

上图是使用乘法注意力的Seq2Seq模型，显示了解码器的第三步。注意，为了可读性，我们不描绘前一个组合输出与解码器输入的连接。

给定源语言中的一个句子，我们从词嵌入矩阵中查找单词嵌入，得到  $\mathbf{x}_1, \dots, \mathbf{x}_m | \mathbf{x}_i \in \mathbb{R}^{e \times 1}$ ，其中  $m$  为源语句的长度， $e$  为嵌入大小。我们将这些嵌入提供给双向编码器，为正向( $\rightarrow$ )和反向( $\leftarrow$ )LSTMs生成隐藏状态和单元格状态。前向和后向的版本连接起来，以得到隐藏状态  $\mathbf{h}_i^{\text{enc}}$  和单元格状态  $\mathbf{c}_i^{\text{enc}}$

$$\begin{aligned} \mathbf{h}_i^{\text{enc}} &= [\overleftarrow{\mathbf{h}_i^{\text{enc}}}; \overrightarrow{\mathbf{h}_i^{\text{enc}}}] \text{ where } \mathbf{h}_i^{\text{enc}} \in \mathbb{R}^{2h \times 1}, \overleftarrow{\mathbf{h}_i^{\text{enc}}}, \overrightarrow{\mathbf{h}_i^{\text{enc}}} \in \mathbb{R}^{h \times 1} \quad 1 \leq i \leq m \\ \mathbf{c}_i^{\text{enc}} &= [\overleftarrow{\mathbf{c}_i^{\text{enc}}}; \overrightarrow{\mathbf{c}_i^{\text{enc}}}] \text{ where } \mathbf{c}_i^{\text{enc}} \in \mathbb{R}^{2h \times 1}, \overleftarrow{\mathbf{c}_i^{\text{enc}}}, \overrightarrow{\mathbf{c}_i^{\text{enc}}} \in \mathbb{R}^{h \times 1} \quad 1 \leq i \leq m \end{aligned}$$

然后，我们使用编码器的最终隐藏状态和最终单元状态的线性投影，初始化解码器的第一个隐藏状态  $\mathbf{h}_0^{\text{dec}}$  和单元状态  $\mathbf{c}_0^{\text{dec}}$

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

初始化解码器之后，现在必须用目标语言为它提供匹配的句子。在第  $t$  步，我们查找第  $t$  个单词的嵌入， $\mathbf{y}_t \in \mathbb{R}^{e \times 1}$ 。然后，我们将  $\mathbf{y}_t$  与前一个时间步的 combined-output 组合输出向量

$\mathbf{o}_{t-1} \in \mathbb{R}^{h \times 1}$  连接起来(我们将在下一页解释这是什么!), 得到  $\overline{\mathbf{y}}_t \in \mathbb{R}^{(e+h) \times 1}$ 。注意, 对于第一个目标单词(即 start 标记),  $\mathbf{o}_0$  是一个零向量。然后将  $\overline{\mathbf{y}}_t$  作为输入输入到解码器 LSTM 中。

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

然后我们用  $\mathbf{h}_t^{\text{dec}}$  来计算在  $\mathbf{h}_0^{\text{enc}}, \dots, \mathbf{h}_m^{\text{enc}}$  上的乘法注意

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

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

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

现在, 我们将注意力输出  $\alpha_t$  与解码器隐藏状态  $\mathbf{h}_t^{\text{dec}}$  连接起来, 并将其通过线性层 Tanh 和 Dropout 来获得组合输出向量  $\mathbf{o}_t$ 。

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

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

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

然后, 在第  $t$  个时间步长时, 得到目标词的概率分布  $\mathbf{P}_t$

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

这里,  $V_t$  是目标词汇表的大小。最后, 为了训练网络, 我们计算了  $\mathbf{P}_t, \mathbf{g}_t$  之间的 softmax 交叉熵损失,  $\mathbf{g}_t$  是时间步  $t$  的目标词的 one-hot 向量

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

在这里,  $\theta$  代表所有的模型参数,  $J_t(\theta)$  是解码器第  $t$  步的损失。现在我们已经描述了该模型, 让我们尝试将其实现为西班牙语到英语的翻译!

