Assignment 4: Language Processing with RNN-Based Autoencoders

Deadline: Sunday, June 15th, by 9pm.

Submission: Submit a PDF export of the completed notebook as well as the ipynb file.

In this assignement, we will practice the application of deep learning to natural language processing. We will be working with a subset of Reuters news headlines that are collected over 15 months, covering all of 2019, plus a few months in 2018 and in a few months of this year.

In particular, we will be building an **autoencoder** of news headlines. The idea is similar to the kind of image autoencoder we built in lecture: we will have an **encoder** that maps a news headline to a vector embedding, and then a **decoder** that reconstructs the news headline. Both our encoder and decoder networks will be Recurrent Neural Networks, so that you have a chance to practice building

- a neural network that takes a sequence as an input
- a neural network that generates a sequence as an output

This assignment is organized as follows:

- Question 1. Exploring the data
- Question 2. Building the autoencoder
- Question 3. Training the autoencoder using *data augmentation*
- Question 4. Analyzing the embeddings (interpolating between headlines)

Furthermore, we'll be introducing the idea of **data augmentation** for improving of the robustness of the autoencoder, as proposed by Shen et al [1] in ICML 2020.

[1] Shen, Tianxiao, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. "Educating text autoencoders: Latent representation guidance via denoising." In International Conference on Machine Learning, pp. 8719-8729. PMLR, 2020.

In [186…
```
import torch
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim

import matplotlib.pyplot as plt
import numpy as np
import random
```

In [187…
```
# helper class to print nice outputs
class txt:
   BOLD = '\033[1m'
   UNDERLINE = '\033[4m'
   END = '\033[0m'

def make_bold(text):
    return txt.BOLD+ text + txt.END

def make_underline(text):
    return txt.UNDERLINE+ text + txt.END
```

Question 1. Data (20 %)

Download the files `reuters_train.txt` and `reuters_valid.txt`, and upload them to Google Drive.

Then, mount Google Drive from your Google Colab notebook:

In [188…
```
from google.colab import drive
drive.mount('/content/gdrive')
train_path = '/content/gdrive/My Drive/Intro_to_Deep_Learning/assignment4/reuters_train.txt' # Update me
valid_path = '/content/gdrive/My Drive/Intro_to_Deep_Learning/assignment4/reuters_valid.txt' # Update me
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", f
orce_remount=True).
```

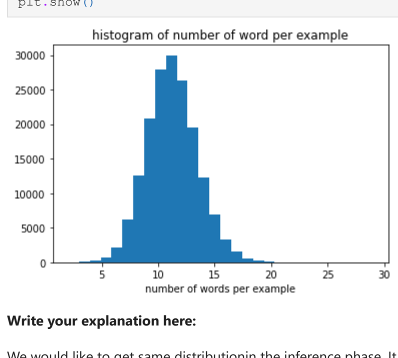
As we did in some of our examples (e.g., training transformers on IMDB reviews) will be using PyTorch's `torchtext` utilities to help us load, process, and batch the data. We'll be using a `TabularDataset` to load our data, which works well on structured CSV data with fixed columns (e.g. a column for the sequence, a column for the label). Our tabular dataset is even simpler: we have no labels, just some text. So, we are treating our data as a table with one field representing our sequence.

In [189…
```
import torchtext.legacy.data as data

# Tokenization function to separate a headline into words
def tokenize_headline(headline):
    """Returns the sequence of words in the string headline. We also
    prepend the "<bos>" or beginning-of-string token, and append the
    "<eos>" or end-of-string token to the headline.
    """
    return ("<bos> " + headline + " <eos>").split()

# Data field (column) representing our *text*.
text_field = data.Field(
    sequential=True,            # this field consists of a sequence
    tokenize=tokenize_headline, # how to split sequences into words
    include_lengths=True,       # to track the length of sequences, for batching
    batch_first=True,           # similar to batch_first=True used in nn.RNN demonstrated in lecture
    use_vocab=True)             # to turn each character into an integer index
train_data = data.TabularDataset(
    path=train_path,                # data file path
    format="tsv",                   # fields are separated by a tab
    fields=[('title', text_field)]) # list of fields (we have only one)
```

Part (a) -- 5%

Draw histograms of the number of words per headline in our training set. Excluding the `<bos>` and `<eos>` tags in your computation. Explain why we would be interested in such histograms.

In [190…
```
from collections import Counter

words_per_example=[(len(example.title)-2) for  example in train_data]
plt.hist(words_per_example, bins=max(words_per_example)-1)
plt.title("histogram of number of word per example")
plt.xlabel('number of words per example')
plt.show()
```



Write your explanation here:

We would like to get same distributionin the inference phase. It is important to see that most headline have between 5 to 20 words. We would expect that the new headline we will get will be also with 5-20 words.

Part (b) -- 5%

How many distinct words appear in the training data? Exclude the `<bos>` and `<eos>` tags in your computation.

In [191…
```
all_words=[]
for  example in train_data:
  all_words+=(example.title)

print(make_underline("number of distinct words in training data:"),len(set(all_words))-2)
```

```
number of distinct words in training data: 51298
```

Part (c) -- 5%

The distribution of *words* will have a long tail, meaning that there are some words that will appear very often, and many words that will appear infrequently. How many words appear exactly once in the training set? Exactly twice? Print these numbers below

In [192…
```
# Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values
c_words=Counter(all_words)
print(make_underline("number of words with 1 occurrence:"),len([word for word, count in c_words.items() if cour
print(make_underline("number of words with 2 occurrences:"),len([word for word, count in c_words.items() if cou
```

```
number of words with 1 occurrence: 19854
number of words with 2 occurrences: 7193
```

Part (d) -- 5%

We will replace the infrequent words with an `<unk>` tag, instead of learning embeddings for these rare words. `torchtext` also provides us with the `<pad>` tag used for padding short sequences for batching. We will thus only model the top 9995 words in the training set, excluding the tags `<bos>`, `<eos>`, `<unk>`, and `<pad>`.

