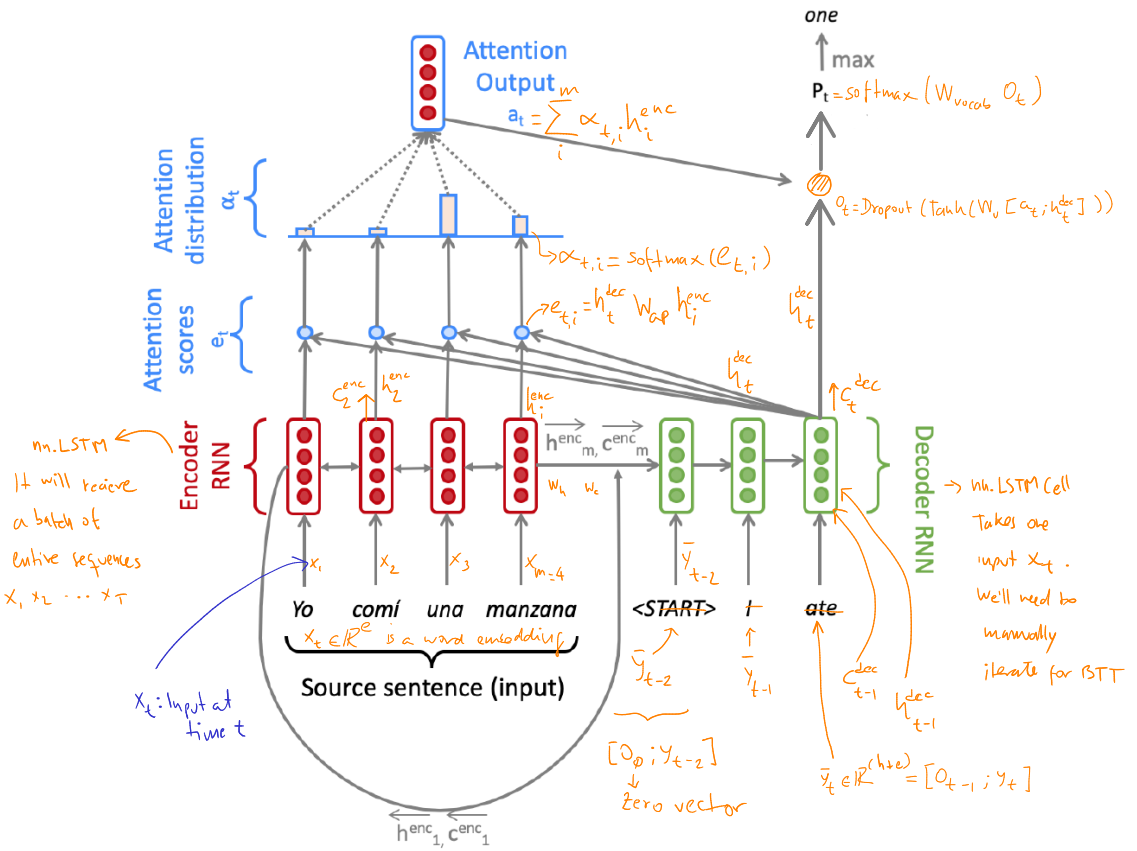
Architecture Review



Math and Dimensions Review

$$\mathbf{h}_i^{\text{enc}} = [\overleftarrow{\mathbf{h}_i^{\text{enc}}}; \overrightarrow{\mathbf{h}_i^{\text{enc}}}] \text{ where } \mathbf{h}_i^{\text{enc}} \in \mathbb{R}^{2h\times 1}, \overleftarrow{\mathbf{h}_i^{\text{enc}}}, \overrightarrow{\mathbf{h}_i^{\text{enc}}} \in \mathbb{R}^{h\times 1} \qquad 1 \le i \le m$$

$$\mathbf{c}_i^{\text{enc}} = [\overleftarrow{\mathbf{c}_i^{\text{enc}}}; \overrightarrow{\mathbf{c}_i^{\text{enc}}}] \text{ where } \mathbf{c}_i^{\text{enc}} \in \mathbb{R}^{2h\times 1}, \overleftarrow{\mathbf{c}_i^{\text{enc}}}, \overrightarrow{\mathbf{c}_i^{\text{enc}}} \in \mathbb{R}^{h\times 1} \qquad 1 \le i \le m$$

