计算机工程与科学
Computer Engineering & Science
ISSN 1007-130X,CN 43-1258/TP

《计算机工程与科学》网络首发论文

题目：信息传递增强的神经机器翻译
作者：史小静，宁秋怡，季佰军，段湘煜
收稿日期：2020-03-17
网络首发日期：2020-10-15
引用格式：史小静，宁秋怡，季佰军，段湘煜．信息传递增强的神经机器翻译[J/OL]．计算机工程与科学.
https://kns.cnki.net/kcms/detail/43.1258.TP.20201015.0908.002.html

CN 43-1258/TP
ISSN 1007-130X
计算机工程与科学
Computer Engineering & Science

信息传递增强的神经机器翻译*

史小静，宁秋怡，季佰军，段湘煜
（苏州大学，自然语言处理实验室，苏州 215006）

摘　要：神经机器翻译领域中多层神经网络模型结构能够显著提升翻译效果，但是多层神经网络结构存在信息传递的退化问题。为了缓解这一问题，提出了层间和子层间信息融合传递增强的方法。该方法有助于增强多层神经网络的层与层之间信息传递的能力。通过引入“保留门”机制来控制融合信息的传递权重，将融合信息与当前层的输出信息连接共同作为下一层的输入，使得信息传递更加充分。在目前最先进的多层神经网络结构 Transformer 上进行相关的实验，在中英和德英翻译任务上的实验结果表明，该信息传递增强方法相比于基线系统 BLEU 值分别提高 0.66 和 0.42。

关键词：神经网络；神经机器翻译；信息传递；信息退化；残差网络；门机制
中图分类号：H085　　**文献标识码：**A

Enhancing Information Transfer in Neural Machine Translation

SHI Xiao-jing, NING Qiu-yi, JI Bai-jun, DUAN Xiang-yu
(Natural Language Processing Laboratory, Soochow University, Suzhou, 215006, China)

Abstract: In the field of Neural Machine Translation(NMT), the multi-layer neural network model structure can significantly improve the translation performance. However, the structure of multi-layer neural network has an inherent problem with information transfer degeneracy. To alleviate this problem, we propose a novel method which is favorable to improve information transfer by fusing layers information and sublayers information. By introducing a retention gate mechanism to control the fused information transfer weight, which is aggregated with the output of the current layer and then serves as the input of the next layer, making fuller information transfer between layers. We carry out the experiments on the most advanced NMT model Transformer. Results on the Chinese-English and German-English show that our method indeed facilitates the flow of semantic information and improves the strong baseline by 0.66,and 0.42 BLEU respectively.

Key words: neural network；neural machine translation；information transfer；information degeneracy；residual network；gate mechanism

1 引言

神经机器翻译（NMT, Neural Machine Translation）[21][22][23]是完全采用神经网络完成

* **收稿日期:**2020－03－17；**修回日期：**2020－05－03
基金项目:目国家自然科学基金项目（项目号 61673289）
通信地址:215006 江苏省苏州市苏州大学计算机科学与技术学院
Address: School of Computer Science and Technology, Soochow University, Suzhou 215006,Jiangsu, P.R.China

源语言到目标语言的端到端的翻译系统，吸引了众多学者的关注和研究。最初，神经机器翻译模型是基于循环神经网络（RNN, Recurrent Neural Networks）[1]。注意力机制[2]是对神经机器翻译编码器-解码器框架的完善，使得模型能够在解码时只关注源语言句子中的一部分区域，动态地生成源语言信息，极大地提高了翻译译文的质量。为了提高神经机器翻译的表达能力，多层神经网络的结构被采纳使用，基于卷积神经网络（CNN，Convolutional Neural Network）[3]和自注意力机制(Self-Attention)[4]的多层神经网络的出现，使得神经机器翻译在多项翻译任务中达到了最先进的水平。

编码器-解码器（Encoder-Decoder）[5][6]框架是神经机器翻译模型的经典框架，可以由不同的神经网络实现，如循环神经网络（RNN）[1]，门限循环单元（GRU, Gated Recurrent Unit）[2][6]，长短期记忆神经网络（LSTM, Long Short-Term Memory）[7]，卷积神经网络（CNN）[3]，Transformer[4]等。一般地，编码端将源端的输入句子由低层到高层逐层地生成语义向量表示，解码端根据编码端最后一层的输出和已经解码生成的词的隐藏层表示作为输入通过一系列计算得到对应的目标端的译文。多层神经网络中逐层传递的过程往往存在梯度消失或者梯度爆炸的问题，残差网络(Residual Network)[8]的引入有效地缓解了梯度消失和梯度爆炸的问题，但多层神经网络的逐层信息传递过程中的信息退化问题仍然不能得到全部解决。