What percentage of total word count(whole dataset) will be supported? Alternatively, what percentage of total word count(whole dataset) in the training set will be set to the `<unk>` tag?

In [193…
```
# Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values
most_common_words=[count for word,count in c_words.most_common(9997)[2:]]
all_words=[count for word,count in c_words.most_common()[2:]]
p_supported=sum(most_common_words)/sum(all_words)
print(make_underline("persentege of total supported word count"), p_supported*100,"%")
print(make_underline("persentege of total supported word count"), (1-p_supported)*100 ,"%")
```

```
persentege of total supported word count 93.97857393100142 %
persentege of total supported word count 6.021426068998581 %
```

The `torchtext` package will help us keep track of our list of unique words, known as a **vocabulary**. A vocabulary also assigns a unique integer index to each word.

In [194…
```
# Build the vocabulary based on the training data. The vocabulary
# can have at most 9997 words (9995 words + the <bos> and <eos> token)
text_field.build_vocab(train_data, max_size=9997)

# This vocabulary object will be helpful for us
vocab = text_field.vocab
print(vocab.stoi["hello"]) # for instances, we can convert from string to (unique) index
print(vocab.itos[10])      # ... and from word index to string

# The size of our vocabulary
vocab_size = len(text_field.vocab.stoi)

# Here are the two tokens that torchtext adds for us:
print(vocab.itos[0]) # <unk> represents an unknown word not in our vocabulary
print(vocab.itos[1]) # <pad> will be used to pad short sequences for batching
```

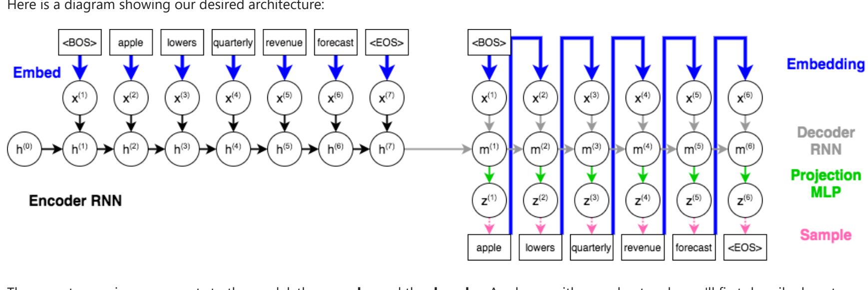
```
0
on
<unk>
<pad>
```

Question 2. Text Autoencoder (40%)

Building a text autoencoder is a little more complicated than an image autoencoder like we did in class. So we will need to thoroughly understand the model that we want to build before actually building it. Note that the best and fastest way to complete this assignment is to spend time upfront understanding the architecture. The explanations are quite dense, but it is important to understand the operation of this model. The rationale here is similar in nature to the `seq2seq` RNN model we discussed in class, only we are dealing with unsupervised learning here rather than machine translation.

Architecture description

Here is a diagram showing our desired architecture:



There are two main components to the model: the **encoder** and the **decoder**. As always with neural networks, we'll first describe how to make **predictions** with of these components. Let's get started:

The **encoder** will take a sequence of words (a headline) as *input*, and produce an embedding (a vector) that represents the entire headline. In the diagram above, the vector $\mathbf{h}^{(7)}$ is the vector embedding containing information about the entire headline. This portion is very similar to the sentiment analysis RNN that we discussed in lecture (but without the fully-connected layer that makes a prediction).

The **decoder** will take an embedding (in the diagram, the vector $\mathbf{h}^{(7)}$) as input, and uses a separate RNN to **generate a sequence of words**. To generate a sequence of words, the decoder needs to do the following:

1. Determine the previous word that was generated. This previous word will act as $\mathbf{x}^{(t)}$ to our RNN, and will be used to update the hidden state $\mathbf{m}^{(t)}$. Since each of our sequences begin with the `<bos>` token, we'll set $\mathbf{x}^{(1)}$ to be the `<bos>` token.
2. Compute the updates to the hidden state $\mathbf{m}^{(t)}$ based on the previous hidden state $\mathbf{m}^{(t-1)}$ and $\mathbf{x}^{(t)}$. Intuitively, this hidden state vector $\mathbf{m}^{(t)}$ is a representation of *all the words we still need to generate*.
3. We'll use a fully-connected layer to take a hidden state $\mathbf{m}^{(t)}$, and determine *what the next word should be*. This fully-connected layer solves a *classification problem*, since we are trying to choose a word out of $K =$ `vocab_size` distinct words. As in a classification problem, the fully-connected neural network will compute a *probability distribution* over these `vocab_size` words. In the diagram, we are using $\mathbf{z}^{(t)}$ to represent the logits, or the pre-softmax activation values representing the probability distribution.
4. We will need to *sample* an actual word from this probability distribution $\mathbf{z}^{(t)}$. We can do this in a number of ways, which we'll discuss in question 3. For now, you can imagine your favourite way of picking a word given a distribution over words.
5. This word we choose will become the next input $\mathbf{x}^{(t+1)}$ to our RNN, which is used to update our hidden state $\mathbf{m}^{(t+1)}$, i.e., to determine what are the remaining words to be generated.

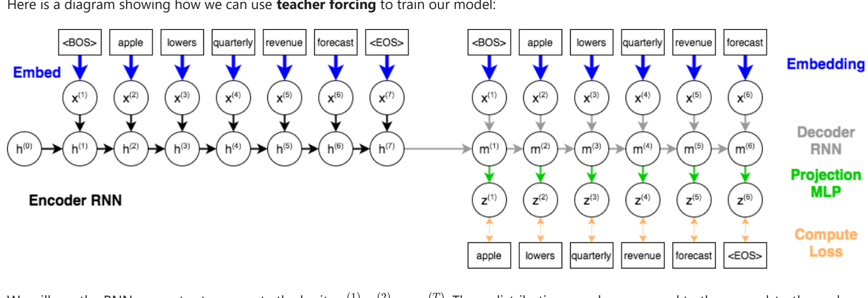
We can repeat this process until we see an `<eos>` token generated, or until the generated sequence becomes too long.