Linear projection: we are reducing the dim. of $[\overleftarrow{h_1^{enc}}; \overrightarrow{h_m^{enc}}]$ by dot product with W_h

$$\mathbf{h}_0^{\text{dec}} = \mathbf{W}_h[\overleftarrow{\mathbf{h}_1^{\text{enc}}}; \overrightarrow{\mathbf{h}_m^{\text{enc}}}] \text{ where } \mathbf{h}_0^{\text{dec}} \in \mathbb{R}^{h\times 1}, \mathbf{W}_h \in \mathbb{R}^{h\times 2h}$$

$$\mathbf{c}_0^{\text{dec}} = \mathbf{W}_c[\overleftarrow{\mathbf{c}_1^{\text{enc}}}; \overrightarrow{\mathbf{c}_m^{\text{enc}}}] \text{ where } \mathbf{c}_0^{\text{dec}} \in \mathbb{R}^{h\times 1}, \mathbf{W}_c \in \mathbb{R}^{h\times 2h}$$

$h\times 1$; $h\times 2h$; $1\times 2h$

Input of a nn.Linear

$$\mathbf{h}_t^{\text{dec}}, \mathbf{c}_t^{\text{dec}} = \text{Decoder}(\overline{\mathbf{y}_t}, \mathbf{h}_{t-1}^{\text{dec}}, \mathbf{c}_{t-1}^{\text{dec}}) \text{ where } \mathbf{h}_t^{\text{dec}} \in \mathbb{R}^{h\times 1}, \mathbf{c}_t^{\text{dec}} \in \mathbb{R}^{h\times 1}$$

Another projection

$$\mathbf{e}_{t,i} = (\mathbf{h}_t^{\text{dec}})^T \mathbf{W}_{\text{attProj}} \mathbf{h}_i^{\text{enc}} \text{ where } \mathbf{e}_t \in \mathbb{R}^{m\times 1}, \mathbf{W}_{\text{attProj}} \in \mathbb{R}^{h\times 2h} \qquad 1 \le i \le m$$

$$\alpha_t = \text{Softmax}(\mathbf{e}_t) \text{ where } \alpha_t \in \mathbb{R}^{m\times 1}$$

$$\mathbf{a}_t = \sum_i^m \alpha_{t,i} \mathbf{h}_i^{\text{enc}} \text{ where } \mathbf{a}_t \in \mathbb{R}^{2h\times 1}$$

Another projection

$$\mathbf{u}_t = [\mathbf{a}_t; \mathbf{h}_t^{\text{dec}}] \text{ where } \mathbf{u}_t \in \mathbb{R}^{3h\times 1}$$

$$\mathbf{v}_t = \mathbf{W}_u \mathbf{u}_t \text{ where } \mathbf{v}_t \in \mathbb{R}^{h\times 1}, \mathbf{W}_u \in \mathbb{R}^{h\times 3h}$$

$$\mathbf{o}_t = \text{Dropout}(\text{Tanh}(\mathbf{v}_t)) \text{ where } \mathbf{o}_t \in \mathbb{R}^{h\times 1}$$

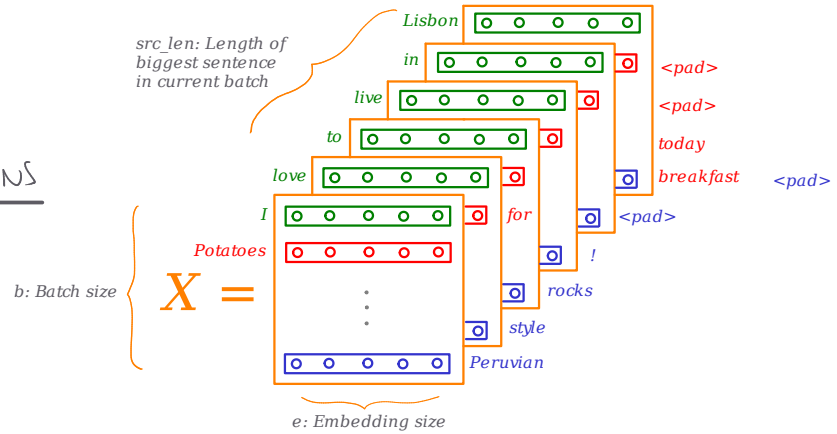
$$\mathbf{P}_t = \text{Softmax}(\mathbf{W}_{\text{vocab}} \mathbf{o}_t) \text{ where } \mathbf{P}_t \in \mathbb{R}^{V_t\times 1}, \mathbf{W}_{\text{vocab}} \in \mathbb{R}^{V_t\times h}$$

