注意力链接：一种高效的基于注意力的低资源机器翻译架构

Zeping Min北京大学数学
科学学院中国北京
zpm@pku.edu.cn

arXiv:2302.00340v1 [cs.CL] 1 Feb 2023

摘要—**Transformer 在机器翻译领域取得了巨大成功，但基于 Transformer 的神经机器翻译模型通常需要数百万规模的双语平行语料库进行训练。本文提出了一种名为注意力连接（AL）的新型架构，以提升 Transformer 模型的性能，尤其是在低训练资源条件下。我们从理论上证明了注意力链接架构在低训练资源条件下的优越性。此外，我们进行了大量实验，包括在 IWSLT14 数据集上的英语-德语、德语-英语、英语-法语、英语-意大利语、意大利语-英语和英语-罗马尼亚语翻译任务，以及在 CVIT PIB 数据集上的孟加拉语-古吉拉特语和古吉拉特语-泰米尔语翻译任务，以测试其在真实低资源场景中的表现。所有实验结果都表明，注意力连接功能强大，能够带来显著提升。此外，通过结合注意力连接与其他先进方法，我们在 IWSLT14 德语-英语任务上取得了 37.9 的 BLEU 分数，达到新的最先进水平。**

索引词—**Transformer、低资源、机器翻译、理论分析**

I. 引言

机器翻译任务是自然语言处理领域最重要的任务之一。一般而言，机器翻译有三种范式：基于规则的机器翻译、统计机器翻译和神经机器翻译。基于规则的机器翻译，例如 [1], [2], [3],[4] 和 [5]，是最传统的机器翻译方法，通常依赖专家提取的固定规则来运行，例如统计机器翻译 [6], [7] [8] 和 [9]。统计机器翻译通常通过对大量平行语料库进行统计分析来构建统计翻译模型，然后利用该模型进行翻译。近年来，神经机器翻译（NMT）技术取得了显著突破。在注意力机制出现之前，最常用的神经机器翻译模型是循环神经网络[10], 、长短期记忆网络 [11], 、门控循环单元 [12] ，以及许多基于这些模型的研究，例如 [13], [14], [15]和 [16]。尽管基于循环神经网络和长短期记忆网络的模型都取得了巨大成功，但它们仍面临一些棘手的问题。例如，两者在翻译长句时都表现不佳。

A. 基于注意力的神经机器翻译

[17] 尝试在机器翻译任务中使用注意力机制。注意力机制的设计灵感源于模仿人类注意力，从大量信息中快速筛选出高价值信息。 [18]提出了一种完全基于注意力机制的神经机器翻译模型。该模型在 WMT数据集上取得了出色的结果，展现了注意力机制在神经机器翻译任务中的巨大潜力。此后，各种基于注意力机制的神经机器翻译模型成为机器翻译任务中的主流模型。例如， [19], [20], [21] 和 [22] 尝试让 Transformer 更轻松地处理长句。 [23] 和 [24] 尝试让 Transformer 训练得更快、占用更少内存。 [25] 尝试训练深层 Transformer。毫无疑问，Transformer模型在机器翻译领域取得了巨大成功，但基于 Transformer 的语言模型通常需要数百万个双语平行语料库样本进行训练，而且往往需要仔细调整超参数。对于某些特定语言对，很难获得足够的平行语料库。本文提出了一种易于实现且有效的架构，称为注意力连接。理论分析和实验结果证明了注意力链接架构的有效性。我们认为，注意力连接能够轻松捕获句子中更通用的语义信息，并且具有较强的抗噪能力。此外，我们的架构可以轻松部署在基于 Transformer 的模型中，而无需增加额外参数。我们的贡献如下：

* 我们提出了注意力连接（AL），这是一种新颖的架构，可帮助提升 Transformer模型的性能，尤其是在训练资源有限的情况下。 * 我们从理论上阐释了注意力链接架构的优越性。 * 大量实验验证了注意力连接的有效性。所有实验结果均表明，我们提出的注意力连接性能强大，能够带来显著提升，尤其是在低资源翻译条件下。此外，我们在 IWSLT14德英数据集上取得了新的最先进水平。