Training the architecture

While our autoencoder produces a sequence, computing the loss by comparing the complete generated sequence to the ground truth (the encoder input) gives rise to multiple challanges. One is that the generated sequence might be longer or shorter than the actual sequence, meaning that there may be more/fewer $\mathbf{z}^{(t)}$s than ground-truth words. Another more insidious issue is that the **gradients will become very high-variance and unstable**, because **early mistakes will easily throw the model off-track**. Early in training, our model is unlikely to produce the right answer in step $t = 1$, so the gradients we obtain based on the other time steps will not be very useful.

At this point, you might have some ideas about "hacks" we can use to make training work. Fortunately, there is one very well-established solution called **teacher forcing** which we can use for training: instead of *sampling* the next word based on $\mathbf{z}^{(t)}$, we will forget sampling, and use the **ground truth** $\mathbf{x}^{(t)}$ as the input in the next step.

Here is a diagram showing how we can use **teacher forcing** to train our model:



We will use the RNN generator to compute the logits $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots \mathbf{z}^{(T)}$. These distributions can be compared to the ground-truth words using the cross-entropy loss. The loss function for this model will be the sum of the losses across each $t \in \{1, \ldots, T\}$.

We'll train the encoder and decoder model simultaneously. There are several components to our model that contain tunable weights:

- The word embedding that maps a word to a vector representation. In theory, we could use GloVe embeddings, as we did in class. In this assignment we will not do that, but learn the word embedding from data. The word embedding component is represented with blue arrows in the diagram.
- The encoder RNN (which will use GRUs) that computes the embedding over the entire headline. The encoder RNN is represented with black arrows in the diagram.
- The decoder RNN (which will also use GRUs) that computes hidden states, which are vectors representing what words are to be generated. The decoder RNN is represented with gray arrows in the diagram.
- The **projection MLP** (a fully-connected layer) that computes a distribution over the next word to generate, given a decoder RNN hidden state. The projection is represented with green arrows

Part (a) -- 20%

Complete the code for the AutoEncoder class below by:

1. Filling in the missing numbers in the `__init__` method using the parameters `vocab_size`, `emb_size`, and `hidden_size`.
2. Complete the `forward` method, which uses teacher forcing and computes the logits $\mathbf{z}^{(t)}$ of the reconstruction of the sequence.

You should first try to understand the `encode` and `decode` methods, which are written for you. The `encode` method bears much similarity to the RNN we wrote in class for sentiment analysis. The `decode` method is a bit more challenging. You might want to scroll down to the `sample_sequence` function to see how this function will be called.

You can (but don't have to) use the `encode` and `decode` method in your `forward` method. In either case, be careful of the input that you feed into ether `decode` or to `self.decoder_rnn`. Refer to the teacher-forcing diagram. **bold text** Notice that batch_first is set to True, understand how deal with it.

In [195…
```
class AutoEncoder(nn.Module):
    def __init__(self, vocab_size, emb_size, hidden_size):
        """
        A text autoencoder. The parameters
            - vocab_size: number of unique words/tokens in the vocabulary
            - emb_size: size of the word embeddings $x^{(t)}$
            - hidden_size: size of the hidden states in both the
                           encoder RNN ($h^{(t)}$) and the
                           decoder RNN ($m^{(t)}$)
        """
        super().__init__()
        self.embed = nn.Embedding(num_embeddings=vocab_size, # TODO
                                  embedding_dim=emb_size)   # TODO
        self.encoder_rnn = nn.GRU(input_size=emb_size, #TODO
                                  hidden_size=hidden_size, #TODO
                                  batch_first=True)
        self.decoder_rnn = nn.GRU(input_size=emb_size, #TODO
                                  hidden_size=hidden_size, #TODO
                                  batch_first=True)
        self.proj = nn.Linear(in_features=hidden_size, # TODO
                              out_features=vocab_size) # vocab_size in numm of classes

    def encode(self, inp):
        """
        Computes the encoder output given a sequence of words.
        """
        emb = self.embed(inp)
        out, last_hidden = self.encoder_rnn(emb)
        return last_hidden

    def decode(self, inp, hidden=None):
        """
        Computes the decoder output given a sequence of words, and
        (optionally) an initial hidden state.
        """
        emb = self.embed(inp)
        out, last_hidden = self.decoder_rnn(emb, hidden)
        out_seq = self.proj(out)
        return out_seq, last_hidden

    def forward(self, inp):
        """
        Compute both the encoder and decoder forward pass
        given an integer input sequence inp with shape [batch_size, seq_length],
        with inp[a,b] representing the (index in our vocabulary of) the b-th word
        of the a-th training example.

        This function should return the logits $z^{(t)}$ in a tensor of shape
        [batch_size, seq_length - 1, vocab_size], computed using *teaching forcing*.

        The (seq_length - 1) part is not a typo. If you don't understand why
        we need to subtract 1, refer to the teacher-forcing diagram above.
        """

        last_hidden = self.encode(inp)
        # [batch_size,seq_length-1]
        new_inp=inp[:,:-1]
        out_seq, last_hidden = self.decode(new_inp,last_hidden)
        return  out_seq
```

Part (b) -- 10%

To check that your model is set up correctly, we'll train our autoencoder neural network for at least 300 iterations to memorize this sequence:

In [196…
```
headline = train_data[42].title
input_seq = torch.Tensor([vocab.stoi[w] for w in headline]).long().unsqueeze(0)
```

We are looking for the way that you set up your loss function corresponding to the figure above. Be careful of off-by-one errors here.

Note that the Cross Entropy Loss expects a rank-2 tensor as its first argument (the output of the network), and a rank-1 tensor as its second argument (the true label). You will need to properly reshape your data to be able to compute the loss.

