This assignment is split into two sections: *Neural Machine Translation with RNNs* and *Analyzing NMT Systems*. The first is primarily coding and implementation focused, whereas the second entirely consists of written, analysis questions. If you get stuck on the first section, you can always work on the second. That being said, the NMT system is more complicated than the neural networks we have previously constructed within this class and takes about **4 hours to train on a GPU**. Thus, we strongly recommend you get started early with this assignment. Finally, the notation and implementation of the NMT system is a bit tricky, so if you ever get stuck along the way, please come to Office Hours so that the TAs can support you.

1 Neural Machine Translation with RNNs (45 points)

(a)(b)(c)(d)(e)(f)

Please see corresponding code.

(g)

It sets the value at all positions whose corresponding value is 1 in 'masks' to -inf.
Because $\exp^{-\inf} = 0$, so during attention these dummy hiddens won't affect the final result of attention outputs.

(h)(i)

Please see the code.

(j)

- **dot product attention**
  Advantages: No additional parameters, easy to compute.
  Disadvantages: Requires s_t and h_i to have the same length.
- **multiplicative product attention**
  Advantages: Can combine two vectors with different lengths.
  Disadvantages: Requires additional parameter matrix W
- **multiplicative product attention**
  Advantages: Provides more parameters for training
  Disadvantages: Overhead caused by such additional parameters

2 Analyzing NMT Systems (30 points)

(a) (12 points)

Here we present a series of errors we found in the outputs of our NMT model (which is the same as the one you just trained). For each example of a Spanish source sentence, reference (i.e., 'gold') English translation, and NMT (i.e., 'model') English translation, please:

1 Identify the error in the NMT translation.

2 Provide a reason why the model may have made the error (either due to a specific linguistic construct or specific model limitations).

3 Describe one possible way we might alter the NMT system to fix the observed error

Below are the translations that you should analyze as described above. Note that out-of-vocabulary words are underlined.

i. (2 pts) Specific linguistic construct

ii. (2 pts) Specific linguistic construct

iii. (2 pts) Specific model limitations (cannot understand unknown word)

iv. (2 pts) Specific model limitations (custom expression)

v. (2 pts) Specific model limitations (attention bias)

vi. (2 pts) Specific model limitations (unit conversion error)

(b) (4 points)

(c) (14 points)

BLEU Score is the most commonly used automatic evaluation metric for NMT systems. It is usually calculated across the entire test set, but here we will consider BLEU defined for a single example.3 Suppose we have a source sentence s, a set of k reference translations **r1,...,rk**, and a candidate translation **c**. To compute the BLEU score of **c**, we first compute the modified n-gram precision pn of c, for each of n = 1,2,3,4:

i. (5 points) Please consider this example:
Source Sentence **s**: **el amor todo lo puede**
Reference Translation r1: love can always find a way
Reference Translation r2: love makes anything possible
NMT Translation c1: the love can always do
NMT Translation c2: love can make anything possible

For c_1 and $n = 1$:

1-gram	$Count_{r_1}(1-gram)$	$Count_{r_2}(1-gram)$	$Count_c(1-gram)$
the	0	0	1
love	1	1	1
can	1	0	1
always	1	0	1
do	0	0	1

$$p_1^{c_1} = \frac{0+1+1+1+0}{5} = 0.6 \tag{1}$$

For c_1 and $n = 2$:

2-gram	$Count_{r_1}(1-gram)$	$Count_{r_2}(1-gram)$	$Count_c(1-gram)$
the love	0	0	1
love can	1	0	1
can always	1	0	1
always do	0	0	1

$$p_2^{c_1} = \frac{0+1+1+0}{4} = 0.5 \tag{2}$$

$$c = 5 \tag{3}$$

$$r^* = 4 \tag{4}$$

$$BP = 1 \tag{5}$$

$$BLEU^{c_1} = BP \times \exp(\sum_{n=1}^{4} \lambda_n \log p_n) = \exp(0.5 \log 0.6 + 0.5 \log 0.5) = 0.7699 \tag{6}$$

For c_2 and $n = 1$:

1-gram	$Count_{r_1}(1-gram)$	$Count_{r_2}(1-gram)$	$Count_c(1-gram)$
love	1	1	1
can	1	0	1
make	0	0	1
anything	0	1	1
possible	0	1	1

$$p_1^{c_2} = \frac{1+1+0+1+1}{5} = 0.8 \tag{7}$$

For c_2 and $n = 2$:

2-gram	$Count_{r_1}(1-gram)$	$Count_{r_2}(1-gram)$	$Count_c(1-gram)$
love can	1	0	1
can make	0	0	1
make anything	0	0	1
anything possible	0	1	1

$$p_2^{c_2} = \frac{1+0+0+1}{4} = 0.5 \tag{8}$$

$$c = 5 \tag{9}$$

$$r^* = 4 \tag{10}$$

$$BP = 1 \tag{11}$$

$$BLEU^{c_2} = BP \times \exp(\sum_{n=1}^{4} \lambda_n \log p_n) = \exp(0.5 \log 0.8 + 0.5 \log 0.5) = 0.8196 \tag{12}$$

Because $BLEU^{c_1} < BLEU^{c_2}$, c_2 is considered the better translation.

ii. (5 points)

$$p_1^{c_1} = \frac{0+1+1+1+0}{5} = 0.6 \tag{13}$$

$$p_2^{c_1} = \frac{0+1+1+0}{4} = 0.5 \tag{14}$$

$$c = 5 \tag{15}$$

$$r^* = 4 \tag{16}$$

$$BP = 1 \tag{17}$$

$$BLEU^{c_1} = BP \times \exp(\sum_{n=1}^{4} \lambda_n \log p_n) = \exp(0.5 \log 0.6 + 0.5 \log 0.5) = 0.7699 \tag{18}$$

$$p_1^{c_2} = \frac{1+1+0+0+0}{5} = 0.4 \tag{19}$$

$$p_2^{c_2} = \frac{1+0+0+1}{4} = 0.5 \tag{20}$$

$$c = 5 \tag{21}$$

$$r^* = 4 \tag{22}$$

$$BP = 1 \tag{23}$$

$$BLEU^{c_1} = BP \times \exp(\sum_{n=1}^{4} \lambda_n \log p_n) = \exp(0.5 \log 0.4 + 0.5 \log 0.5) = 0.7051 \tag{24}$$

This time c_1 becomes the better translation. I don't agree it's a better translation.

iii. (2 points)
Because for each translation there are many possible reference results. However it can never ensure that the given reference is the best expression. Besides, a better translation should be similar to most of referenced translation from intuition but may look unlike a few of them. Increasing the number of reference translations can assure the accuracy of BLEU score.

iv. (2 points)
Advantages:

* Pure mathematical definition, which is convenient to calculate and can save human efforts
* Mathematically defined, an objective criteria that rules out human factors

Disadvantages:

* Machine based criteria, which can cause deviation from real human languages, i.e. may give some bad translation a high score
* Mathematical based criteria that contains many parameters to tune, for different tasks and models we need different parameter settings, causing it difficult to use.