本文提出各层间信息融合传递增强和子层间信息融合传递增强的方法，在残差网络的基础上能够进一步补充多层神经网络逐层传递过程中的退化信息，保留之前所有层或子层的输出信息，再经过一个“保留门”机制来控制之前所有层的输出融合后的信息保留的权重比例，该“保留门”是通过网络自主学习得到的，与当前层的输出做一个连接，共同作为下一层网络的输入，使得多层网络中层与层之间的信息传递更加充分，优化和增强了层与层之间信息的传递的能力，本文提出的子层间信息融合传递增强方法在中英和德英翻译任务上提升了0.66 和 0.42 BLEU[9]，实验结果表明作者提出的方法，有效地提升了神经机器翻译的性能。

2 相关工作

近年来，神经机器翻译在具有大规模的平行语料上训练获得了较好的翻译性能，超越了传统的统计机器翻译（SMT，Statistical Machine translation）[25][26][27]。神经机器翻译仅需要句子级的平行语料采用神经网络来实现翻译过程[22][23][10]，便于训练大规模的翻译模型，具有很高的实用价值，吸引了众多研究者的关注。

注意力机制[2]是对神经机器翻译编码器-解码器框架的完善，它允许在不考虑输入和输出序列中的距离的情况下对依赖关系进行建模，使得模型能够利用平行语料学习到词与词之间的对齐信息。Vaswani[4]等人首次将自注意力机制引入神经机器翻译模型，该模型仅依赖于注意力机制就能完成源语言到目标语言的翻译，不仅在机器翻译方面达到了最先进的效果，而且在多个自然语言处理任务[10][11][12]上达到了最好的效果。自 Transformer 提出以来，众多研究者进一步提出不同的方法对该模型的各层之间的信息进行利用，Wang[13]等研究者提出通过添加另外一层网络来将其他层的层间信息进行

融合，以此来捕获被丢失的信息，He[14]提出编码端和解码端层与层之间的协调使得在解码时能够关注编码端不同层的输出信息，而不只是关注利用编码端最后一层的输出信息，该方法的使用超出了基准 Transformer 的翻译性能，Wang[15]等人提出通过对层正则化的改进并提出不同层的输出信息通过线性变换进行动态连接使得能够训练更深层次的 Transformer 模型，从而得到更好的效果。

基于编码器-解码器框架的神经机器翻译模型多是依赖于多层神经网络，然而多层神经网络结构在层与层之间信息传递过程中往往存在信息退化的问题，残差网络[8]的引入有效地缓解了逐层之间信息传递过程中的信息退化的问题，但是残差网络只是添加了最邻近前面一层的输出信息，未能利用其他层的输出信息。多层网络中固有的信息捕获不充分的问题并没有很好的解决。

与该文方法较相似的工作是 Yang[16]等人为了使得自注意力机制获得更好的词间的关联性提出利用不同的策略来捕获多层神经网络的上下文信息，经过一个门控权重将上下文信息引入到注意力模型输入(query 和 key)中，以使得自注意力机制能够捕获到更好的词间关联信息。

该文提出的各层和子层间信息融合传递增强的方法，在残差网络的基础上能够进一步补充多层神经网络逐层传递过程中退化的语义信息。通过保留之前所有层或子层的输出信息并进行融合，再经过一个“保留门”机制来控制融合信息作为下一层输入信息的权重，与当前层的输出做一个连接，共同作为下一层网络的输入，不仅对自注意力机制的输入 query 和 key 进行补充，同时对 value 的信息也进行增强，使得多层网络中层与层之间的信息传递更加充分，优化和增强了层与层之间信息传递的能力。在中英和德英翻译任务上的实验结果表明，本文提出的方法，有效地提升了神经机器翻译的性能。

