探究 transformer 解码器中编码器不同层的输出对翻译效果影响

摘要

随着模型结构的快速演化，神经机器翻译取得了显著的进展。本文在 transformer 框架下探究不同编码器层输出对机器翻译任务中性能的影响。具体来说，相较于 transformer 基线模型下只有编码器最后一层输出被利用，本文探究从底层到与解码器层对应的编码器层的每层加权和以及与解码器层对应的编码器层到顶层的每层加权和这两种方式对性能的影响。实验表明，在 IWSLT 14 英德数据集上，这两种方法给模型带来不同程度的性能下降。

1 介绍

神经机器翻译是近年来备受关注的一项具有挑战性的任务，并且神经机器翻译模型的结构也得到了迅速的发展。在神经网络的范式下，第一种神经机器翻译模型的设计是基于循环神经网络(RNN)，然后引入注意力机制更好地建模源语和目标语之间的对齐关系，接着采用更深层的架构来增加神经机器翻译模型的表达能力。接下来，基于卷积神经网络(CNN)和基于自注意力(Transformer)的模型被提出，在许多被广泛采用的翻译任务中都达到最先进的性能。

即使这些模型使用不同的基本构建块，但他们都属于典型的编码器-解码器框架，即编码器将源语言标记作为输入，从低级别到高级别逐层输出隐藏表示。然后解码器将来自编码器层的最后一层输出(最高级别表示)作为输入，为每个目标语位置生成从低级别到高级别的隐藏层表示，最后基于最后一层表示，生成目标语。从中可以看出，针对目标语隐藏层表示的生成，无论是高层还是底层，都是基于源语言的最高层表示。

在可视化 transformer 结构的工作中，有人提出编码器中每层的输出表示语言的不同信息，比如较低层的表示更多与单词相关，较高层的表示与语义语法的相关程度更多。而有前人提出，将编码器与解码器的每层表示协调在一起会对翻译任务的效果有所改善。那么问题自然产生：既然编码器顶层输出以及当前层输出有效，那么二者之间的所有层的输出是否有效果呢？从底层到当前层的输出对结

果的影响呢？为什么仅把解码器的一层关联到解码器的每一层，而不是多层输出的结合呢？正是这些问题激发这次的实验尝试。

在这个报告中，我们使用 transformer 模型中编码器和解码器的不同层输出进行关联。模型中的编码器和解码器具有相同的层数，关于编码器和解码器之间的注意力机制，本文探索两种方式。第一种方式下，解码器的第 0 层至第 i 层的输出进行不同形式的加权求和，再与解码器的第 i 层相关联。第二种方式下，解码器的第 i 层至第 n 层的输出进行不同形式的加权求和，再与解码器的第 i 层相关联。解码器和编码器内部之间的注意力机制不做变动。

解码器在生成目标语时可以利用更细粒度的源语言信息，并且可以使用编码器中不同层的信息，不仅仅是之前的对应层或者是编码器最高层的输出。然而，实验效果很不理想，在 IWSLT 英德任务中，基线模型 BELU 得分为 35.81，但无论是第一种还是第二种尝试，得分都会下降 2-3 个 BELU。

2 背景

2.1 编码器和解码器框架

给定一个双语句子对(x,y)，神经机器翻译模型通过最大化对数似然$P(y|x;\theta)$来优化参数θ。$P(y|x;\theta)$可以被分解为对每个目标语单词的条件概率的乘积：

$$P(y|x;\theta) = \prod_{t=1}^{m} P(y_t|y < t, x;\theta)$$

其中，m 代表句子 y 的长度，$y < t$ 为位置 t 之前的目标标记。

通常采用编码器和解码器框架来建模条件概率$P(y|x;\theta)$，编码器将输入语句 x 映射成一组隐藏表示 h，解码器使用先前生成的目标语 y<t 以及源语言表示 h，去生成目标令牌y_t。编码器和解码器都可以由不同的结构的神经网络模型实现，如 RNN、CNN 以及 Transformer。