II. 相关工作

A. 通过额外语料库改进

提升低资源神经机器翻译性能的一种方法是使用额外语料库，包括数据增强方法，以及其他更容易获取的语料库。获取某种语言的单语语料库，要比获取某个语言对的双语平行语料库容易得多。预训练 [26] 和回译 [27], [28] 和 [29] 是利用单语语料库的两种方法。预训练通过学习良好的表示来提升性能。

例如， [30]. [31] 通过预测句子中的掩码词来学习表示，这种方法已在众多自然语言处理任务中取得了巨大成功。

回译[27], [32]是一种经典且常用的方法。回译 [27] 的主要思路是利用单语语料库来提升模型的性能。回译通常能显著提升低资源条件下的机器翻译性能，并且已有许多关于回译方法的研究。

B. 通过模型改进

此外，还有许多研究致力于改进模型，使其更适用于低资源条件。

改进模型最简单的方法之一是选择合适的超参数。 [33] 和 [34] 通过选择合适的 Transformer 超参数提升了性能。此外，也有一些研究专注于改进模型本身。 [35] 将编码器和解码器中的所有自注意力矩阵替换为固定的高斯分布。然而，所得模型的性能并未显著下降。[36] 固定 Transformer 编码器的注意力矩阵，可使 BLEU 分数提高 3 分。在这项工作中，我们也将重点放在注意力矩阵上。针对低资源条件，我们从注意力矩阵的角度提出了一种自适应架构——注意力连接。

III. 注意力链接架构

本节首先介绍我们设计注意力连接的见解与动机，然后详细描述注意力链接架构及其数学公式。

A. 动机

一般而言，低资源带来的困难很大一部分源于难以获取充足的信息。过度训练更容易导致严重的过拟合。具体来说，当我们尝试在低资源条件下训练基于 Transformer 的神经机器翻译系统时，Transformer模型可能难以提取语义信息，而只能简单地记忆样本。过度记忆可能导致模型过于关注少量词语（子词）之间的相关性，而缺乏全局视野。Transformer模型中的注意力矩阵反映了模型对词语（子词）之间相关性的判断。

我们首先从注意力矩阵入手，以验证我们的推测。首先，我们从英语-法语 IWSLT14 数据集中随机选取约1万和3万句平行语料库，以模拟低资源条件，分别在1万、3万训练资源和完整数据集上训练Transformer模型。随后，我们在图1中将不同训练资源水平下训练的Transformer模型在一个随机选取的句对上的注意力矩阵进行可视化。我们可以观察到，对于在低资源条件下训练的Transformer模型，其自注意力矩阵和交叉注意力矩阵都明显更加稀疏。这意味着它们将大部分注意力集中在少数词对上，缺乏全局视野。这在某些句对上可能表现良好（例如模型记忆的样本），但由于缺乏全局视角，模型将缺乏泛化能力。

受 ResNet [37], 中快捷结构的启发，我们提出了一种注意力链接结构。由于注意力矩阵旨在捕捉词语之间的关系，为减轻低资源的影响，我们尝试借鉴 ResNet [37] 中的快捷结构，将相邻层的注意力矩阵连接起来。通过注意力连接，注意力矩阵之间的信息可以共享，从而降低单层注意力矩阵受到噪声干扰的可能性。

B. Transformer 和注意力连接

我们首先简要介绍 Transformer模型，明确一些符号，然后给出我们的注意力链接架构。近年来，Transformer模型 [18]已成为许多领域的基石。Transformer模型的编码器和解码器分别由编码器层和解码器层串联而成。编码器层和解码器层主要由自注意力部分、交叉注意力部分和前馈神经网络部分组成。这些层将 $\underline{\boldsymbol{X}} \in \mathbb{R}^{d\times n}$ 映射为 $\boldsymbol{X} \in \mathbb{R}^{d\times n}$。数学上，如 [38],中的公式所示，我们有