3 模型结构

3.1 Transformer 模型

Transformer[4]模型架构首先是由 Vaswani 等研究者提出，属于编码器-解码器结构，编码器与解码器之间通过注意力机制实现连接。Transformer 的编码器端和解码器端分别是由 N 层相同的网络层组成，编码器端的每一层都包含两个子层，多头自注意力机制和全连接的前馈网络层，每一个子层的输出都进行了残差连接和层归一化处理。与编码器端相似，解码器端的每一层中除了包含自注意力机制层和全连接的网络层之外，还包含了一个编码器-解码器注意力子层，同样地，对解码端的每一个子层的输出进行了残差连接和层归一化处理。

Transformer 中使用多头注意力机制，多头注意力机制使得模型在不同位置联合处理来自不同表示空间的信息，将向量Q、向量K和向量V的切分成更小维度的向量，希望通过不同视角获取多样的注意力信息。通常是将向量Q、向量K和向量V维度设置为 512 维，分为 8 个头，即每个向量被分为 8 份，每一份维度为 64 维。然后计算每个头的注意力信息，最终将各个头的注意力信息进行拼接再次映射成 512 维。多头注意力机制的核心是缩放点乘注意力如公式(1,2,3)。

$$MultiHead = Concat(head_1, \dots, head_h)W^o \quad (1)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

其中映射矩阵 $W_i^Q \in \mathbb{R}^{d_{model}*d_k}$，$W_i^K \in \mathbb{R}^{d_{model}*d_k}$，$W_i^V \in \mathbb{R}^{d_{model}*d_v}$，$W^O \in \mathbb{R}^{hd_v*d_{model}}$

Transformer 模型完全基于注意力机制，为了充分利用位置信息，添加了位置编码操作，位置编码采用正余弦位置编码方式如公式(4,5)。

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (4)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (5)$$

其中 pos 表示当前词在该序列中的位置编号，i 指向量表示的某一维度。

3.2 损失函数

Transformer 中使用交叉熵作为损失函数，具体地，给定一对平行语料，其中源端 $X = \{x_1, x_2 \dots x_n\}$，目标端为$Y = \{y_1, y_2 \dots y_m\}$，在训练过程中通过最大化似然估计来生成目标端候选词的最大化概率$\mathrm{P}(y_i|y_{[1:(i-1)]}, X)$，训练过程中的损失函数公式如公式(6)。

$$L = \frac{1}{m}\sum_{i=1}^{m} -\log(P(y_i|y_{<i}; X)) \quad (6)$$

其中 m 表示目标端句子的长度，$p(y_i|y_{<i}; X)$表示的是目标端的第 i 个词的概率。

4 层间信息传递增强的方法

Transformer 是多层神经网络结构，存在信息传递的过程中的信息退化问题。本文提出利用 Transformer 中的每一层的层间输出信息和每一个子层的输出信息融合来增强不同层和子层之间的信息传递，使得模型能够捕获更丰富的语义信息，弥补逐层之间信息传递过程中丢失的信息，进一步减缓了多层神经网络中的信息退化问题。

4.1 各层间信息传递增强

基准系统 Transformer 中编码器和解码器通常均为六个相同层的堆叠，当前层的输入仅基于最邻近的上一层的输出信息，而忽略了更底层的输出信息，这样很容易导致信息传递过程中的退化问题。针对这一问题，本文提出各层间信息融合传递增强的方法。

各层间信息融合传递增强的模型结构如图 1 所示。图中每一个完整层计算之后都有一个流动箭头表示对当前第 i 层的输出信息$layer_i$进行保留并流入未来每一层的输入中。在进行下一层的计算之前，将保留的低层输出语义信息$[layer_1, layer_2, \dots, layer_{i-1}]$进行融合，与当前层的输出$layer_i$通过保留门机制进行连接，共同作为下一层的输入信息。利用补充的低层信息，能够增强神经网络的表示特征的捕获能力，使得神经网络能够捕获更丰富的语义信息。其中保留门控制低层融合后的信息流入下一层计算的保留权重。本文使用的信息融合的方法包括线性变换融合和算术平均融合。具体如公式(7,8,9)。

