# CSC413 Assignment 3. Text Denoising Autoencoder for News Headlines

**Deadline**: TBD

**Submission**: Submit a PDF report containing your code, outputs, and your written solutions. You may export the completed notebook, but if you do so **it is your responsibly to make sure that your code and answers do not get cut off**.

**Late Submission**: Please see the syllabus for the late submission criteria.

**Working with a partner**: You may work with a partner for this assignment. If you decide to work with a partner, please create your group on Markus by the deadline date, even if you intend to use grace tokens. Markus does not allow you to create groups past the deadline, even if you have grace tokens remaining.

In this assignment, we'll explore a more advanced use of deep learning on a natural language process (NLP) task involving news headlines. In particular, we'll be working with a dataset of Reuters news headlines collected over a span of 15 months, covering some of 2018, 2019, and early 2020. This assignment will combine several of the concepts that we discussed in class, including recurrent neural networks, data augmentation, autoencoders (soon), and working with embeddings.

To be more specific, we'll be building an **autoencoder** of news headlines. We will build an **encoder** model that maps a news headline to a vector embedding, and a **decoder** that reconstructs the news headline. By building a model that learns to reconstruct the news headlines from the vector embedding, the model will learn good embeddings of these headlins.

We'll see a similar idea with image autoencoders and image VAEs, but both our encoder and decoder networks will be Recurrent Neural Networks. You'll have a chance build networks that takes a sequence as an input, and a network that generates a sequence as an output.

This project is organized as follows:

- Question 1. Data exploration
- Question 2. Background Math
- Question 3. Building the autoencoder
- Question 4. Training the autoencoder using *data augmentation*
- Question 5. Analyzing the embeddings (interpolating between headlines)
- Question 6. Work Allocation

Much of the idea behind this assignment is movtivated by Shen et al [1]. We'll use the data augmentation rules proposed in that work to improve the robustness of the autoencoder.

[1] Shen et al (2019) "Educating Text Autoencoders: Latent Representation Guidance via Denoising" https://arxiv.org/pdf/1905.12777.pdf

```
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

%matplotlib inline
```

## Question 1

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:

```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
train_path = '/content/gdrive/My Drive/CSC321/reuters_train.txt' # Update me
valid_path = '/content/gdrive/My Drive/CSC321/reuters_valid.txt' # Update me
```

We will be using PyTorch's `torchtext` utilities to help us load, process, and batch the data. This package is useful, but takes a bit of time to get used to.

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.

```
import torchtext

# 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 = torchtext.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 in nn.RNN demonstrated in lecture
    use_vocab=True)             # to turn each character into an integer index
train_data = torchtext.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) – 2 points

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.

```
# Include your histogram and your written explanations

# Here is an example of how to plot a histogram in matplotlib:
# plt.hist(np.random.normal(0, 1, 40), bins=20)

# Here are some sample code that uses the train_data object:
print(train_data[5].title)
for example in train_data:
    print(example.title)
    break
```

### Part (b) – 2 points

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

```
# Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values

# You might find the python class Counter from the collections package useful
from collections import Counter
```

### Part (c) – 2 points

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?

```
# Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values
```

### Part (d) – 2 points

Explain why we may wish to replace these infrequent words with an `<unk>` tag, instead of learning embeddings for these rare words. (Hint: Consider words in the validation set that might not appear in training)

```
# Include your explanation here
```

### Part (e) – 2 points

We will only model the top 9995 words in the training set, excluding the tags `<bos>`, `<eos>`, and other possible tags we haven't mentioned yet (including those, we will have a vocabulary size of exactly 10000 tokens).

What percentage of word occurrences will be supported? Alternatively, what percentage of word occurrences in the training set will be set to the `<unk>` tag?

```
# Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values
```

Our `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. You can interpret these indices as sparse representations of one-hot vectors.