与典型的编码器和解码器框架不同，本文设计的机理可以使解码器中每一层利用编码器对应层以及以下的层数的输出或对应层以及以上的层数的输出，而不是利用解码器最后一层或者解码器相对应层的隐藏表示。

2.2 自注意力机制

自注意力机制在之前的很多工作中都有使用，transformer 的提出也利用这一机制。对于单个自注意力层，利用交叉位置的自注意力从整个句子中的标记中提取信息，然后利用一个前馈神经网络来增加非线性表示。自注意机制的表述如下：

$$\text{Attention}(\text{Q},\text{K},\text{V}) = softmax(\frac{QK^T}{\sqrt{d_{model}}})V$$

其中，d_{model}表示隐藏层表示的维度。将词嵌入大小，自注意的输入和输出大小均设置为d_{model}。对于编码器中的自注意机制，$\text{Q},\text{K},\text{V} \in R^{n*d_{model}}$，对于解码器中的自注意机制，$\text{Q},\text{K},\text{V} \in R^{m*d_{model}}$，其中 n 和 m 分别为源语句子和目标语句子的长度。对从源语到目标语中编码器和解码器的交叉注意，$\text{Q} \in R^{m*d}$，$\text{K},\text{V} \in R^{n*d_{model}}$。所有的 Q，K，V 都来自于编码器或解码器中的隐藏层表示，但是由于不同的参数矩阵投影W_Q，W_K和W_V。基于位置的前馈神经网络有两层线性变换和 ReLU 激活函数组成：

$$\text{FFN}(\text{x}) = \max(0, xW_1 + b_1)\,W_2 + b_2$$

在编码器和解码器中的每一层都采用前馈网络。

3 方法

本文仅从一个方面修改了 transformer 的结构，解码器的每一层都关注编码器的相对应层以及向上所有层或者向下所有层的加权输出。编码器和解码器具有相同的层数，且解码器的第 i 层可以直接从编码器层相对应层及其向上所有层或者向下所有层提取信息，而不是标准 transformer 那样从编码器的最后一层提取信息。

本文尝试了两种方式。第一种方式，解码器的第 i 层与编码器的第 0 层到第 i 层每层输出的加权和相关联。如图 1 所示，解码器的第 3 层与编码器的第 0 层、第 1 层、第 2 层与第 3 层相关联。加权的方式包含三种：(1)对相关联的层数进行平均求和；(2)给予较高的层更大的权重；(3)给予较低的层更大的权重。

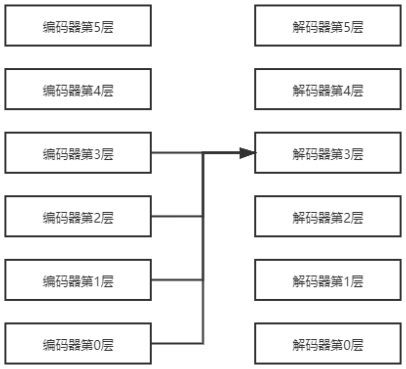

图 1：第一种方式示意图

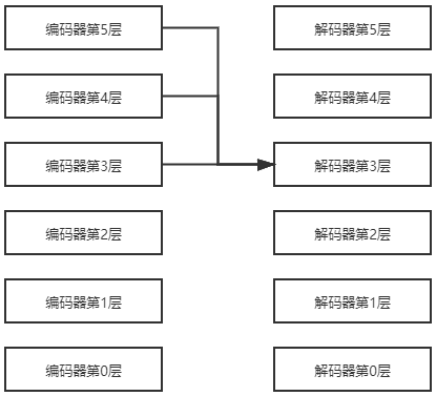

图 2：第二种方式示意图

第二种方式，解码器的第 i 层与编码器的第 i 层到第 n 层每层输出的加权和相关联。如图 2 所示，解码器的第 3 层与编码器的第 3 层、第 4 层与第 5 层相关联。加权的方式包含三种：(1)对相关联的层数进行平均求和；(2)给予较高的层更大的权重；(3)给予较低的层更大的权重。