$$fusion_{i-1} = f(layer_1, layer_2, .., layer_{i-1}) \quad (7)$$

$$g = \sigma(W[fusion_{i-1}; layer_i] + b) \quad (8)$$

$$layer_{i+1} = LN\big((1-g)*layer_i + g*fusion_{i-1}\big) \quad (9)$$

其中，$layer_i$表示第 i 层的输出，公式(7)表示对保留的信息进行融合，f 表示信息融合函数，比如算术平均或线性变换等，$fusion_{i-1}$表示网络前 i-1 层的输出融合后的信息，公式(8)表示将融合后的信息与当前层的输出信息进行连接，为了更好地控制融合信息的保留权重，通过 sigmoid 函数将其值介于 0-1 之间，$W \in \mathbb{R}^{d_{model}*2}$为可训练权重矩阵，公式(9)表示当前层信息和融合后的信息的以不同的概率连接，LN表示层正则化函数，*表示对应元素乘积，我们引入的保留门机制是通过网络自主学习之前每一层的保留信息权重，模型在训练过程中会自动学习需要多少比例的之前每一层的输出信息，控制每一层的信息保留程度。

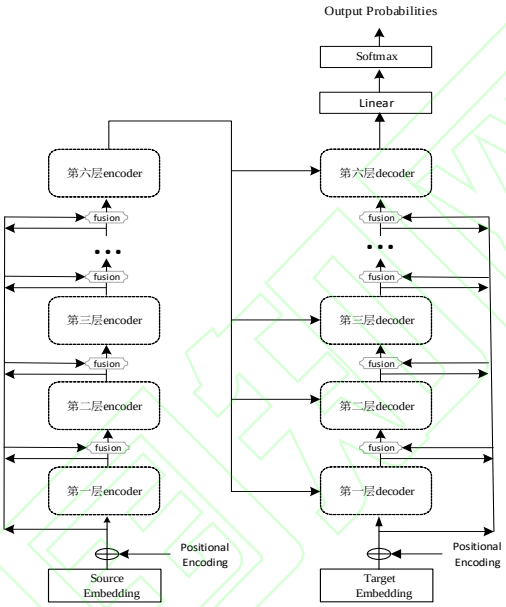

Figure 1 The structure of layers information transfer enhance

图 1 层间信息传递增强结构

4.2 子层间信息传递增强

Transformer 模型中每一个完整的层中包含若干个子层，每一个子层的输出进行残差连接作为下一个子层的输入信息。为了更充分地利用之前所有子层的输出信息，本文提出子层间信息传递增强的方法。

子层间信息传递增强的方法如图 2 所示，图中的虚线部分表示基准系统 Transformer 原有的残差网络，实线部分表示增加的子层信息传递增强的方法。图中每一个子层的输出均有一个实线流动箭头表示对第 i 层的第 j 个子层信息$sublayer_{i,j}$进行保留并流入未来每一子层的输入中，在进行下一个子层的计算之前，将前面保留的所有子层的输出信息$[sublayer_{1,1}, sublayer_{1,2i}, \dots, sublayer_{i,j-1}]$进行融合，然后通过门机制与当前子层的输出$sublayer_{i,j}$进行连接，共同作为下一子层的输入信息，具体如公式(10,11,12)。

$$fusion_{ij-1} = f\big(sublayer_{1,1}, \dots, sublayer_{i,j-1}\big) \quad (10)$$

$$g = \sigma\big(W[fusion_{ij-1}, sublayer_{ij}] + b\big) \quad (11)$$

$$sublayer_{ij+1} = LN\left((1-g)*sublayer_{ij} + g*sublayer_{ij-1}\right) \quad (12)$$

其中，$sublayer_{ij}$表示第 i 层的第 j 个子层的输出，公式(10)表示对保留的信息进行融合，f表示融合函数，$fusion_{ij-1}$表示第 i 层第 j-1 个子层之前所有子层的输出融合后的信息，公式(11)表示将融合后的信息与当前子层的输出信息进行连接，为了控制融合后的信息保留权重，使用 sigmoid 函数将其数值限制与 0-1 之间，$W \in \mathbb{R}^{d_{model}*2}$是可训练权重矩阵，通过网络自动学习融合信息的保留权重，公式(12)表示将融合后的信息和当前的输出信息以一定的概率进行连接，LN为层正则化函数，*表示对应元素乘积。