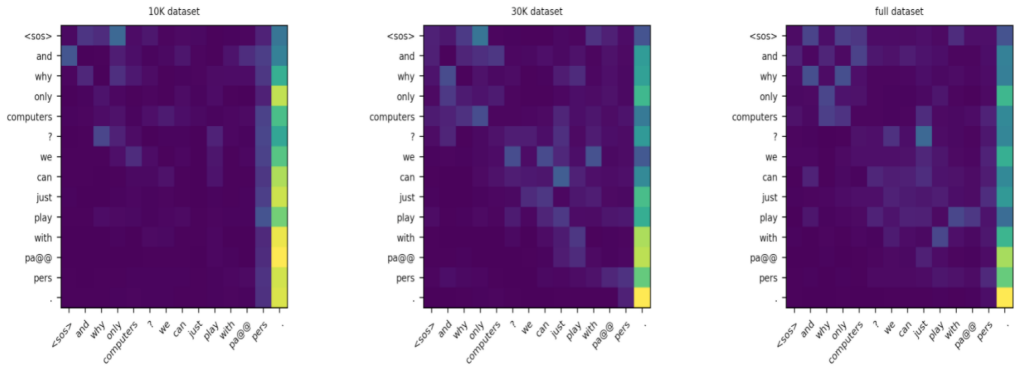
$$\begin{aligned}\mathrm{Self_Attn}_n(\boldsymbol{X}) = \sum_{i=1}^{h} \boldsymbol{W}_{n,O}^i \boldsymbol{W}_{n,V}^i \boldsymbol{X} \cdot \\ S\left[\left(\boldsymbol{W}_{n,K}^i \boldsymbol{X}\right)^T \boldsymbol{W}_{n,Q}^i \boldsymbol{X}\right]\end{aligned} \tag{1}$$

$$\begin{aligned}\mathrm{Cross_Attn}_n(\boldsymbol{X}) = \sum_{i=1}^{h} \boldsymbol{W}_{n,O}^i \boldsymbol{V} \cdot \\ S\left[(\boldsymbol{K})^T \boldsymbol{W}_{n,Q}^i \boldsymbol{X}\right]\end{aligned} \tag{2}$$

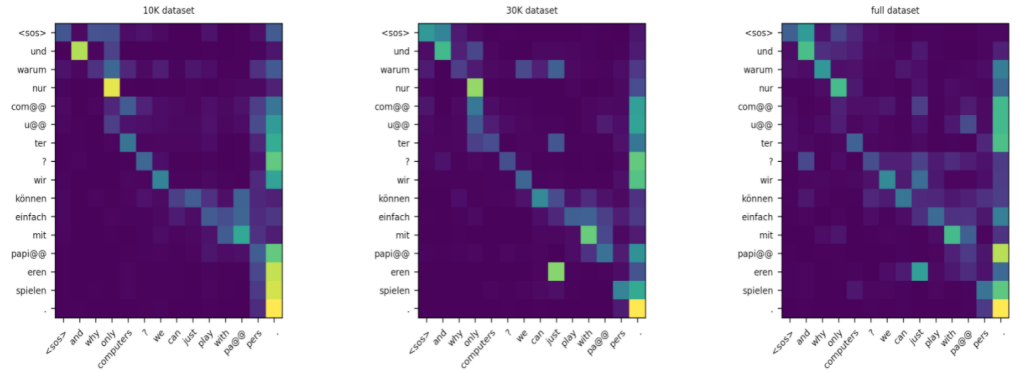
$$\mathrm{FFN}(\boldsymbol{X}) = \boldsymbol{W}_2 \cdot \mathrm{ReLU}\left(\boldsymbol{W}_1 \cdot (\boldsymbol{X}) + \boldsymbol{b}_1 \mathbf{1}^T\right) + \boldsymbol{b}_2 \mathbf{1}^T \tag{3}$$