In [197…
```
model = AutoEncoder(vocab_size, 128, 128)
optimizer = optim.Adam(model.parameters(), lr=0.001)
criterion = nn.CrossEntropyLoss()
print(make_bold("loss per iteration"))
for it in range(300):
    optimizer.zero_grad()
    pred = model(input_seq)
    pred = pred.view(-1,vocab_size)
    input_without_bos = input_seq[:,1:].view(-1)
    loss = criterion(pred, input_without_bos)
    loss.backward()
    optimizer.step()

    if (it+1) % 50 == 0:
        print("[Iter %d] Loss %f" % (it+1, float(loss)))
```

```
loss per iteration
[Iter 50] Loss 0.102723
[Iter 100] Loss 0.026540
[Iter 150] Loss 0.016716
[Iter 200] Loss 0.011711
[Iter 250] Loss 0.008742
[Iter 300] Loss 0.006823
```

Part (c) -- 4%

Once you are satisfied with your model, encode your input using the RNN encoder, and sample some sequences from the decoder. The sampling code is provided to you, and performs the computation from the first diagram (without teacher forcing).

Note that we are sampling from a multi-nomial distribution described by the logits $z^{(t)}$. For example, if our distribution is [80%, 20%] over a vocabulary of two words, then we will choose the first word with 80% probability and the second word with 20% probability.

Call `sample_sequence` at least 5 times, with the default temperature value. Make sure to include the generated sequences in your PDF report.

In [198…
```
def sample_sequence(model, hidden, max_len=20, temperature=1):
    """
    Return a sequence generated from the model's decoder
        - model: an instance of the AutoEncoder model
        - hidden: a hidden state (e.g. computed by the encoder)
        - max_len: the maximum length of the generated sequence
        - temperature: described in Part (d)
    """
    # We'll store our generated sequence here
    generated_sequence = []
    # Set input to the <BOS> token
    inp = torch.Tensor([text_field.vocab.stoi["<bos>"]]).long()
    for p in range(max_len):
        # compute the output and next hidden unit
        output, hidden = model.decode(inp.unsqueeze(0), hidden)
        # Sample from the network as a multinomial distribution
        output_dist = output.data.view(-1).div(temperature).exp()
        top_i = int(torch.multinomial(output_dist, 1)[0])
        # Add predicted word to string and use as next input
        word = text_field.vocab.itos[top_i]
        # Break early if we reach <eos>
        if word == "<eos>":
            break
        generated_sequence.append(word)
        inp = torch.Tensor([top_i]).long()
    return generated_sequence
```

In [199…
```
# Your solutions go here
print(make_bold(make_underline("sample_sequence with  default temperature value:")))
hidden = model.encode(input_seq)
for i in range(5):
  print(make_bold(str(i+1)+")"),sample_sequence(model,hidden))
```

```
sample_sequence with  default temperature value:
1) ['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
2) ['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
3) ['zambian', 'president', 'swears', 'in', 'new', 'in', 'new', 'army', 'chief']
4) ['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
5) ['zambian', 'president', 'swears', 'in', 'new', 'army', 'chief']
```

Part (d) -- 6%

The multi-nomial distribution can be manipulated using the `temperature` setting. This setting can be used to make the distribution "flatter" (e.g. more likely to generate different words) or "peakier" (e.g. less likely to generate different words).

Call `sample_sequence` at least 5 times each for at least 3 different temperature settings (e.g. 1.5, 2, and 5). Explain why we generally don't want the temperature setting to be too **large**.

Explanation: As the tempature rise the multi-nomial distribution is getting "flatter" hence it is more likely to generate different words. We generally don't want the temperature setting to be too large because we would genetrate really different word in randomness and the senteses would not make sense.