#### Pytorch Bidirectional RNNs Note

Pytorch 中的 RNNs, 返回的 **out** 的 shape 为 (seq\_len, batch, num\_directions \* hidden\_size)

- 转换为 (seq\_len, batch, num\_directions, hidden\_size) 后, num\_directions 中的顺序是先 forward 再 backward, 并且 forward 和 backward 的 hidden state 的顺序是相反的, 即  $\text{out}[0][0][0]$  是 forward 的第一个时间步的结果, 而  $\text{out}[0][0][1]$  是 backward 的最后一个时间步的结果。此外, **out** 只包含最后一层的结果

但对于 **h\_n** (**c\_n** 同理) 而言, shape 为 (num\_layers \* num\_directions, batch, hidden\_size), 保存的是 forward 和 backward 的最后一个时间步的结果。

- 转换为 (num\_layers, num\_directions, batch, hidden\_size) 后, 第一维的 num\_layers 和真实的 layer 层数——对应, 即  $h_n[1][0][0]$  与  $\text{out}[-1][0][0]$  相等,  $h_n[1][1][0]$  与  $\text{out}[0][0][1]$ 。

### ? Question 1.g

首先解释(大约三句话) masks 对整个注意力计算有什么影响。然后(用一两句话)解释为什么有必要这样使用 masks。

### Answer 1.g

- 使用 masks 将句子中的 pad token 的分数赋值为  $-\infty$ ，从而使得 softmax 作用后获得的 attention 分布中，pad token 的 attention 概率值近似为 0
- attention score / distributions 计算的是 decoder 中某一时间步上的 target word 对 encoder 中的所有 source word 的注意力概率，而 pad token 只是用于 mini-batch，并没有任何语言意义，target word 无须为其分散注意力，所以需要 masks 过滤掉 pad token

### ? Question 1.j

在课堂上，我们学习了点积注意、乘法注意和加法注意。请就其他两种注意机制中的任何一种，提供每种注意机制可能的优点和缺点

- 点积注意  $\mathbf{e}_{t,i} = \mathbf{s}_t^T \mathbf{h}_i$
- 乘法注意  $\mathbf{e}_{t,i} = \mathbf{s}_t^T \mathbf{W} \mathbf{h}_i$
- 加法注意  $\mathbf{e}_{t,i} = \mathbf{v}^T (\mathbf{W}_1 \mathbf{h}_i + \mathbf{W}_2 \mathbf{s}_t)$

### Answer 1.j

	优点	缺点
点积注意	不需要额外的线性映射层	$\mathbf{s}_t, \mathbf{h}_t$ 必须有同样的维度
乘法注意	$\mathbf{s}_t, \mathbf{h}_t$ 不需要有同样的维度并且因为可以使用高效率的矩阵乘法，比加法注意力要更快更省内存	增加了训练参数
加法注意	高维时的表现更好	训练参数更多（两个参数矩阵以及注意力的维度）

## 2. Analyzing NMT Systems

### ? Question 2.a

这里，我们展示了在NMT模型的输出中发现的一系列错误(与您刚刚训练的模型相同)。对于西班牙语源句的每个示例，标准英文翻译，以及NMT(即，“模型”)，请你：

- 识别NMT翻译中的错误
- 提供模型可能出错的原因(由于特定的语言构造或特定的模型限制)
- 描述一种可能的方法，我们可以改变NMT系统，以修复观察到的错误

下面是您应该按照上面描述的那样分析的翻译。请注意，标记了下划线的单词是词汇表外的单词

- 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”.*
- 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.*
- 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>*
- 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.*
- 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.*
- 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.*

## Answer 2.a

- 
- Error: “ **favorite** of my favorites”
  - Reason: 特定的语言构造，低资源语言对
  - Possible fix: 尝试在这类语言对上添加更多的训练数据
- 
- Error: “ **more reading** in the U.S.” 语义错误
  - Reason: 特定的语言构造，模型对语义的理解不足，需要增大模型的容量以增强理解能力
  - Possible fix: 增大Hidden\_size
- 
- Error: “Richard \<unk>”
  - Reason: 模型限制，Bolingbroke 是词表外的单词
  - Possible fix: 对此类姓名中出现的词加以处理，比如直接添加到词表中
- 
- Error: “go back to the apple “

- Reason: 模型限制, “manzana” 有丰富的含义, 包括 apple 苹果和 block 街区。“block”在西班牙语中的表达方式比 “apple” 在西班牙语中的表达方式更多。然而, 在训练集中, “manzana”更多地表示“apple”, 而不是“block”。
- Possible fix: 在训练集中添加更多的关于 “manzana” 表示 “block” 的数据, 保持多重含义的训练不失衡

- 
- Error: “go to the bathroom in the women’s room”
  - Reason: 模型限制, 由于在数据集中, 女性比专业人员(教师)的出现频率要更高, 所以导致翻译具有来自训练数据的偏见 bias
  - Possible fix: 添加更多 profesore 的训练样本

