Stanford CS 224n Assignment 4

Akash Gupta

September 5, 2020

1 Neural Machine Translation with RNNs (45 points)

In Machine Translation, our goal is to convert a sentence from the source language (e.g. Spanish) to the target language (e.g. English). In this assignment, we will implement a sequence-to-sequence (Seq2Seq)network with attention, to build a Neural Machine Translation (NMT) system. In this section, we describe the training procedure for the proposed NMT system, which uses a Bidirectional LSTM Encoder and a Unidirectional LSTM Decoder.

(g) (3 points) (written) The `generate_sent_masks()` function in `nmt_model.py` produces a tensor called `enc_masks`. It has shape (batch size, max source sentence length) and contains 1s in positions corresponding to 'pad' tokens in the input, and 0s for non-pad tokens. Look at how the masks are used during the attention computation in the `step()` function (lines 295-296). First explain (in around three sentences) what effect the masks have on the entire attention computation. Then explain (in one or two sentences) why it is necessary to use the masks in this way.

Answer: The input sentences are padded to equal lengths with a token that contains no information. Thus a mask is used to mark the locations of the padding. The `step()` function sets the padded locations in the attention vector e_t to $-\inf$ to zero out those locations in the attention distribution α_t after calculation.

(i) (4 points) Once your model is done training (this should take about 4 hours on the VM),execute the following command to test the model: `sh run.sh test`. Please report the model's corpus BLEU Score. It should be larger than 21

Answer: We left the model training locally on RTX 2080 Ti overnight. It hit early stopping at epoch 14. Test Corpus BLEU is 35.89

(j) (3 points) (written) In class, we learned about dot product attention, multiplicative attention, and additive attention. Please explain one advantage and one disadvantage of dot product attention compared to multiplicative attention. Then explain one advantage and one disadvantage of additive attention compared to multiplicative attention. As a reminder, dot product attention is $e_{t,i} = s_t^{\intercal} h_i$, multiplicative attention is $e_{t,i} = s_t^{\intercal} W hi$, and additive attention is $e_{t,i} = v \tanh(W_1 h_i + W_2 s_t)$

Answer:

	Advantages	**Disadvantages**
Dot Product	Simple, easy to calculate. No extra w variables No extra w variables	Assumes s_t and h_i in same dimension. Returns a scalar thus less information
Multiplicative	s_t and h_i don't have to be in same dimension. more representation than dot product	Costly in high dimension New W parameter
Additive	Similar computational complexity as Multiplicative Better computational efficiency in high dimensions	W_1 and W_2 parameters Dimension is another hyperparameter.

Reference: Sebatian Ruder's blog post and Lilian Weng's blog post

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 possible reason(s) why the model may have made the error (either due to a specific linguistic construct or a specific model limitation).
3. Describe one possible way we might alter the NMT system to fix the observed error. There are more than one possible fixes for an error. For example, it could be tweaking the size of the hidden layers or changing the attention mechanism.

Below are the translations that you should analyze as described above. Note that out-of-vocabulary words are underlined. Rest assured that you don't need to know Spanish to answer these questions. You just need to know English! The Spanish words in these questions are similar enough to English that you can mostly see the alignments. If you are uncertain about some words, please feel free touse resources like Google Translate to look them up.

i (2 points) **Source Sentence:** *Aqu ı otro de mis favoritos, "La noche estrellada".*
Reference Translation: *So another one of my favorites, "The Starry Night".*
NMT Translation: *Here's another favorite of my favorites, "The Starry Night".*

Answer:
Error: *favorite of my favorites.*
Reason: Model limitations?
Fix: More training data of similar sentence structures.

ii (2 points) **Source Sentence:** *Ustedes saben que lo que yo hago es escribir para los ni nos, y ,de hecho, probablemente soy el autor para ni nos, ms ledo en los EEUU.*
Reference Translation: *You know, what I do is write for children, and I'm probably America's most widely read children's author, in fact.*
NMT Translation: *You know what I do is write for children, and in fact, I'm probably the author for children, more reading in the U.S.*

Answer:
Error: *more reading in the US* didnt capture the relationships.
Reason: Model limitation. It didn't capture relationships between punctuation fragments and directly translates phrase *ms ledo en los EEUU* into *more reading in the US*
Fix: Increase size of hidden units. More training data of similar sentence structures.

iii (2 points) **Source Sentence:***Un amigo me hizo eso – Richard Bolingbroke.*
Reference Translation:*A friend of mine did that – Richard Bolingbroke.*
NMT Translation:*A friend of mine did that – Richard¡unk¿*