```
# 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 is actually 10000
vocab_size = len(text_field.vocab.stoi)
print(vocab_size) # should be 10000

# The reason is that torchtext adds two more tokens 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
```

## Question 2

Choosing the right model architecture is key for any successful deep learning system. In this question, we will compare the learning performance of RNNs and GRUs from the perspective of the vanishing/exploding gradient problem that arises during backpropagation.

## Part (a) – 4 pts

First, we will analyze the recurrent weight matrix of an RNN using Singular Value Decomposition (SVD). SVD says that that any real matrix $M \in \mathbb{R}^{m \times n}$ can be written as $W = U\Sigma V^T$ where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are square orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times n}$ is a rectangular diagonal matrix. Recall that the values of $\Sigma$ are the *eigenvalues* of $M$.o

(For a quick overview of eigenvalues and eigenvectors, see https://www.youtube.com/watch?v=PFDu9oVAE-g 0:44-5:22 and 13:05-end. The last section explains visually why we are interested in working with eigenvalues when working with an RNN)

Consider a simple simple RNN-like architecture that computes $x_{t+1} = sigmoid(Wx_t)$. You can view this architecture as a deep fully connected network that uses the same weight matrix at each layer. Suppose the largest singular value of the weight matrix is $\sigma_{max}(W) = \frac{1}{2}$.

Show that the largest singular value of the input-output Jacobian has the following bound: $0 \leq \sigma_{max}(\frac{\partial x_n}{\partial x_1}) \leq (\frac{1}{2})^n$. *Hint*: if $C = AB$, then $\sigma_{max}(C) \leq \sigma_{max}(A)\sigma_{max}(B)$. Also, the input-output Jacobian is the multiplication of layerwise Jacobians).

What does this tell us about the input-output Jacobian $\frac{\partial x_n}{\partial x_1}$ as $n \to \infty$ ?

## Part (b) – 4 pts

We will now compare the gradients of a vanilla RNN unit and a Gated Recurrent Unit (GRU). For both parts (b) and (c), assume that all weights are scalars, and that we have an input sequence of length $T$ with $x_1 = 1$ and $x_t = 0$ for all other $t$. Also, assume that $h_0 = 0$ and that after $T$ timesteps, we calculate the squared loss $L = \frac{1}{2}(y_T - o_T)^2$ where $o_T$ is the target at timestep $T$.

Consider the vanilla RNN units that compute $h_t$ at each timestep $t$ as follows:

$$m_t = W_x x_t + W_h h_{t-1}$$
$$h_t = \tanh(m_t)$$
$$y_t = W_y h_t$$

Compute $\frac{\partial L}{\partial W_x}$ using backpropagation. You should obtain an expression in terms of the quantities given (like $o_T$, $y_T$, etc. . . )

Do you see a vanishing gradient problem? What about exploding gradient? Explain.

## Part (c) – 4 pts

Now, let's consider GRU units that uses a gating mechanism, and computes $h_t$ at each timestep $t$ as follows:

$$z_t = \sigma(W_x x_t + U_z h_{t-1})$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$
$$\hat{h}_t = \tanh(W_h x_t + U_h r_t h_{t-1})$$
$$h_t = (1 - z_t)h_{t-1} + z_t \hat{h}_t$$
$$y_t = W_y h_t$$

Where $\sigma$ is the sigmoid function.

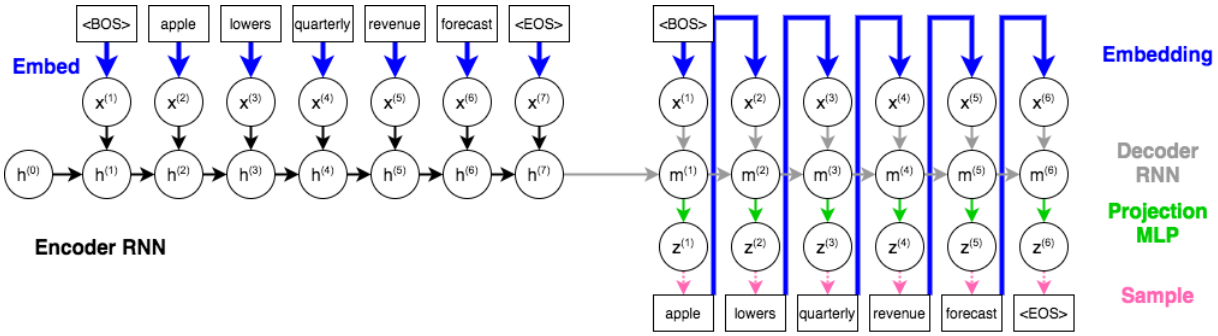
Compute $\frac{\partial L}{\partial W_x}$ using backpropagation.

Can the vanishing gradient problem be prevented? *Hint* : Consider the term $\frac{\partial h_t}{\partial h_{t-1}}$, what role does $z_t$ play in this gradient? Can it help alleviate the vanishing gradient problem?