- 
- Error: “100,000 acres.”
  - Reason: 模型限制, 常识错误, hectáreas 表示公顷, acres 表示英亩 (acre的复数)。模型并未理解两个单位制之间的转换关系, 由于 acres 在训练集中的出现频率更高, 直接采用 acres 并且使用 hectáreas 附近的数字直接修饰 acres
  - Possible fix: 添加关于 hectáreas 的训练数据

### ? Question 2.b

现在是时候探索您所训练的模型的输出了! 问题 1-i 中生成的模型的测试集翻译应该位于output /test\_output.txt中。请找出你的模型产生的两个错误示例。你发现的两个例子应该是不同的错误类型, 并且与前一个问题中提供的例子不同。对于每个例子, 你应该:

- 写下西班牙语原文句子。源语句在 en\_es\_data/test.es 中
- 写下参考译文, 参考译文在en\_es\_data/test.en中
- 写下NMT模型的英文翻译, 模型翻译的句子位于output /test\_output .txt中
- 识别NMT翻译中的错误
- 提供模型可能出错的原因(由于特定的语言构造或特定的模型限制)
- 描述一种可能的方法, 我们可以改变NMT系统, 以修复观察到的错误

### Answer 2.b

- 
- Source Sentence: El 5 de noviembre de 1990
  - Reference Translation: On November 5<sup>th</sup>, 1990
  - NMT Translation: On **five** of November 1990
  - Error: five
  - Reason: 模型限制, 模型没有数据集中充分学习到日期格式的转换
  - Possible Fix: 增加更多关于西班牙语与英语之间的日期格式转换的数据样本

- Source Sentence: Y mis amigos hondureos me pidieron que dijera: "Gracias TED".
- Reference Translation: And my friends from Honduras asked me to say thank you, TED.
- NMT Translation: My friends were asked to say, "Thank you."
- Error : 说话的对象错误, 说话的人是我而不是我的朋友
- Reason: 句法结构有误并且有缺译现象
- Possible Fix: 尝试为模型的添加更有效的对齐方式, 如优化注意力模型

### Question 2.c

BLEU评分是NMT系统中最常用的自动评价指标。它通常在整个测试集中计算, 但这里我们将考虑为单个示例定义的BLEU。假设我们有一个源句  $s$ , 一组  $k$  个参考译文  $\mathbf{r}_1, \dots, \mathbf{r}_k$  和一个候选翻译  $\mathbf{c}$ 。为了计算  $\mathbf{c}$  的BLEU分数, 我们首先为  $\mathbf{c}$  计算修改后的  $n$ -gram 精度  $p_n$ , 对于  $n = 1, 2, 3, 4$ :

$$p_n = \frac{\sum_{\text{ngram} \in \mathbf{c}} \min(\max_{i=1, \dots, k} \text{Count}_{r_i}(\text{ngram}), \text{Count}_{\mathbf{c}}(\text{ngram}))}{\sum_{\text{ngram} \in \mathbf{c}} \text{Count}_{\mathbf{c}}(\text{ngram})}$$

这里, 对于出现在候选翻译  $\mathbf{c}$  中的每个  $n$ -gram, 我们计算它在任何一个参考译文中出现的最大次数, 并以它出现在  $\mathbf{c}$  中的次数为上限(这是分子), 再除以  $\mathbf{c}$  的  $n$ -gram (分母)

接下来, 我们计算简洁代价 brevity penalty BP。令  $c$  作为  $\mathbf{c}$  的长度, 让  $r^*$  作为最接近  $\mathbf{c}$  的参考翻译的长度(在两个相等接近的参考翻译长度的情况下, 选择较短的参考翻译的长度作为  $r^*$ )

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

最后, 候选翻译  $\mathbf{c}$  关于  $\mathbf{r}_1, \dots, \mathbf{r}_k$  的BLEU分数为:

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

其中,  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$  是总和为1的权重

#### ? Question 2.c.i

请考虑这个例子:

Source Sentence  $s$ : 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

分别计算  $c_1, c_2$  的BLEU分数。令  $\lambda_i = 0.5$  for  $i \in \{1, 2\}$ ,  $\lambda_i = 0$  for  $i \in \{3, 4\}$ 。当计算BLEU分数时，显示你的计算过程(展示  $p_1, p_2, c, r^*, BP$  的计算值)。