Answer:
Error: model fail to remember *Bolingbroke.*
Reason: Embedding size too small and truncated infrequent word relationships.
Fix: Make embedding size larger to include *Bolingbroke* into the dictionary.

iv (2 points) **Source Sentence:** *Solo tienes que dar vuelta a la manzana para verlo como unaepifan ıa.*
Reference Translation: *You've just got to go around the block to see it as an epiphany.*
NMT Translation: *You just have to go back to the apple to see it as an epiphany.*

Answer:
Error: *go around the block* mistranslated as *go back to the apple.*
Reason: *manzana* is a homonym that means either *apple* or *block.* Probably not enough training samples with *manzana* as *block.*
Fix: include more training samples where *manzana* is translated as *block.*

v (2 points) **Source Sentence:***Ella salv o mi vida al permitirme entrar al ba no de la sala de profesores.*
Reference Translation:*She saved my life by letting me go to the bathroom in the teachers' lounge.*
NMT Translation: *She saved my life by letting me go to the bathroom in the women's room.*

Answer:
Error: *teachers' lounge* mistranslated as *women's room.*
Reason: bias in training sample data that doesn't have enough training samples with pairs of *ella* associating with *profesores*
Fix: include more gender unbiased data into the training set.

vi (2 points) **Source Sentence:***Eso es m as de 100,000 hect areas.*
Reference Translation:*That's more than 250 thousand acres.*
NMT Translation:*That's over 100,000 acres.*

Answer:
Error: accounting unit error, doesn't translate from *100,00 hect areas* into *250 thousand acres.*
Reason: model limitations makes it hard to translate across different numbering schemes and accounting units.
Fix: more training data regarding simple arithmetics and unit conversions. flattening out the numbering schemes in training samples, e.g., instead of *250 thousand*, do *250,000.* But even then, not sure how well will the model capture basic logic and arithmetic relationships.

(b) (4 points) Now it is time to explore the outputs of the model that you have trained! The test-set translations your model produced in question 1-i should be located in `outputs/testoutputs.txt`.Please identify **2 examples** of errors that your model produced. The two examples you find should be different error types from one another and different error types than the examples provided in the previous question. For each example you should:

1. Write the source sentence in Spanish. The source sentences are in the `enesdata/test.es`.
2. Write the reference English translation. The reference translations are in the `enesdata/test.en`.
3. Write your NMT model's English translation. The model-translated sentences are in the `outputs/testoutputs.txt`.
4. Identify the error in the NMT translation.
5. Provide a reason why the model may have made the error (either due to a specific linguistic construct or specific model limitations).
6. Describe one possible way we might alter the NMT system to fix the observed error

Answer:

i **Source**: *Aqu lo tienen.*
Reference: *So here it is*
NMT Translation: *Here you have*
Error: Missing pronoun *it.*
Reason: model limitation, language construction.
Fix: More training samples these super short languages.

ii **Source**: *y tuve que irme de la ciudad.*
Reference: *And I had to get out of town.*
NMT Translation: *And I had to go from the city.*
Error: idioms/linguistic construct, *que irme de la ciudadd* literal translates to *go from the city* but is not *get out of town.*
Reason: model limitation, language construction.
Fix: Not sure. Perhaps including more idioms from both English and Spanish will help.

(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. Suppose we have a source sentences, a set of k reference translations $r_1, ..., r_k$, and a candidate translation c. To compute the BLEU score of c, we first compute the *modified n-gram precision* p_n of c, for each of $n = 1, 2, 3, 4$, where n is the n in n-gram:

$$p_n = \frac{\sum_{ngram \in c} \min(\max_{i=1,...,k} Count_{r_i}(ngram), Count_c(ngram))}{\sum_{ngram \in c} Count_c(ngram)}$$

Here, for each of the n-grams that appear in the candidate translation c, we count the maximum number of times it appears in any one reference translation, capped by the number of times it appears in c (this is the numerator). We divide this by the number of n-grams in c (denominator).