In [200…
```
# Include the generated sequences and explanation in your PDF report.
print(make_bold(make_underline("sample_sequence with  different temperature value:")))

hidden = model.encode(input_seq)
for temp in [1.5,2,5]:
  print(make_bold("temp "+str(temp)+":"))
  for i in range(5):
    print(make_bold(str(i+1)+")"),sample_sequence(model,hidden,temperature=temp))
  print("=====================================")
```

```
sample_sequence with  different temperature value:
temp 1.5:
1) ['zambian', 'zambian', 'president', 'swears', 'in', 'new', 'year-on-year', 'hurricane', 'donbass', 'bahn',
 'shape', 'valued', 'homebuilder', '_num_-ryanair', 'denounce', 'chief', 'in', 'packing', 'madoff', 'having']
2) ['zambian', 'debacle', 'in', 'new', 'army', 'fold']
3) ['zambian', 'president', 'swears', 'uncertainties', 'new', 'new', 'army', 'sensor', 'turkey', 'overseas', 'h
ill', 'farmers', 'ocean', 'cpi', 'zambian', 'in', 'dampens', 'tunnel', 'nile', 'prefers']
4) ['zambian', 'president', 'swears', 'in', 'new', 'army', 'producing']
5) ['zambian', 'pardon', 'swears', 'ease', 'collision', 'stoltenberg', 'skid', 'sparking', '1,200', 'chief', 'i
slamists', 'dp', 'woodland', 'everything', 'army', 'mnuchin', 'spotlight', 'roads', '_num_-7', 'carnival']
=====================================
temp 2:
1) ['zambian', 'moon', 'celebrations', ';', 'yonhap', 'epidemic', 'monitors', 'act', 'restructures', 'nationali
st', 'belt', 'stabbings', 'liquidation', 'vuitton', 'anglo', 'nissan', 'shilling', 'drove', 'pre-election', 'fe
in']
2) ['_num_-singapore', 'ousts', 'swears', 'clings', '_num_-intel', 'a321xlr', 'pepsico', 'events', 'successfu
l', 'loading', 'steer', 'amber', 'billion', 'florida', 'blind', 'watching', 'army', 'figures', 'crossing', 'cha
nnels']
3) ['bryant', 'cenbank', 'france-klm', 'frets', 'friendly', 'check', 'triple', 'tornadoes', 'vegas', 'sueddeuts
che', 'grace', 'xerox', 'wildcard', 'defending', 'so', 'markets-asian', 'army', 'chief', 'fighter', 'exec']
4) ['nomura', 'unilever', 'in', 'ola', 'plots', 'sudden', 'recognise', 'list', 'notes', 'driving', 'suu', 'thun
der', 'serbian', 'yearly', 'americas', 'embraer', 'secondary', 'vodafone', 'u.s.-japan', 'infection']
5) ['lied', 'situation', 'hna', '-cftc', 'tornadoes', 'fashion', 'weapon', 'cancer-causing', 'chance', 'disposa
l', 'workers', 'shorter', 'thanks', '_num_-japan', 'table-uae', 'anti-doping', 'spokesman', 'zambian', 'era',
 'davos']
=====================================
temp 5:
1) ['trap', 'where', 'background', 'provinces', 'grew', 'irving', 'scrapped', 'hill', 'technical', "'re", 'nada
l', '_num_-global', 'tea', 'lights', 'heightening', 'heads', 'drove', 'another', 'elections', 'fueling']
2) ['cds', 'met', 'rba', 'fiery', 'protecting', 'synagogue', 'operate', 'largely', '_num_-nike', 'platform', 'p
act', 'sears', 'perform', 'reviews', 'sdf', 'stoking', 'trash', 'cooperate', 'rosneft', 'barneys']
3) ['kushner', 'u.s', 'actress', 'dayton', 'reported', 'evacuates', 'overturn', 'booms', 'relatives', 'prompt
s', 'honor', 'kidnapped', 'extreme', 'extinguished', 'mainline', 'yemen', 'treasurer', 'downgrade', 'replacemen
t', 'killings']
4) ['how', 'northwest', 'companies', 'peru', 'robot', 'bold', '_num_-walmart', 'tyson', 'old', 'sway', 'spacecr
aft', 'nile', 'dull', 'bomb', 'macri', 'pensions', 'salvadoran', 'leviathan', 'texans', 'rematch']
5) ['-state', 'rivals', 'cb', 'prayers', 'hook', '_num_-brexit', 'due', 'turns', 'highlights', '_num_-man', '-c
fo', 's.', 'picket', 'catastrophe', 'drifts', '_num_-southwest', 'huffman', 'zuckerberg', 'pay', 'justices']
=====================================
```

We can tell tempature 5 is too big and indead we get some random words and not a logical sentence.

Question 3. Data augmentation (20%)

It turns out that getting good results from a text auto-encoder is very difficult, and that it is very easy for our model to **overfit**. We have discussed several methods that we can use to prevent overfitting, and we'll introduce one more today: **data augmentation**.

The idea behind data augmentation is to artificially increase the number of training examples by "adding noise" to the image. For example, during AlexNet training, the authors randomly cropped 224×224 regions of a 256×256 pixel image to increase the amount of training data. The authors also flipped the image left/right. Machine learning practitioners can also add Gaussian noise to the image.

When we use data augmentation to train an *autoencoder*, we typically to only add the noise to the input, and expect the reconstruction to be *noise free*. This makes the task of the autoencoder even more difficult. An autoencoder trained with noisy inputs is called a **denoising auto-encoder**. For simplicity, we will *not* build a denoising autoencoder today.

Part (a) -- 5%

We will add noise to our headlines using a few different techniques:

1. Shuffle the words in the headline, taking care that words don't end up too far from where they were initially
2. Drop (remove) some words
3. Replace some words with a blank word (a `<pad>` token)
4. Replace some words with a random word

The code for adding these types of noise is provided for you:

In [201…
```
def tokenize_and_randomize(headline,
                           drop_prob=0.1,  # probability of dropping a word
                           blank_prob=0.1, # probability of "blanking" out a word
                           sub_prob=0.1,   # probability of substituting a word with a random one
                           shuffle_dist=3): # maximum distance to shuffle a word
    """
    Add 'noise' to a headline by slightly shuffling the word order,
    dropping some words, blanking out some words (replacing with the <pad> token)
    and substituting some words with random ones.
    """
    headline = [vocab.stoi[w] for w in headline.split()]
    n = len(headline)
    # shuffle
    headline = [headline[i] for i in get_shuffle_index(n, shuffle_dist)]

    new_headline = [vocab.stoi['<bos>']]
    for w in headline:
        if random.random() < drop_prob:
            # drop the word
            pass
        elif random.random() < blank_prob:
            # replace with blank word
            new_headline.append(vocab.stoi["<pad>"])
        elif random.random() < sub_prob:
            # substitute word with another word
            new_headline.append(random.randint(0, vocab_size - 1))
        else:
            # keep the original word
            new_headline.append(w)
    new_headline.append(vocab.stoi['<eos>'])
    return new_headline

def get_shuffle_index(n, max_shuffle_distance):
    """ This is a helper function used to shuffle a headline with n words,
    where each word is moved at most max_shuffle_distance. The function does
    the following:
       1. start with the *unshuffled* index of each word, which
          is just the values [0, 1, 2, ..., n]
       2. perturb these "index" values by a random floating-point value between
          [0, max_shuffle_distance]
       3. use the sorted position of these values as our new index
    """
    index = np.arange(n)
    perturbed_index = index + np.random.rand(n) * 3
    new_index = sorted(enumerate(perturbed_index), key=lambda x: x[1])
    return [index for (index, pert) in new_index]
```

Call the function `tokenize_and_randomize` 5 times on a headline of your choice. Make sure to include both your original headline, and the five new headlines in your report.