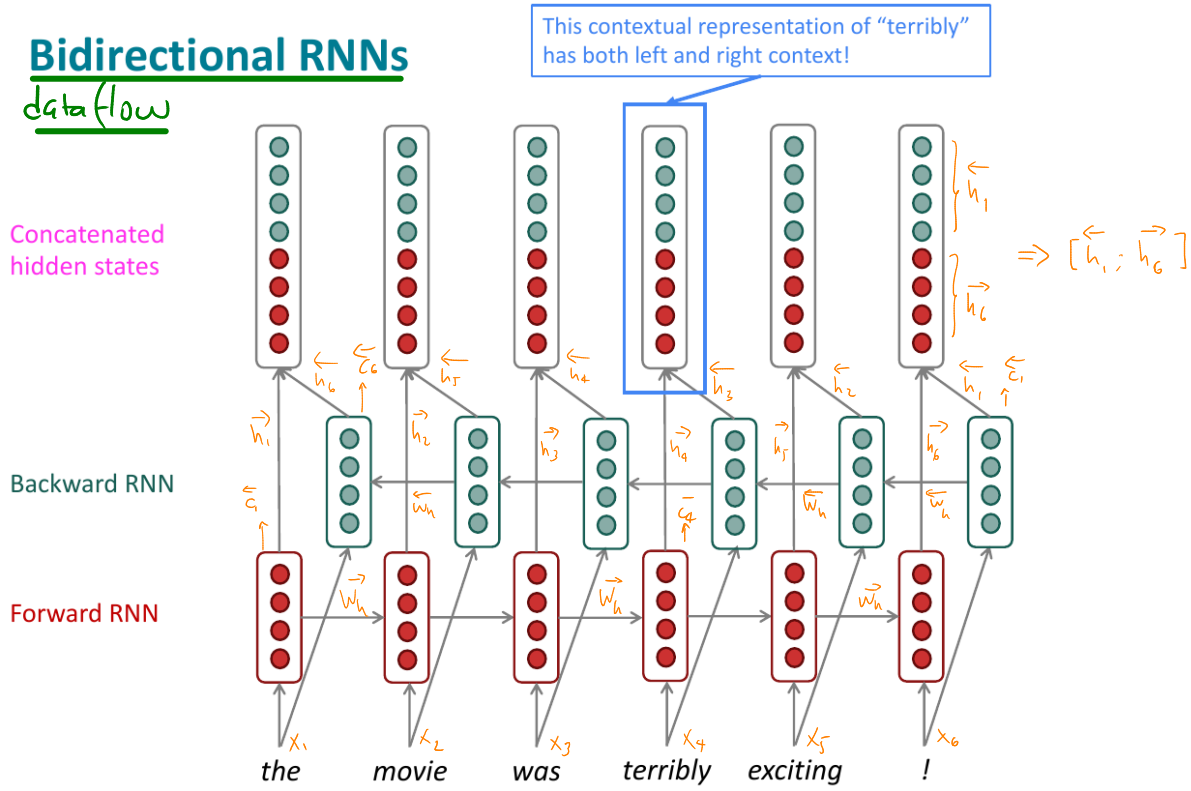
Yes, this is also a projection

$$J_t(\theta) = CE(\mathbf{P}_t, \mathbf{g}_t)$$

loss at time step t ; 1-hot vector of y_t

Input Dimensions





(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.

src_sentence = [Δ, song, for, nelly, <pad>, <pad>]

enc_mask = [0, 0, 0, 0, 1, 1]

e_t = [Δ, Δ, Δ, Δ, $-\infty$, $-\infty$]

α_t = Softmax(e_t) = [✱, ✱, ✱, ✱, 0, 0]

The attention output $a_t = \sum_{i=1}^{m} \alpha_{t,i} h_i^{enc}$ won't pay attention the pad tokens.

This is needed to avoid polluting the decoders predictions with the pad tokens.