## Question 3

Building a text autoencoder is a little more complicated than an image autoencoder, so we'll need to thoroughly understand the model that we want to build before actually building our model. Note that the best and fastest way to complete this assignment is to spend a *lot* of time upfront understanding the architecture. The explanations are quite dense, and you might need to stop every sentence or two to understand what's going on. You won't feel productive for a while since you won't be writing code, but this initial investment will help you become more productive later on. Understanding this architecture will also help you understand other machine learning papers you might come across. So, take a deep breath, and let's do this!

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 $K = 10000$ distinct words. As in a classification problem, the fully-connected neural network will compute a *probability distribution* over these 10,000 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 4. 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.

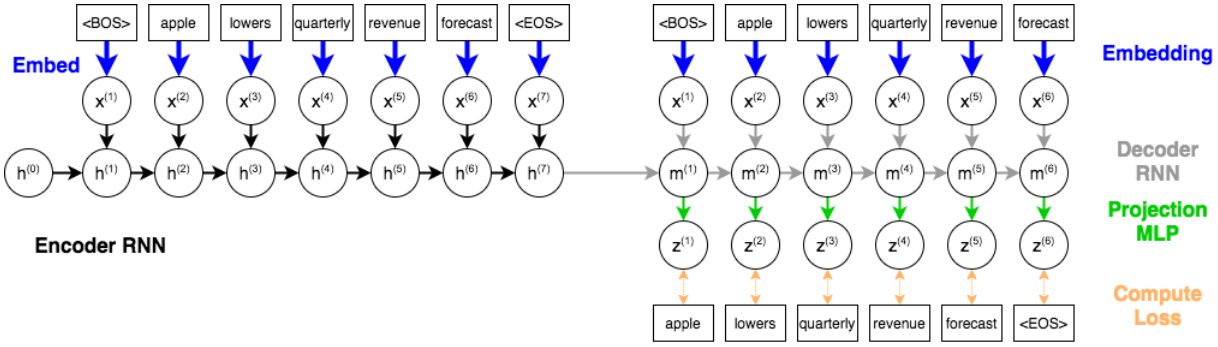
Unfortunately, we can't *train* this autoencoder in the way we just described. That is, we can't just compare our generated sequence with our ground-truth sequence, and get gradients. Both sequences are **discrete** entities, so we won't be able to compute gradients at all! In particular, **sampling is a discrete process**, and so we won't be able to back-propagate through any kind of sampling that we do.

You might wonder whether we can get away with computing gradients by comparing the distributions $\mathbf{z}^{(t)}$ with the ground truth words at each time step. Like any multi-class classification problem, we can represent the ground-truth words as a one-hot vector, and use the cross-entropy loss.

In theory, we can do this. In practice, there are a few issues. 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 forgo sampling, and use the **ground truth** $\mathbf{x}^{(t)}$ 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 $t$. (This is similar to what we did in a pixel-wise prediction problem.)

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, or initialize our parameters to GloVe embeddings. To prevent students who don't have Colab access from having to download a 1GB file, we won't do that. The word embedding component is represented with blue arrows in the diagram.
- The encoder RNN (which will use Gated Recurrent Units) 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 Gated Recurrent Units) 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** (one fully-connected layer) that computes a distribution over the next word to generate, given a decoder RNN hidden state.

## Part (a) – 8 pts

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`. (4 points)
2. Complete the `forward` method, which uses teacher forcing and computes the logits $z^{(t)}$ of the reconstruction of the sequence. (4 points)

You should first try to understand the `encode` and `decode` methods, which are written for you. The `encode` method mimics a discriminative RNN (see the sentiment analysis notebook). The `decode` method is a generative RNN and is a bit more complex (see the text generation tutorial notebook). 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 very careful of the input that you feed into ether `decode` or to `self.decoder_rnn`. Refer to the teacher-forcing diagram.