In [202…
```
headline = train_data[42].title
new_headline=' '.join(headline[1:-1])
print(make_bold(make_underline("tokenize_and_randomize on headline: \""+new_headline+"\"")))

for i in range(5):
  print(make_bold(str(i+1)+")"),[vocab.itos[x] for x in tokenize_and_randomize(new_headline)])
```

```
tokenize_and_randomize on headline: "zambian president swears in new army chief"
1) ['<bos>', 'zambian', 'president', 'swears', 'in', 'will', 'chief', 'army', '<eos>']
2) ['<bos>', 'zambian', 'parade', 'swears', 'in', 'army', 'new', 'chief', '<eos>']
3) ['<bos>', 'zambian', 'swears', 'president', 'in', 'new', 'army', 'chief', '<eos>']
4) ['<bos>', '<pad>', 'rail', 'swears', 'in', 'chief', '<eos>']
5) ['<bos>', '<pad>', '<pad>', 'in', '<pad>', 'new', '<pad>', 'army', '<eos>']
```

Part (b) -- 8%

The training code that we use to train the model is mostly provided for you. The only part we left blank are the parts from Q2(b). Complete the code, and train a new AutoEncoder model for 1 epoch. You can train your model for longer if you want, but training tend to take a long time, so we're only checking to see that your training loss is trending down.

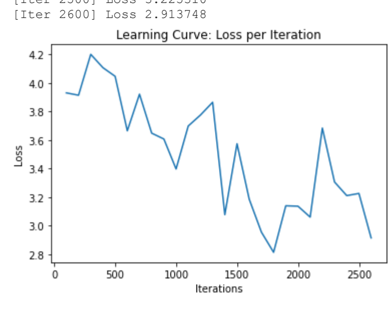
If you are using Google Colab, you can use a GPU for this portion. Go to "Runtime" => "Change Runtime Type" and set "Hardware acceleration" to GPU. Your Colab session will restart. You can move your model to the GPU by typing `model.cuda()`, and move other tensors to GPU (e.g. `xs = xs.cuda()`). To move a model back to CPU, type `model.cpu`. To move a tensor back, use `xs = xs.cpu()`. For training, your model and inputs need to be on the *same device*.

In [203…
```
def train_autoencoder(model, batch_size=64, learning_rate=0.001, num_epochs=10):
    optimizer = optim.Adam(model.parameters(), lr=learning_rate)
    criterion = nn.CrossEntropyLoss()
    losses=[]
    iters=[]
    for ep in range(num_epochs):
        # We will perform data augmentation by re-reading the input each time
        field = data.Field(sequential=True,
                           tokenize=tokenize_and_randomize, # <-- data augmentation
                           include_lengths=True,
                           batch_first=True,
                           use_vocab=False, # <-- the tokenization function replaces this
                           pad_token=vocab.stoi['<pad>'])
        dataset = data.TabularDataset(train_path, "tsv", [('title', field)])
        # This BucketIterator will handle padding of sequences that are not of the same length
        train_iter = data.BucketIterator(dataset,
                                         batch_size=batch_size,
                                         sort_key=lambda x: len(x.title), # to minimize padding
                                         repeat=False)

        for it, ((xs, lengths), _) in enumerate(train_iter):
          xs=xs.cuda()
          optimizer.zero_grad()
          pred = model(xs)
          pred = pred.reshape(-1,vocab_size)
          input_without_bos = xs[:,1:].reshape(-1)
          loss = criterion(pred, input_without_bos)
          loss.backward()
          optimizer.step()
          if (it+1) % 100 == 0:
            iters.append(it)
            losses.append(loss)
            print("[Iter %d] Loss %f" % (it+1, float(loss)))
    return iters,losses
```

In [204…
```
# change name of original signature
def plot_learning_curve(iters, train_losses):
    """
    Plot the learning curve.
    """
    plt.title("Learning Curve: Loss per Iteration")
    plt.plot(iters, train_losses, label="Train")
    plt.xlabel("Iterations")
    plt.ylabel("Loss")
    plt.show()
```

In [205…
```
model = AutoEncoder(vocab_size, 128, 128).cuda()
iters, losses = train_autoencoder(model,num_epochs=1)
plot_learning_curve(iters, losses)
```

```
[Iter 100] Loss 3.929831
[Iter 200] Loss 3.913346
[Iter 300] Loss 4.199826
[Iter 400] Loss 4.106308
[Iter 500] Loss 4.045897
[Iter 600] Loss 3.663965
[Iter 700] Loss 3.920741
[Iter 800] Loss 3.647631
[Iter 900] Loss 3.606497
[Iter 1000] Loss 3.396665
[Iter 1100] Loss 3.697790
[Iter 1200] Loss 3.774533
[Iter 1300] Loss 3.863959
[Iter 1400] Loss 3.076610
[Iter 1500] Loss 3.573408
[Iter 1600] Loss 3.185116
[Iter 1700] Loss 2.954397
[Iter 1800] Loss 2.812865
[Iter 1900] Loss 3.138788
[Iter 2000] Loss 3.135981
[Iter 2100] Loss 3.059474
[Iter 2200] Loss 3.683353
[Iter 2300] Loss 3.305320
[Iter 2400] Loss 3.210059
[Iter 2500] Loss 3.225310
[Iter 2600] Loss 2.913748
```



Part (c) -- 7%

This model requires many epochs (>50) to train, and is quite slow without using a GPU. You can train a model yourself, or you can load the model weights that we have trained, and available on the course website (AE_RNN_model.pk).

Assuming that your `AutoEncoder` is set up correctly, the following code should run without error.

In [206…
```
model = AutoEncoder(10000, 128, 128)

checkpoint_path = '/content/gdrive/MyDrive/Intro_to_Deep_Learning/assignment4/AE_RNN_model.pk' # Update me
model.load_state_dict(torch.load(checkpoint_path))
```

Out[206…
```
<All keys matched successfully>
```

Then, repeat your code from Q2(d), for `train_data[10].title` with temperature settings 0.7, 0.9, and 1.5. Explain why we generally don't want the temperature setting to be too **small**.