此外，为了验证模型低层的输出在层与层之间信息流动中的重要性，基于基准系统，对所有保留的子层信息$[sublayer_{1,1},\dots,sublayer_{i,j-1}]$进行下一层运算时没有及时添加，而是仅在最后一层的输出上进行权重连接的实验，作者称之为子层信息连接，来验证逐层信息增强的重要性，其计算方法如公式(13,14,15,16)。

$$sublayers = \left[sublayer_{1,1},\dots,sublayer_{i,j-1}\right] \quad (13)$$

$$fusion = f(sublayers) \quad (14)$$

$$g = \sigma\left(W\left[fusion, sublayer_{t-1,3}\right] + b\right) \quad (15)$$

$$layer_t = LN\left((1-g)*layer_{t-1,3} + g*fusion\right) \quad (16)$$

其中，t为解码器的堆叠层数，$sublayers$为保留的解码器端所有子层的输出信息，f为融合函数，$layer_t$解码器端的最后一层的输出，LN为层正则化函数。

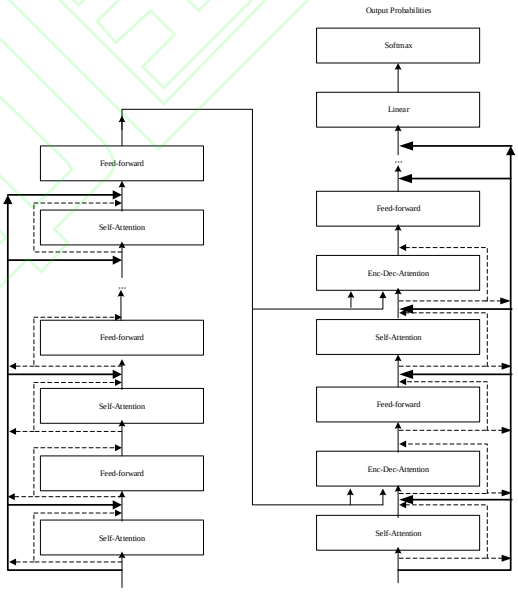

Figure 2 The structure of sublayers information fusion transfer enhance

图 2 子层间信息融合传递增强

5 实验

5.1 数据集

本文在中英和德英翻译任务上进行实验来

验证作者方法的有效性。

中英翻译任务中使用的训练数据集是从LDC（Linguistic Data Consortium，简称：LDC）里抽取的125万句的中英平行语句对，测试集采用的是美国国家标准与技术研究院 2002 年的数据NIST02、NIST03、 NIST04、 NIST05和 NIST08，验证集采用的是NIST06，分别对中英语料进行BPE编码处理[18]，中文和英文分别使用BPE表，表的大小均为32000，中文设置词表大小为4万，英文设置词表大小为3万，不在词表中的低频词使用特殊符号“<UNK>”替换。

德英翻译任务中使用的数据是IWSLT2016平行语料，训练集包含18.5万的口语语料平行句对，验证集包含8415句平行语句对，测试集包括7883句平行语句对，分别对德英进行BPE编码处理[18]，表的大小均为32000，德文和英文词表大小均设置为3万，不在词表中的低频词使用特殊符号“<UNK>”替换。

5.2 实验设置

本文的所有的实验方法的实现都是基于开源代码 fairseq[19]，将训练模型设置为Transformer。所有模型的训练和测试均是基于NVIDIA GeForce GTX 1080 GPUs，设置一个batch中最多包含5000个词，在所有的实验中均使用相同的超参数，采用 Adam[20]优化器和逆平方根学习率计划，初始化学习率为0.0005，$\beta 1 = 0.9$，$\beta 2 = 0.99$，$\varepsilon = 10^{-9}$，在训练期间，作者在所系统上的实验均设置dropout为0.3并且设置了标签平滑[17]的值为 0.1。在解码的时候，采用束搜索（Beam Search）的解码方式，搜索宽度设置为5，其他实验设置与Vaswani等人[4]相同。

5.3 实验结果及分析

5.3.1 中英实验结果分析

中英翻译任务上的实验结果如表1所示。从表1中可以看出，基准系统在5个测试集上的平均BLEU得分为44.47。

仅在解码器端添加层间信息和子层间信息增强平均得分分别为44.81和44.98，相比于基准系统分别提升0.34和0.51，可以得出子层信息增强的方法更优于层间信息增强，此外，对比层间信息增强和子层间信息增强的不同的信息融合方法，可以看出，算术平均融合的效果更优。

仅在编码器端添加层间信息和子层间信息增强平均得分分别为44.65和44.73，该实验结果相比于基准系统略微提升，但是对比实验结果仍可得出子层信息增强的方法更优于层间信息增强，并且，算术平均融合的效果更优，这与仅在解码器端添加信息增强得到的结果一致。