```
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=None, # TODO
                                  embedding_dim=None)  # TODO
        self.encoder_rnn = nn.GRU(input_size=None, #TODO
                                  hidden_size=None, #TODO
                                  batch_first=True)
        self.decoder_rnn = nn.GRU(input_size=None, #TODO
                                  hidden_size=None, #TODO
                                  batch_first=True)
        self.proj = nn.Linear(in_features=None, # TODO
                              out_features=None) # TODO

    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.
        """

        # TODO
```

**Part (b) – 5 pts**

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:

```
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 very careful of off-by-ones.**

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

```
model = AutoEncoder(vocab_size, 128, 128)
optimizer = optim.Adam(model.parameters(), lr=0.001)
criterion = nn.CrossEntropyLoss()

for it in range(300):

    # TODO

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

### Part (c) – 2 pt

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.

```
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
```

```
# Your solutions go here
```

### Part (d) – 3 pt

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**.

```
# Include the generated sequences and explanation in your PDF report.
```

## Question 4

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 $224 \times 224$ regions of a $256 \times 256$ pixel image to increase the amount of training data. The authors also flipped the image left/right (but not up/down—why?). 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) – 3pt

Give three more examples of data augmentation techniques that we could use if we were training an **image** autoencoder. What are different ways that we can change our input?

```
# Include your three answers
```

### Part (b) – 2pt

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:

```
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.

```
# Report your values here. Make sure that you report the actual values,
# and not just the code used to get those values
```

**Part (c) – 3 pt**

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*.

```
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()

    for ep in range(num_epochs):
        # We will perform data augmentation by re-reading the input each time
        field = torchtext.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 = torchtext.data.TabularDataset(train_path, "tsv", [('title', field)])

        # This BucketIterator will handle padding of sequences that are not of the same length
        train_iter = torchtext.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):

            # Fill in the training code here

            if (it+1) % 100 == 0:
                print("[Iter %d] Loss %f" % (it+1, float(loss)))

        # Optional: Compute and track validation loss
        #val_loss = 0
        #val_n = 0
        #for it, ((xs, lengths), _) in enumerate(valid_iter):
        #    zs = model(xs)
        #    loss = None # TODO
        #    val_loss += float(loss)

# Include your training curve or output to show that your training loss is trending down
```

**Part (d) – 2 pt**

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 https://www.cs.toronto.edu/~lczhang/321/files/p4model.pk (11MB).

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

```
model = AutoEncoder(10000, 128, 128)
checkpoint_path = '/content/gdrive/My Drive/CSC321/p4model.pk' # Update me
model.load_state_dict(torch.load(checkpoint_path))
```

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**.

```
# 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()

# ...
```

## Question 5

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:

```
valid_data = torchtext.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) – 2 pt**

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.

```
# Write your code here
# Show that your resulting PyTorch tensor has shape `[19046, 128]`
```

**Part (b) – 2 pt**

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

```
# Write your code here. Make sure to include the actual 5 closest headlines.
```

**Part (c) – 2 pt**

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

```
# Write your code here.
# Make sure to include the original headline and the 5 closest headlines.
```

**Part (d) – 4 pts**

Choose two headlines from the validation set, and find their embeddings. We will **interpolate** between the two embeddings.

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, while generating sequences that makes sense.

```
# Write your code here. Include your generated sequences.
```

## Question 6. Work Allocation – 2 pts

This question is to make sure that if you are working with a partner, that you and your partner contributed equally to the assignment.

Please have each team member write down the times that you worked on the assignment, and your contribution to the assignment.

```
# Your answer goes here
```