In [207…
```
# Include the generated sequences and explanation in your PDF report.

headline = train_data[10].title
input_seq = torch.Tensor([vocab.stoi[w] for w in headline]).unsqueeze(0).long()
print(make_bold("input: \""+' '.join(headline)+"\""))
hidden = model.encode(input_seq)
for temp in [0.7,0.9,1.5]:
  print(make_bold(make_underline("temp "+str(temp)+":")))
  for i in range(5):
    print(make_bold(str(i+1)+")"),sample_sequence(model,hidden,temperature=temp))
  print("==============================")
```

```
input: "<bos> wall street rises , limps across the finish line of a turbulent year <eos>"
temp 0.7:
1) ['wall', 'street', 'rises', ',', 'limps', 'across', 'the', 'finish', 'line', 'of', 'a', 'turbulent', 'year']
2) ['wall', 'street', 'rises', ',', 'limps', 'across', 'the', 'finish', 'line', 'of', 'a', 'turbulent', 'year']
3) ['wall', 'street', 'rises', ',', 'limps', 'open', 'across', 'woman', '_num_-opec', 'to', 'september', 'fir
e', 'finds']
4) ['wall', 'street', 'rises', ',', 'limps', 'die', 'win', 'at', 'of', 'sciences', '<pad>', 'presidential', 'am
id']
5) ['wall', 'street', 'rises', ',', 'limps', 'dead', 'anthem', 'with', '<pad>', "'s", 'change', 'for', 'singapo
re']
==============================
temp 0.9:
1) ['wall', 'street', 'rises', ',', 'limps', 'heist', 'join', 'searches', '_num_', 'deal', 'clear', 'after', 's
ight']
2) ['wall', 'street', 'rises', ',', 'limps', 'open', 'detroit', ',', 'circus', 'says', 'share', ':', 'finds']
3) ['wall', 'street', 'rises', ',', 'limps', 'die', 'win', 'at', 'of', 'sciences', '<pad>', 'four', 'war']
4) ['wall', 'street', 'rises', ',', 'scales', 'finish', 'dead', 'on', 'trash', 'hong', 'norway', 'whole', 'cor
e']
5) ['wall', 'street', 'rises', ',', 'limps', 'form', 'atlantia', 'after', 'to', 'moratorium', '<pad>', 'year',
 'turbulent']
==============================
temp 1.5:
1) ['wall', 'street', 'rises', ',', 'jd.com', 'passage', 'australia', 'after', 'primed', '<pad>', 'three', 'ind
ia', 'narrows']
2) ['wall', 'street', 'rises', ',', 'kurz', 'bottleneck', 'banks', 'in', 'gilead', 'block', 'fall', 'u.s.', 'ab
ove']
3) ['wall', 'street', 'rises', ',', 'ditches', 'longer', 'australia', 'families', 'in', 'energy', 'after', 'isr
aelis', 'another']
4) ['wall', 'street', 'softbank', 'rally', 'socialists', 'floating', 'form', 'in', '<pad>', 'raises', 'future
s', 'northern', 'year']
5) ['wall', 'street', 'sticks', 'exposure', 'modi', 'science', '<unk>', 'ways', ',', 'party', '<pad>', 'investm
ents', 'four', 'land']
==============================
```

Here for tempature 0.7 and 0.9 we get almost the same sentence as the original inputs- hence the tempature is too low.

Write your explanation here: As the tempature decrease the distribution becomes peakier . for tempature too high we will get a "delta function" and the generation would me almost deterministic and we would just memorize the input sentences.

Question 4. Latent space manipulations (20%)

In parts 2-3, we've explored the decoder portion of the autoencoder. In this section, let's explore the **encoder**. In particular, the encoder RNN gives us embeddings of news headlines!

First, let's load the **validation** data set:

In [208…
```
valid_data = data.TabularDataset(
    path=valid_path,                # data file path
    format="tsv",                   # fields are separated by a tab
    fields=[('title', text_field)]) # list of fields (we have only one)
```

Part (a) -- 4%

Compute the embeddings of every item in the validation set. Then, store the result in a single PyTorch tensor of shape `[19046, 128]`, since there are 19,046 headlines in the validation set.

In [209…
```
# Write your code here
# Show that your resulting PyTorch tensor has shape `[19046, 128]`
model.cuda()
embeddings=[]
for xs in valid_data:
  headline=xs.title
  input_seq = torch.Tensor([vocab.stoi[w] for w in headline]).unsqueeze(0).long().cuda()
  encoding = model.encode(input_seq).detach().squeeze(0)
  embeddings.append(encoding)

word_emb = torch.cat(embeddings, dim=0)
print(word_emb.size())
```

```
torch.Size([19046, 128])
```

Part (b) -- 4%

Find the 5 closest headlines to the headline `valid_data[13]`. Use the cosine similarity to determine closeness. (Hint: You can use code from assignment 2)

In [217…
```
def print_five_closet(example_num,word_emb):
  headline = valid_data[example_num].title
  input_seq = torch.Tensor([vocab.stoi[w] for w in headline]).unsqueeze(0).long()
  encoding = model.encode(input_seq).detach().squeeze(0)

  word_emb=word_emb.cpu()
  norms = np.linalg.norm(word_emb, axis=1)
  emb_headlines_norm = (word_emb.T / norms).T
  enc_norm = encoding.numpy() / np.linalg.norm(encoding.numpy(), axis=1)
  similarities = np.matmul(enc_norm, emb_headlines_norm.T)
  five_max_indices = (-similarities.squeeze()).argsort()[1:6]

  # print
  print(make_underline(make_bold("5 closetst headlines to \""+' '.join(valid_data[example_num].title)+"\":\n"))
  for i,idx in enumerate(five_max_indices):
    print(make_bold(str(i+1)+")"),make_bold("similarity:"),similarities[0][idx].numpy(),make_bold(" || headline
```

```
model.cpu()
print_five_closet(13,word_emb)
```