(j) (3 points) In class, we learned about dot product attention, multiplicative attention, and additive attention. Please provide one possible advantage and disadvantage of each attention mechanism, with respect to either of the other two attention mechanisms. As a reminder, dot product attention is $\mathbf{e}_{t,i} = \mathbf{s}_t^T \mathbf{h}_i$, multiplicative attention is $\mathbf{e}_{t,i} = \mathbf{s}_t^T \mathbf{W} \mathbf{h}_i$, and additive attention is $\mathbf{e}_{t,i} = \mathbf{v}^T (\mathbf{W}_1 \mathbf{h}_i + \mathbf{W}_2 \mathbf{s}_t)$.

↳ decoder hidden state

↳ weight vector

	Advantage	Disadvantage
Dot product Attention	Encoder/decoder relationship is captured	The relationship is not learned because the absence of parameters
Multiplicative Attention	Encoder/decoder relationship is captured and learned	Might lose implicit relationships
Additive Attention	More parameters can better capture relationship and manner	Extra complexity. Improvements justify the extra cost?

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 points) **Source Sentence**: *Aquí 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".*

Error — relationship captured

I guess this is because a limitation in the model. It correctly captures that "otro" is the same type of "favoritos" but it just uses the same word which probably has a higher ranking in the beam search hypothesis. Some extra params should be added to also have into account that the source word "otro" yields a one-to-many alignment with a different word to avoid redundancy.

ii. (2 points) **Source Sentence**: *Ustedes saben que lo que yo hago es escribir para los niños, y, de hecho, probablemente soy el autor para niños, 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.*

It is a complex sentence and the NMT is translating in strict sequence which in this case won't produce the best results in English. Also "más leído" is translated with an incorrect verb tense. Perhaps the training corpus can include more examples of this kind of translations

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>*

Two things. First even the gold translation is missing that the action is directed to the speaker, that means that "me hizo eso" should be translated as "did that to me"

Second problem is that the target vocabulary lacks the lastname "Bolingbroke". To solve it just add some examples with it.

iv. (2 points) **Source Sentence:** *Solo tienes que dar vuelta a la manzana para verlo como una epifanía.*
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 a epiphany.*

The problem here is that the model is making a literal translation probably because there aren't training examples for the "block" meaning for apple.

v. (2 points) **Source Sentence:** *Ella salvó mi vida al permitirme entrar al baño 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.*

Clearly the problem is using women's instead of teachers'. It seems to be a gender bias issue in the training data. To solve it replace those training examples.

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

The issue is that the NMT should use hectares instead of acres. It is treating both words as synonyms. Perhaps the training examples are incorrectly using acres, or the training data lacks examples doing the numeric conversion between hectares to acres.

(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/test_outputs.txt`. Please identify **2 examples** of errors that your model produced.[2] 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 `en_es_data/test.es`.
2. Write the reference English translation. The reference translations are in the `en_es_data/test.en`.
3. Write your NMT model's English translation. The model-translated sentences are in the `outputs/test_outputs.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.

First example:

Source sentence: Poco tiempo después, una organización donde era voluntaria, All hands volunteers, estuvo en el lugar trabajando como parte del equipo de respuesta

Reference transl.: Soon after, an organization I volunteer with, All hands volunteers, were on the ground, within days, working as part of the response efforts.
?

NMT translation: Soon time later, an organization where he was voluntary, -- there was <unk><unk> in the place working as part of the response team.

Error: Extra dashes appeared without a reason and English words were not added without translating them.

Why: Lack of enough training examples containing words in the target language to learn how to treat them.

Fix: Add training examples

Second example:

Source sentence: Es una comunidad de vacaciones

Ref. translation: It's a vacation community

NMT translation: It's a vacation

Error: It didn't include a capital word => community

Why: This seems to be a specific model limitation. The word "community" might appear as hypothesis but maybe with a slightest worse value so it was not selected.