Next, we compute the *brevity penalty* BP. Let *len(c)* be the length of c and let *len(r)* be the length of the reference translation that is closest to *len(c)* (in the case of two equally-close reference translation lengths, choose *len(r)* as the shorter one).

$$BP = \begin{cases} 1, & \text{if } len(c) \geq len(r) \\ \exp\left(1 - \frac{len(r)}{len(c)}\right), & \text{otherwise} \end{cases}$$

Lastly, the BLEU score for candidate c with respect to $r_1, ..., r_k$ is:

$$BLEU = BP \exp\left(\sum_{n=1}^{4} \lambda_n \log p_n\right)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weights that sum to 1. The log here is natural log.

i (5 points) Please consider this example:
Source Sentences: el amor todo lo puede
Reference Translation r_1: love can always find a way
Reference Translation r_2: love makes anything possible
NMT Translation c_1: the love can always do
NMT Translation c_2: love can make anything possible
Please compute the BLEU scores for c_1 and c_2. Let $\lambda_i = 0.5$ for $i \in 1, 2$ and $\lambda_i = 0$ for $i \in 3, 4$ (this means we ignore 3-grams and 4-grams, i.e., don't compute p_3 or p_4). When computing BLEU scores, show your working (i.e., show your computed values for p_1, p_2, *len(c)*, *len(r)* and *BP*). Note that the BLEU scores can be expressed between 0 and 1 or between 0 and 100. The code is using the 0 to 100 scale while in this question we are using the 0 to 1 scale. Which of the two NMT translations is considered the better translation according to the BLEU score? Do you agree that it is the better translation?

Answer:

Unigram	max($Count_{r_i}, Count_c$)
the	0
love	1
can	1
always	1
do	0

bigram	max($Count_{r_i}, Count_c$)
the love	0
love can	1
can always	1
always do	0

c_1: $p_1 = 0.6$, $p_2 = 0.5$, $len(c) = 5$, $len(r) = 4$. $BP = 1$.
BLEU c_1 = `np.exp(0.5*np.log(0.65) + 0.5*np.log(0.5))` = **0.5477**

Unigram	max($Count_{r_i}, Count_c$)
love	1
can	1
make	0
anything	1
possible	1

bigram	max($Count_{r_i}, Count_c$)
love can	1
can make	0
make anything	0
anything possible	1

c_2: $p_1 = 0.8$, $p_2 = 0.5$, $len(c) = 5$, $len(r) = 4$. $BP = 1$.
BLEU $c_2 =$ `np.exp(0.5*np.log(0.8) + 0.5*np.log(0.5))` $=$ **0.6324**

According to BLEU score, c_2 is a better translation. And I agree with the score.

ii (5 points) Our hard drive was corrupted and we lost Reference Translation r_2. Please recompute BLEU scores for c_1 and c_2, this time with respect to r_1 only. Which of the two NMT translations now receives the higher BLEU score? Do you agree that it is the better translation?

Answer:

Unigram	max($Count_{r_i}, Count_c$)
the	0
love	1
can	1
always	1
do	0

bigram	max($Count_{r_i}, Count_c$)
the love	0
love can	1
can always	1
always do	0

c_1: $p_1 = 0.6$, $p_2 = 0.5$, $len(c) = 5$, $len(r) = 6$. $BP =$ `np.exp(1-6/5)` $= 0.8187$.
BLEU $c_1 =$ `np.exp(1-6/5) * np.exp(0.5*np.log(0.65) + 0.5*np.log(0.5))` $=$ **0.4484**

Unigram	max($Count_{r_i}, Count_c$)
love	1
can	1
make	0
anything	0
possible	0

bigram	max($Count_{r_i}, Count_c$)
love can	1
can make	0
make anything	0
anything possible	0

c_2: $p_1 = 0.4$, $p_2 = 0.25$, $len(c) = 5$, $len(r) = 6$. $BP =$ `np.exp(1-6/5)` $= 0.8187$.
BLEU $c_2 =$ `np.exp(1-6/5) * np.exp(0.5*np.log(0.4) + 0.5*np.log(0.25))` $=$ **0.2589**

According to BLEU score, c_1 is a better translation. And I disagree with the score.

iii (2 points) Due to data availability, NMT systems are often evaluated with respect to only a single reference translation. Please explain (in a few sentences) why this may be problematic.

Answer: With only a single reference, good translations might receive a low BLEU score due to little n-gram overlaps. With more reference sentences, the target space increases with more n-grams available for the model to generate and receive a good BLEU score.

iv (2 points) List two advantages and two disadvantages of BLEU, compared to human evaluation, as an evaluation metric for Machine Translation

Answer:
advantages:

1 Fully automated and quantitative evaluation, faster than human.

2 Simple and easy implementation. Human evaluation would have to understand both source and target languages.

disadvantages:

1 Only measures n-gram overlaps, not quality of the sentences produced.

2 Not measuring semantics, log, grammar, etc