其中，$\mathrm{Self_Attn}_n$ 表示第 n 个编码器层或第 n 个解码器层的自注意力部分。$\mathrm{Cross\ Attn}_n$表示第 n 个解码器层的自注意力部分。$\boldsymbol{W}_{n,O}^i \in \mathbb{R}^{d\times d_v}$、$\boldsymbol{W}_{n,V}^i \in \mathbb{R}^{d_v\times d}$、$\boldsymbol{W}_{n,K}^i \in \mathbb{R}^{d_k\times d}$、$\boldsymbol{W}_{n,Q}^i \in \mathbb{R}^{d_q\times d}$ $\boldsymbol{W}_2 \in \mathbb{R}^{d\times d_{hidden}}$ $\boldsymbol{W}_1 \in \mathbb{R}^{d_{hidden}\times d}$ $\boldsymbol{b}_2 \in \mathbb{R}^d$ $\boldsymbol{b}_1 \in \mathbb{R}^{d_{hidden}}$. 以及 d、d_q、d_k、d_v、d_{hidden}、h 是 6 个主要的超参数。 d 表示文本嵌入维度， d_q表示查询向量维度， d_k 表示键向量维度， d_v 表示值向量维度， d_{hidden} 表示隐藏层维度， h表示 Transformer头的数量。 S 表示 Softmax函数。 K 表示最后一个编码器的键张量。 V表示最后一个编码器的值张量。$S\left[\left(\boldsymbol{W}_{n,K}^i \boldsymbol{X}\right)^T \boldsymbol{W}_{n,Q}^i \boldsymbol{X}\right]$ 表示第 n 个编码器层或第 n 个解码器层中第 i 个头的自注意力矩阵。$S\left[(\boldsymbol{K})^T \boldsymbol{W}_{n,Q}^i \boldsymbol{X}\right]$ 表示第 n 个解码器层中第 i 个头的交叉注意力矩阵。

受 ResNet [37], 中快捷连接结构的启发，我们的注意力链接结构旨在连接注意力矩阵



(a) 编码器中的自注意力矩阵



(b) 解码器中的交叉注意力矩阵

Fig. 1. 不同资源条件下训练的Transformer模型的注意力矩阵（英语） -法语翻译任务

之间的相邻层。通过注意力连接，注意力矩阵之间的信息得以共享，从而降低单层注意力矩阵受到噪声干扰的可能性。从数学上讲，我们有

$$\begin{aligned}\text{Linked_Self_Attn}_n(\boldsymbol{X}) = \sum_{i=1}^{h} \boldsymbol{W}^i_{n,O}\boldsymbol{W}^i_{n,V}\boldsymbol{X}\cdot \\ S[(\boldsymbol{W}^i_{n,K}\boldsymbol{X})^T \boldsymbol{W}^i_{n,Q}\boldsymbol{X} + (\boldsymbol{W}^i_{n-1,K}\boldsymbol{X})^T \\ \boldsymbol{W}^i_{n-1,Q}\boldsymbol{X}]\end{aligned} \tag{4}$$

$$\begin{aligned}\text{Linked_Cross_Attn}_n(\boldsymbol{X}) = \sum_{i=1}^{h} \boldsymbol{W}^i_{n,O}\boldsymbol{V} \\ \cdot S[(\boldsymbol{K})^T \boldsymbol{W}^i_{n,Q}\boldsymbol{X} + (\boldsymbol{K})^T \boldsymbol{W}^i_{n-1,Q}\boldsymbol{X}]\end{aligned} \tag{5}$$