Fix: I will try increasing the k parameter in beam search and see if that solves the issue

(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 $\mathbf{s}$, a set of k reference translations $\mathbf{r}_1, \ldots, \mathbf{r}_k$, and a candidate translation $\mathbf{c}$. To compute the BLEU score of $\mathbf{c}$, we first compute the *modified n-gram precision* p_n of $\mathbf{c}$, for each of $n = 1, 2, 3, 4$:

$$p_n = \frac{\sum_{\text{ngram} \in \mathbf{c}} \min\left(\max_{i=1,\ldots,k} \text{Count}_{\mathbf{r}_i}(\text{ngram}),\ \text{Count}_{\mathbf{c}}(\text{ngram}) \right)}{\sum_{\text{ngram} \in \mathbf{c}} \text{Count}_{\mathbf{c}}(\text{ngram})} \tag{15}$$

$k=2$ $\quad \text{Count}_c(1\text{-gram}) + \text{Count}_c(2\text{-gram}) + \cdots \text{Count}_c(n\text{-gram})$

i. (5 points) Please consider this example:

Source Sentence $\mathbf{s}$: **el amor todo lo puede**
Reference Translation $\mathbf{r}_1$: *love can always find a way* len(r1) = 6
Reference Translation $\mathbf{r}_2$: *love makes anything possible* len(r2) = 4 as r^*
NMT Translation $\mathbf{c}_1$: *the love can always do* len(c1) = 5
NMT Translation $\mathbf{c}_2$: *love can make anything possible* len(c2) = 5

Please compute the BLEU scores for $\mathbf{c}_1$ and $\mathbf{c}_2$. 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, c, r^* and BP).

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?

BLEU for c_1: $\lambda_1 = \lambda_2 = 0.5$ $\quad \lambda_3 = \lambda_4 = 0$ $\quad \sum \lambda_i = 1$ $\quad r^* = 4$

$$\text{BLEU}_{c_1} = BP_{c_1} \times \exp\left(\sum_{n=1}^{4} \lambda_n \log p_n \right)$$

$$BP_{c_i} = \begin{cases} 1 & \text{if } c_i \geq r_i^* \\ \exp\left(1 - \frac{r_i^*}{c_i}\right) & \text{otherwise} \end{cases}$$

Unique 1-grams	Count	Clipped count
the	0	0
love	1	1
can	1	1
always	1	1
do	0	0

$p_1 = 3/5$

Unique 2-grams	Count	Clipped
the love	0	0
love can	1	1
can always	1	1
always do	0	0

$p_2 = 2/4$

-> max freq. in ref.

$$\text{BLEU}_{c_1} = 1 \times \exp\left(0.5 \log_e 3/5 + 0.5 \log_e 2/4\right)$$

$$\text{BLEU}_{c_1} = 0.548$$

BLEU FOR c_2 $\lambda_1 = \lambda_2 = 0.5$ $\lambda_3 = \lambda_4 = 0$

unique unigrams	count	clipped count
love	1	1
can	1	1
make	0	0
anything	1	1
possible	1	1

$p_1 = \frac{4}{5}$

unique bigrams	count	clipped
love can	1	1
can make	0	0
make anything	0	0
anything possible	1	1

$p_2 = \frac{2}{4}$

$BP_{c_2} = 1$

$BLEU_{c_2} = \exp(0.5 \log 4/5 + 0.5 \log 2/4)$

$BLEU\ c_2 = 0.752$ // Yeah, c_2 is a better candidate

ii. (5 points) Our hard drive was corrupted and we lost Reference Translation $\mathbf{r}_2$. Please recompute BLEU scores for $\mathbf{c}_1$ and $\mathbf{c}_2$, this time with respect to $\mathbf{r}_1$ only. Which of the two NMT translations now receives the higher BLEU score? Do you agree that it is the better translation?

BLEU FOR c_2: $r^* = 6$; $\lambda_1 = \lambda_2 = 0.5$ $\lambda_3 = \lambda_4 = 0$ $BP_c = \exp(1 - 6/5) = 0.819$

$p_1 = 2/5$ $p_2 = 1/4$ $BLEU_{c_1} = 0.819 \times \exp(0.5 \log 2/5 + 0.5 \log 1/4) = 0.259$

BLEU FOR c_1: $p_1 = 3/5$ $p_2 = 2/4$ $BLEU_{c_2} = 0.819 \times \exp(0.5 \log 3/5 + 0.5 \log 2/4) = 0.449$ //

I don't agree; c_1 is worse

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.

This will affect valid translations that are using synonyms of the reference's words. In those cases the BLEU score will be low and a worse translation might be chosen as we saw in previous question where "anything possible" didn't appear in r_1

this is having low n-gram overlap with the reference translations

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

Advantages: - Easy and inexpensive to compute
- Useful to set a baseline for NMT, measure progress and track improvement

Disadvantages: - Might be misleading specially with candidates containing repeated words which might have higher scores.
- Not accurate with only one reference (low n-gram overlap)