predict_sentiment(model,
                  tokenizer,
                  "Hugh Grant and Sandra Bullock are two such likeable actors.")
```

```
0.6016971468925476
```