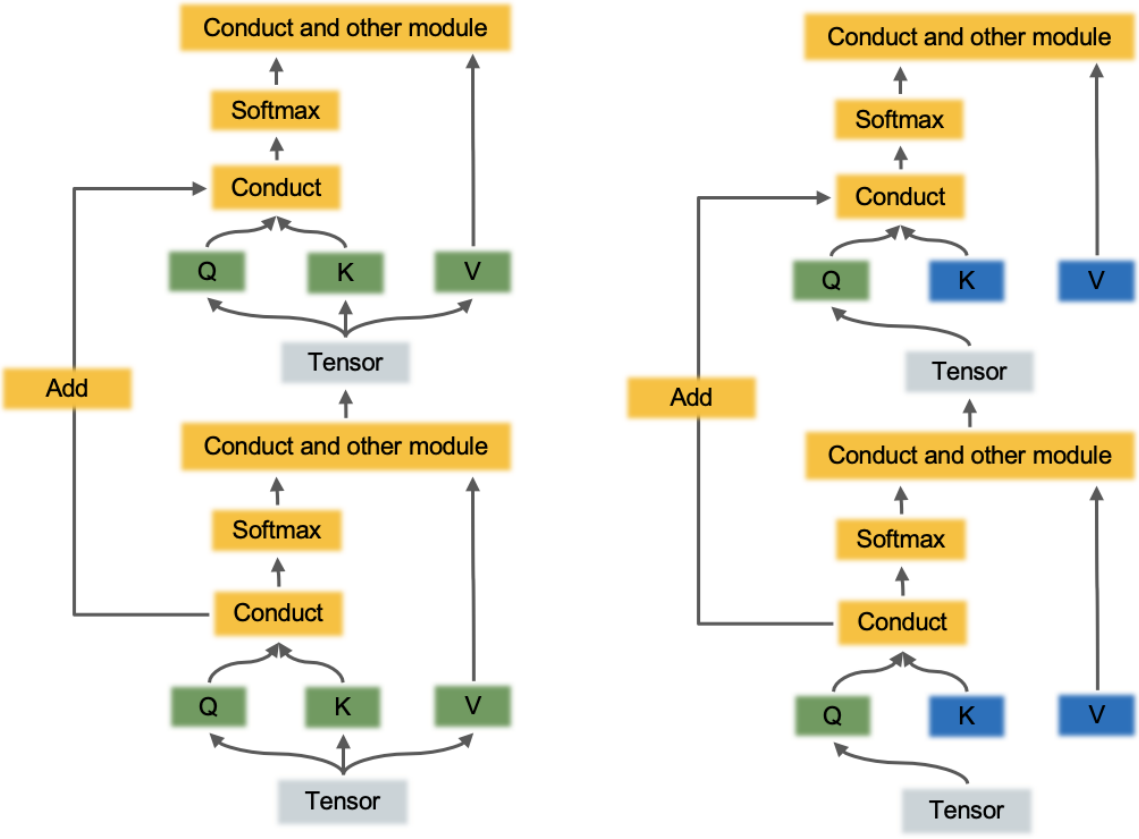
其中，Linked Self Attn$_n$ 表示第 n 个编码器层或第 n 个解码器层的自注意力部分。Linked Cross Attn$_n$ 表示第 n 个解码器层的自注意力部分。其他符号与（1）、（2）和（3）中的相同。

注意力连接的核心思想是，使 Transformer模型中每一层的注意力矩阵不仅取决于当前层的查询张量和键张量，还取决于前一层的查询张量和键张量，从而扩大每一层的感受野，同时缓解资源不足导致的训练偏差对注意力矩阵造成的影响。请注意，注意力链接架构仅将 Transformer模型的编码器层和解码器层中的 Self Attn$_n$ 和 Cross Attn$_n$部分替换为 Linked Self Attn$_n$ 和 Linked Cross Attn$_n$。这不会引入额外参数，而且易于实现。注意力连接的结构如图2所示。

IV. 注意力连接的理论分析

A. 表征能力

我们首先说明，注意力连接不会改变Transformer模型的表征能力。用$T_h(\theta,\cdot)$表示超参数为 h 且可训练参数为 θ的Transformer模型，用 $\tilde{T}_h(\tilde{\theta},\cdot)$表示带有注意力连接、超参数为 h 且可训练参数为 $\tilde{\theta}$的Transformer模型。由于注意力链接结构只是在Transformer模型当前层的 Q,K 乘积中加入前一层的Q,K 乘积，因此只需让 $\tilde{T}_h(\tilde{\theta},\cdot)$和 $T_h(\theta,\cdot)$ 最后一层的注意力矩阵在编码器和解码器中都保持不变。显然，可以通过为 $T_h(\theta,\cdot)$中的编码器或解码器最后一层选择恰当的 Q,K 来实现这一点。总之，我们有



(a) 链接式自注意力　　(b) 链接式交叉注意力

图2。链接式自注意力和链接式交叉注意力

引理1。对于 *Transformer*模型 $T_h(\theta,\cdot)$ 及任意 $\tilde{}$参数 θ，存在参数 θ ，使得链接式 $\tilde{}$ $\tilde{}$*Transformer*模型 $T_h(\theta,\cdot)$ 与 $T_h(\theta,\cdot)$ 相等。也就是说，链接式*Transformer*的表征能力不会低于*Transformer*。