4 实验设置

4.1 数据集

本文基于 fairseq 框架，仅在 IWSLT14 英德翻译任务上评估我们的模型。

4.2 模型配置

对于 IWSLT 英德小数据集，我们选择模型隐藏层大小$d_{model} = 512$，前馈隐藏层大小$d_{ff} = 1024$较小的配置，选择使用与小 transformer 相同的注意力头的数量 4。同时，编码器和解码器的层数都为 6。

4.3 优化器

训练时，选择 Adam 优化器，超参数$\beta_1 = 0.9$，$\beta_2 = 0.98$，$\varepsilon = 10^{-9}$，学习率 lr 与 transformer 基线相同，在训练过程中改变学习率，根据公式：

$$\text{lrate} = d_{model}^{-0.5} \cdot \min(step_{num}{}^{-0.5}, step\_num \cdot warmup\_steps^{-1.5})$$

这对应于线性地增加第一个 warmup_steps 训练步骤的学习速率，并且此后按步数的反平方根比例地减少它。我们使用 warmup_steps = 8000。

5 实验结果

5.1 实验结果

我们在 IWSLT 14 英德翻译任务上评价了修改过地模型，并与小的 transformer 模型进行比较。基线模型在该数据集上达到 35.81 的 BLEU 值(BELU 值越大，翻译效果越好)。在第一种方式下，即与解码器第 i 层相关联的编码器层是从第 0 层到编码器的第 i 层的加权输出。本文做了 3 次尝试：(1)将编码器第 0 层到第 i 层的输出平均，BLEU 得分为 33.02；(2)给予层数高的更多的权重，BLEU 的得分为 33.88；(3) 给予层数低的更多的权重，BLEU 的得分为 33.82。在第二种方式下，即与解码器第 i 层相关联的编码器层是从第 i 层到编码器第 n 层(最高层)的加权输出。同样做了 3 次尝试：(1)将编码器第 i 层到第 n 层的输出平均，BLEU 得分为 34.20；(2)给予层数高的更多的权重，BLEU 的得分为 33.94；(3) 给予层数低的更多的权重，BLEU 的得分为 32.13。

表 1　IWSLT 英德数据集，不同模型 BLEU 得分对比

模型	方法	BLEU
transformer	无	35.81
第一种方式(编码器第 0 层到第 i 层的加权)	平均连接层的输出	33.02
	给层数高的更大权重	33.88
	给层数低的更大权重	33.82
第二种方式(编码器第 i 层到第 n 层的加权)	平均连接层的输出	34.20
	给层数高的更大权重	33.94
	给层数低的更大权重	32.13

注：编码器第 i 层为解码器第 i 层相对应的层，编码器第 n 层为最高层。

5.2 结果分析

在神经机器翻译模型中，目标语的生成依赖于源语的表示和已经生成的目标语。相关研究表明，源语影响生成句子的充分性，而目标语信息影响生成句子的流畅性。这种以编码器第 i 层为界限的划分方式本意是为了区分开不同层的信息，使得传递给解码器第 i 层的信息更加高效，减少冗余信息或者噪声。但从实验结果来看，单单简单的划分编码器的输出不仅不可以该模型带来增益，反而会带来一定的负面效果。

这种结果可能有两种原因：(1)不同层的信息表示有明显区别这种先验知识是不合理的，或者不能简单理解成不同层表示不同信息，它们可能是混合在一起。(2)简单的加权连接破坏了模型的整体效果，做模型更改时，要把相关的地方都做出相应的改动。

6 结论

在本次尝试中，我们通过改变编码器与解码器的连接来尝试改进现有模型。通过编码器多个层的加权表示代替之前的编码器最高层表示，协调模型的编码器和解码器的连接。然而，实验结果表明，在 IWSLT 14 英德数据集上改动的模型无一例外 BLEU 得分均低于基线模型，我们对这一现象做出了简单分析。