5 closetst headlines to "<bos> asia takes heart from new year gains in u.s. stock futures <eos>":

```
1) similarity: 0.9308761  || headline: <bos> italy 's salvini loses aura of invincibility in emilia setback <eos>
2) similarity: 0.930557  || headline: <bos> saudi , russia look to seal deeper output cuts with oil producers <eos>
3) similarity: 0.9298552  || headline: <bos> eu orders quarantine for staff who traveled to northern italy <eos>
4) similarity: 0.92876196  || headline: <bos> update _num_-italy 's prime minister says new government will bicker less <eos>
5) similarity: 0.9280871  || headline: <bos> portugal 's moura pays tribute to cod fishermen at milan fashion close <eos>
```

Part (c) -- 4%

Find the 5 closest headlines to another headline of your choice.

```
print_five_closet(17,word_emb)
```

5 closetst headlines to "<bos> indian manufacturing growth slows in december despite price cuts : pmi <eos>":

```
1) similarity: 0.94242764  || headline: <bos> indian oil plans to shut units at northeast refineries to upgrade fuel : sources <eos>
2) similarity: 0.93822354  || headline: <bos> indian authorities recover bodies of seven climbers from mountain , one still missing <eos>
3) similarity: 0.9336342  || headline: <bos> indian economic growth drops to just _num_ % in jan-march quarter , falls behind china <eos>
4) similarity: 0.9303627  || headline: <bos> china first-quarter gdp growth steady at _num_ percent , beats expectations for slowdown <eos>
5) similarity: 0.9295902  || headline: <bos> china 's first-quarter smartphone sales may halve due to coronavirus : analysts <eos>
```

We can tell that those sentences are indeed close because the original headline has indian as the subject , and the 3 most similiar headlines also had Indian as subject. The next closest headline is with China insteas of indian, when China is also a state to it has similar meanings.

Part (d) -- 8%

Choose two headlines from the validation set, and find their embeddings. We will **interpolate** between the two embeddings like we did in the example presented in class for training autoencoders on MNIST.

Find 3 points, equally spaced between the embeddings of your headlines. If we let e_0 be the embedding of your first headline and e_4 be the embedding of your second headline, your three points should be:

$$e_1 = 0.75e_0 + 0.25e_4$$
$$e_2 = 0.50e_0 + 0.50e_4$$
$$e_3 = 0.25e_0 + 0.75e_4$$

Decode each of e_1, e_2 and e_3 five times, with a temperature setting that shows some variation in the generated sequences. Try to get a logical and cool sentence (this might be hard).

```
# Write your code here. Include your generated sequences.
headlines = [valid_data[13].title,valid_data[17].title]
input_seqs = [torch.Tensor([vocab.stoi[w] for w in headlines[0]]).unsqueeze(0).long(),
             torch.Tensor([vocab.stoi[w] for w in headlines[1]]).unsqueeze(0).long()]
e = [model.encode(input_seqs[0]).detach(),
            model.encode(input_seqs[1]).detach()]

# print(make_bold(make_underline("input e"+str(i+1)+":")))

e1 = 0.75*e[0]+0.25*e[1]
e2 = 0.5*e[0]+0.5*e[1]
e3 = 0.25*e[0]+0.75*e[1]

encs=[e1,e2,e3]
temp=1.5
for i,enc in enumerate(encs):
  print(make_bold(make_underline("input e"+str(i+1)+":")))
  print("temp "+str(temp)+":")
  for i in range(5):
    print(make_bold(str(i+1)+")"),sample_sequence(model,enc,temperature=temp))
  print("=============================")
```

input e1:

```
temp 1.5:
1) ['asia', 'utah', 'april', 'month/month', 'fall', '<unk>', 'third', 'field', 'due', 'to', 'oct']
2) ['asia', 'takes', 'three', 'profit', 'from', 'week', 'russian', 'incentive', 'asia', 'to', 'rees-mogg', 'half']
3) ['markets', 'billion-dollar', 'to', 'over', '%', 'prices', 'media', 'crossing', 'oil', 'shares', 'funding']
4) ['virus', "'s", 'bankruptcy', 'bet', 'in', 'risk', 'three', 'new', 'retailing', 'up', 'dips']
5) ['asia', 'takes', 'than', 'america', 'from', 'first', 'cash', 'heads', 'time', 'deal', 'stocks-tsx']
=============================
```

input e2:

```
temp 1.5:
1) ['vietnam', 'sell', 'after', '_num_-1/2-month', 'in', '_num_', 'activity', '3rd', '<unk>', 'still', 'oct']
2) ['more', 'test', 'slump', 'cleanup', 'fall', 'general', "'s", 'slowing', 'capital', 'one-year', 'wall']
3) ['indian', 'monthly', ':', 'in', 'curbs', 'profit', 'big', 'kenya', '-sources', 'offsets', 'stocks']
4) [',', 'iphone', 'ghosn', 'surge', 'since', 'against', 'reuters', 'may', 'weekend', 'since', 'shift']
5) ['second', 'markets-stocks', 'orders', 'coast', 'below', 'drone', 'south', 'sector', 'financing', 'iran', 'prices']
=============================
```

input e3:

```
temp 1.5:
1) ['indian', 'markets-forint', 'sales', 'job', 'on', 'dubai', 'due', 'five', 'hopes', 'pmis', 'of', 'hovers']
2) ['indian', 'income', 'weakening', '_num_', 'in', 'in', 'reuters', 'despite', 'hopes', 'up', 'value']
3) ['indian', 'impacted', 'growth', 't20', 'in', 'revenues', ':', 'commodities', 'than', 'weak', '.']
4) ['indian', 'impacted', 'purchase', 'in', 'year', 'fall', '_num_-month', '<pad>', 'interest', "'s", 'this', 'steep']
5) ['indian', 'stocks', 'products', 'game', 'after', 'due', 'about', 'factory', '<pad>', 'fy', 'as']
=============================
```