根据BLEU评分，这两种NMT翻译中哪一种被认为是更好的翻译？你同意这是更好的翻译吗？

### Answer 2.c.i

$c_1$

$$\begin{aligned}p_1 &= \frac{0+1+1+1+0}{5} = 0.6 \\p_2 &= \frac{0+1+1+0}{4} = 0.5 \\c &= 5 \\r^* &= 4 \\BP &= 1 \\BLEU_{c_1} &= 1 * \exp(0.5 * \log(0.6) + 0.5 * \log(0.5)) = 0.5477\end{aligned}$$

$c_2$

$$\begin{aligned}p_1 &= \frac{1+1+0+1+1}{5} = 0.8 \\p_2 &= \frac{1+0+0+1}{4} = 0.5 \\c &= 5 \\r^* &= 4 \\BP &= 1 \\BLEU_{c_1} &= 1 * \exp(0.5 * \log(0.8) + 0.5 * \log(0.5)) = 0.632\end{aligned}$$

根据 BLEU 分数， $c_2$  是得分更高的翻译，但我认为  $c_1$  的翻译更加好

### ? Question 2.c.ii

我们的硬盘坏了，我们失去了参考翻译  $r_2$ 。请重新计算  $c_1$  和  $c_2$  的BLEU分数，这次只针对  $r_1$ 。两个NMT分一中，哪一个现在获得了更高的BLEU分数？你同意这是更好的翻译吗？

### Answer 2.c.ii

$c_1$

$$\begin{aligned}
p_1 &= \frac{0+1+1+1+0}{5} = 0.6 \\
p_2 &= \frac{0+1+1+0}{4} = 0.5 \\
c &= 5 \\
r^* &= 6 \\
BP &= \exp\left(1 - \frac{6}{5}\right) = 0.8187 \\
BLEU_{c_1} &= 0.8187 * \exp(0.5 * \log(0.6) + 0.5 * \log(0.5)) = 0.4484
\end{aligned}$$

$c_2$

$$\begin{aligned}
p_1 &= \frac{1+1+0+0+0}{5} = 0.4 \\
p_2 &= \frac{1+0+0+0}{4} = 0.25 \\
c &= 5 \\
r^* &= 6 \\
BP &= \exp\left(1 - \frac{6}{5}\right) = 0.8187 \\
BLEU_{c_1} &= 0.8187 * \exp(0.5 * \log(0.4) + 0.5 * \log(0.25)) = 0.2589
\end{aligned}$$

根据 BLEU 分数,  $c_1$  是得分更高的翻译, 并且我认为这是对的

#### ? Question 2.c.iii

由于数据可用性, NMT系统通常只根据一个参考翻译进行评估。请解释(用几句话)为什么这可能有问题?

#### Answer 2.c.iii

如果我们使用单一参考翻译, 它增加了好翻译由于与单一参考翻译有较低的 n-gram overlap, 而获得较差的BLEU分数的可能性。例如上例中, 如果删去的参考翻译是  $r_1$ , 那么将使得  $c_1$  的BLEU分数变低。

如果我们增加更多的参考翻译, 就会增加一个好翻译中 n-gram overlap 的几率, 这样我们就有可能使好翻译获得相对较高的BLEU分数。

#### ? Question 2.c.iv

列举了BLEU作为机器翻译的评价指标, 相对于人工评价的两个优点和两个缺点。

#### Answer 2.c.iv

优点

- 自动评价, 比人工评价更快, 方便, 快速
- BLEU的使用普及率较高, 方便模型之间的效果对比

缺点



- 结果并不稳定，由于核心思想是 n-gram overlap，所以如果参考翻译不够丰富，会导致出现较好翻译获得较差BLEU分数的情况
- 不考虑语义与句法
- 不考虑词法，例如上例中的make和makes
- 未对同义词或相似表达进行优化


## Reference


- [从SVD到PCA——奇妙的数学游戏](https://my.oschina.net/findbill/blog/535044) [https://my.oschina.net/findbill/blog/535044]
- [alongstar518](https://github.com/alongstar518/CS224NHomeworks) [https://github.com/alongstar518/CS224NHomeworks]
- [NLP 中评价文本输出都有哪些方法？为什么要小心使用 BLEU?](https://www.leiphone.com/news/201901/1ij9vMCBDQ84qJly.html) [https://www.leiphone.com/news/201901/1ij9vMCBDQ84qJly.html]


## 评论


What do you think?


1条回复


 Upvote

 Funny

 Love

 Surprised

 Angry

 Sad

评论 在线社区 1 登录

 推荐  推文  分享 评分最高

开始讨论...

通过以下方式登录

或注册一个 DISQUS 帐号 

姓名