在编码器和解码器端分别添加层间信息和子层间信息增强平均得分分别为45.10和45.13，相比于基准系统分别提升0..63和0.66，再次表明了子层信息增强的方法更优于层间信息增强，以及算术平均融合的效果更好。

为了对比信息传递增强和残差网络在多层神经网络中的影响，作者移除了残差网络仅使用子层间信息传递增强方法，实验结果如表 1中第2行，从表格中的数据可以看出，子层间信息传递增强和残差网络的平均分相差 0.02，结果几乎持平，说明了子层信息传递增强方法的有效性。

5.3.2 德英实验结果分析

德英翻译任务上的实验结果如表 2 所示。从表 2 中可以看出，基准系统在测试集上的 BLEU 得分为 34.50。仅在解码器端添加层间信息和子层间信息增强的得分为 34.61 和 34.73，相比于基准系统提升了 0.23。仅在编码器端添加层间信息和子层间信息增强的得分为 34.56 和 34.68，实验结果略微高于基准系统。在编码器和解码器端分别添加层间信息和子层间信息增强平均得分分别为 34.75 和 34.92，较于基准系统分别提升了 0.25 和 0.42。

通过对中英和德英的实验结果的对比和分析，发现子层间的输出信息对模型的语义信息补充更加充分，这与作者的直观猜想一致，同样也说明了残差网络的重要性。此外，通过实验观察不同的信息融合方法，可以得出，算术平均的方法进行信息融合的效果更好。

Table 2 Experimental results of German-English translation task

表 2 德英翻译任务实验结果

实验系统		Test
基准系统		34.50
解码端信息增强	各层间信息算术平均融合传递增强	34.61
	子层间信息算术平均融合传递增强	34.73
编码器端信息增强	各层间信息算术平均融合传递增强	34.56
	子层间信息算术平均融合传递增强	34.68
编码器和解码器信息增强	各层间信息算术平均融合传递增强	34.75
	子层间信息算术平均融合传递增强	34.92

Table 1 Experimental results of Chinese-English translation task

表 1 中英翻译任务实验结果

	实验系统	NIST02	NIST03	NIST04	NIST05	NIST08	AVG
	基准系统	46.21	46.59	46.43	46.22	36.91	44.47
	子层信息传递增强取代残差网络	46.66	45.84	46.55	46.73	36.47	44.45
解码器端信息增强	子层信息连接	46.68	46.28	46.57	46.63	36.90	44.61
	各层间信息线性变换融合传递增强	46.60	45.98	46.63	46.72	37.01	44.59
	各层间信息算术平均融合传递增强	46.68	46.30	46.70	46.72	37.63	44.81
	子层间信息线性变换融合传递增强	46.52	46.60	46.82	46.76	37.31	44.80
	子层间信息算术平均融合传递增强	47.01	46.86	47.19	46.64	37.19	**44.98**
编码器端信息增强	各层间信息线性变换融合传递增强	46.42	46.23	46.63	46.44	36.91	44,53
	各层间信息算术平均融合传递增强	45.8	46.27	46.85	46.98	37.35	44.65
	子层间信息线性变换融合传递增强	45.86	46.06	46.82	46.91	37.38	44.61
	子层间信息算术平均融合传递增强	46.59	46.45	47.06	46.83	36.72	44.73
编码器端和解码器端信息增强	各层间信息算术平均融合传递增强	46.90	46.67	47.23	47.63	37.10	45.10
	子层间信息算术平均融合传递增强	46.97	46.71	47.27	47.59	37.10	**45.13**

5.4 层间信息融合传递增强对不同句长的翻译影响

为了进一步分析不同的方法对于不同句长输入的翻译质量，作者将所有的测试集根据句子长度分为 6 个不同的部分，分别测试了不同系统对不同长度句子的翻译效果。实验结果如表 3 所示，其中句长为 21 到 30 是测试集中最多的部分共有 1683 句，句长为 10 到 20 和 30 到 40 的分别有 1635 句和 1222 句。从表中可以看出，子层间信息融合传递增强在所有句长上的测试水平均高于基准系统，其中句长大于 50 的句子中提升 0.74BLEU，可见子层间信息融合传递增强处理较长的句子更具优势。

Table 3 BLEU score of different sentence length in Chinese-English translation system

表 3 中英翻译系统不同句子长度的 BLEU 得分

实验系统 \ 句子长度		1-10	11-20	21-30	31-40	41-50	>50
	基准系统	41.30	45.27	45.21	44.80	44.72	43.16
	子层信息传递增强取代残差网络	41.64	45.44	45.14	45.65	45.10	43.48
编码器和解码器	各层间信息算术平均融合传递增强	41.20	46.42	45.71	45.79	45.31	43.09
	子层间信息算术平均融合传递增强	43.3	45.87	45.57	**46.02**	45.09	43.27

Table 4 Translation comparison of different models

表 4 不同模型的译文对比

中文句子	柏林墙 最后 成为 了 历史 ， 从 中 受益 的 不 仅仅 是 德国人 。
参考译文	the berlin wall has finally become history , and it was not only the germans who have benefited from it .
	the berlin wall has finally become history and the germans are not the only ones that benefit from the wall 's fall .
	berlin wall has become history from which not only germans can learn a lesson .
	the berlin wall finally becomes history , and not only the germans benefit .
基准系统	finally , the berlin wall became history , and not only the germans .
编码器和解码器端各层间信息算术平均融合传递增强	the berlin wall finally became history , and not only germans benefited from it .
编码器和解码器端子层间信息算术平均融合传递增强	the berlin wall finally became history and not only the germans benefited .
中文句子	这 场 于 上 月 28 日 凌晨 爆发 的 战斗 是 在 联合国 调查团 调查 利比里亚 政府 用 钻石 换 武器 和 对 利比里亚 实施 制裁 的 情况 下 发生 的 。
参考译文	this fighting that broke out on the early morning of the 28th of last month occurred at a time when a un investigation team was investigating the liberian government 's exchange of diamonds for weapons and imposing sanctions on liberia .
	the current fight broke out at daybreak on february 28th right after united nations (un) investigation of alleged trading of diamonds for weapons by liberian government . un has since imposed sanctions on liberia .
	the war erupted on the 28th last month when the un investigation group was investigating the liberian government 's trading of diamonds for arms against a backdrop of sanctions against liberia .
	the battle exploded on february 28 took place under the circumstance that the un delegation investigated the liberia government 's trading diamonds for arms and the sanction over liberia was in effect .
基准系统	the clash erupted last month 28 when the un investigation team investigated the government of liberia for using diamonds to replace weapons and imposed sanctions on liberia .
编码器和解码器端各层间信息算术平均融合传递增强	the battle broke out **daybreak** on the 28 th of last month when the un investigation team investigated the use of diamond for weapons and the imposition of sanctions against liberia .
编码器和解码器端子层间信息	the battle broke out **early in the morning of** 28 of last month when the un investigative team investigated the liberia

算术平均融合传递增强	government using diamond to replace weapons and impose sanctions against liberia .

5.5 译文质量分析

表 4 是作者从测试集中随机抽取的句子和不同的方法得到的相应的译文，对句子进行分析，基准系统中丢失了原文中“从中受益”的翻译，与之相比作者的各层间信息融合传递增强和各子层间信息融合传递增强分别翻译为“benefited from it”和“benefited”翻译准确无误并和参考译文相符。对于较长句子，基准系统漏翻了“凌晨”一词，作者提出的方法，分别将其翻译为”**daybreak**”和”**early in the morning**”与参考译文相符，并且翻译无误。

6 结束语

本篇文章中，作者提出层间信息融合传递增强和子层间信息融合传递增强的方法，在残差网络的基础上能够进一步补充多层神经网络逐层传递过程中的退化信息。通过保留之前所有层或子层的输出信息并将保留的信息使用不同的方法进行融合，再经过一个“保留门”机制来控制之前所有层或子层的输出融合后的信息保留的权重比例，与当前层或子层的输出做一个连接，共同作为下一层或下一子层网络的输入。使得多层网络中层与层之间的信息传递更加充分，优化了神经网络的信息捕获能力，增强了逐层信息的传递。未来的工作中，作者将探索更加复杂的连接方式，使得模型的效果能够得到进一步提升。最后，感谢“江苏高校优势学科建设工程资助项目”对本论文的支持。

参考文献：

[1] Cho, Kyunghyun and van Merrienboer, Bart and Gulcehre, Caglar,et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar:Association for Computational Linguistics,2014.

[2] Dzmitry Bahdanau,Kyunghyun Cho,and Yoshua Bengio. Neural machine translation by jointly learning to align and translate[C]//In Proc of ICLR, 2015.

[3] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, NSW, Australia: JMLR.org, 2017.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Atten-tion is all you need[C]//Advances in Neural Information Processing Systems 30. Long Beach: Curran Associates, Inc., 2017.

[5] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems. Montréal CANADA :Curran Associates, Inc, 2014.

[6] Luong, Thang and Pham, Hieu and Manning, Christopher D.Effective Approaches to Attention-based Neural Machine Translation.[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.Lisbon, Portugal:Association for Computational Linguistics,2015.

[7] Hochreiter, Sepp, Schmidhuber, Jürgen. Long Short-Term Memory[J]. Neural Computation,1997, 9(8):1735-1780.

[8] He K,Zhang X,Ren S,et al.Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.Nevada,USA,2016.

[9] Kishore Papineni,Salim Roukos and Todd Ward et al,Bleu:a Method for Automatic Evaluation of Machine Translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.Philadelphia, PA, USA:ACL,2002.

[10] Lu, Zhengdong and Li, Hang.A Deep Architecture for Matching Short Texts[J]. Advances in neural information processing systems, 2013:1367-1375.

[11] Hassan, Hany and Aue, Anthony and Chen, Chang,et al, Achieving Human Parity on Automatic Chinese to English News Translation[J]. CoRR,2018, abs/1803.05567: 1803-05567.

[12] Paulus R,Xiong C, Socher R . A Deep Reinforced Model for Abstractive Summarization[C]// 6th International Conference on Learning

Representations, {ICLR}, Vancouver, BC, Canada: Conference Track Proceedings, 2018.

[13] Qiang Wang and Fuxue Li and Tong Xiao et al.Multi-layer Representation Fusion for Neural Machine Translation[C]//Proceedings of the 27th International Conference on Computational Linguistics, {COLING} 2018, Santa Fe, New Mexico, USA:Association for Computational Linguistics.2018.

[14] Tianyu He and Xu Tan and Yingce Xia,et al.Layer-Wise Coordination between Encoder and Decoder for Neural Machine Translation[J].CoRR,2018,8019:7955-7965.

[15] Wang Q , Li B , Xiao T , et al. Learning Deep Transformer Models for Machine Translation[J].CoRR,2019,abs/1906.01787:1810-1822.

[16] Yang B , Li J , Wong D , et al. Context-Aware Self-Attention Networks[C].//. the Association for the Advance of Artificial Intelligence. New York,USA:2019.

[17] Szegedy C,Vanhoucke V,Ioffe S ,et al.Rethinking the Inception Architecture for Computer Vision[J].CoRR,2015,abs/1512.00567: 2818-2826.

[18] Sennrich R , Haddow B , Birch A . Neural Machine Translation of Rare Words with Subword Units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics,2016 .

[19] Myle Ott and Sergey Edunov and Alexei Baevski,et al.fairseq: A Fast, Extensible Toolkit for Sequence Modeling[J]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, {NAACL-HLT} , 2019,abs/1904.01038:48-53.

[20] Kingma, Diederik, Ba, Jimmy. Adam: A Method for Sto-chastic Optimization[C]//3rd International Conference on Learning Representations, {ICLR}.San Diego, CA, USA:Conference Track Proceedings,2015.

附中文参考文献:

[21] 刘洋 . 神经机器翻译前沿进展[J]. 计算机研究与发展,2017,54(6):1144-1149.

[22] 李亚超, 熊德意, 张民. 神经机器翻译综述[J]. 计算机学报, 2018, 41(12):100-121.

[23] 高明虎, 于志强. 神经机器翻译综述[J]. 云南民族大学学报(自然科学版), 2019, 28(1):76-80.

[24] 董陆森. 机器翻译中的常用神经网络模型[J]. 电子技术与软件工程, 2018, 132(10):163.

[25] 刘群.统计机器翻译综述[J].中文信息学报,2003(4):1-12.

[26] 郎君. 统计机器翻译中翻译模型的约简概述[J]. 智能计算机与应用, 2011,(3):17-20.

[27] 戴新宇, 尹存燕, 陈家骏, et al. 机器翻译研究现状与展望[J]. 计算机科学，2004,(11):178-181+186.