表I 分析中的符号

符号	含义 g
$\boldsymbol{x}$	输入向量
$\boldsymbol{y}$	（真实值）输出向量
$\tilde{\boldsymbol{y}}$	原始Transformer输出向量
$\tilde{\tilde{\boldsymbol{y}}}$	带 AL 的 Transformer 输出向量
N	向量长度
$\boldsymbol{T}$	（简化的）Transformer 操作
P	注意力矩阵
θ	模型参数
θ^*	真实模型参数
$\sigma(i,\ j)$	P中位置 (i,j) 的误差
$\tilde{\Delta}$	$\tilde{\boldsymbol{y}}$的误差
$\tilde{\tilde{\Delta}}$	误差 $\tilde{\tilde{\boldsymbol{y}}}$

B. 鲁棒性

在引理1中，我们说明了带有注意力连接的Transformer与Transformer模型具有相同的表征能力。在本节中，我们尝试从理论上解释注意力连接的优势。

a) 符号说明： 我们首先在表I中给出符号说明

b) 分析设定： 为简化分析，我们做出了一些不具普遍性的简化。首先，我们忽略逐点线性层的影响，只分析Transformer模型中的一 层。此外，我们假设 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 都是长度为 N 的向量。因此，从数学上来说，我们有

$$\boldsymbol{T}_\theta(\boldsymbol{x})(j)=\sum_{i=1}^{N}\boldsymbol{x}(i)P(\theta,\boldsymbol{x},i,j) \tag{6}$$

$$\sum_{i=1}^{N} P(\theta, \boldsymbol{x}, i, j) = 1 \tag{7}$$

$P(\theta,\ \boldsymbol{x},\ i,\ j)$ 是参数为 θ、输入为 $\boldsymbol{x}$的注意力矩阵。请注意，$P(\theta,\ \boldsymbol{x},\ i,\ j)$ 在Transformer模型中充当注意力矩阵。我们将参数的真实值记为 θ^*。因此，我们有，

$$\boldsymbol{y}(j) = \sum_{i=1}^{N} \boldsymbol{x}(i) P(\theta^*, \boldsymbol{x}, i, j) \tag{8}$$

c) 低资源下的原始*Transformer*：由于训练资源有限，我们无法获得精确的 θ^* 以及$P(\theta^*,\ \boldsymbol{x},\ i,\ j)$。我们将低资源在$P(\theta^*,\ \boldsymbol{x},\ i,\ j)$中造成的误差记为 $\sigma(i,\ j)$，并假设 $\sigma(i,\ j) \overset{i.i.d}{\sim} N\left(0,\ \sigma_0^2\right)$。因此，从数学上来说，我们有

$$\tilde{\boldsymbol{y}}(j) = \frac{1}{\tilde{c}_j} \sum_{i=1}^{N} \boldsymbol{x}(i) \left(P\left(\theta^*, \boldsymbol{x}, i, j\right) + \sigma(i, j)\right) \tag{9}$$

σ_0^2 是一个常数，$\tilde{c}_j$ 是归一化系数。由于 P 中的误差相对较小，我们有 $\tilde{c}_j \approx 1$。

$$\begin{aligned} &\tilde{\Delta}(\tilde{\boldsymbol{y}}(j)) = |\boldsymbol{y}(j) - \tilde{\boldsymbol{y}}(j)| \approx \\ &\left|\sum_{i=1}^{N} \boldsymbol{x}(i)\sigma(i, j)\right| \leq \frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{x}^2(i) + \sigma^2(i, j)\right) \end{aligned} \tag{10}$$

$$\begin{aligned} &\frac{1}{N}\sum_{j=1}^{N}\tilde{\Delta}(\tilde{\boldsymbol{y}}(j)) \leq \\ &\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{x}^2(i) + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma^2(i, j) \end{aligned} \tag{11}$$

d) 带注意力连接的*Transformer*：对于带连接的Transformer模型，由于训练资源有限，前一层和当前层的注意力矩阵也存在误差，我们将前一层注意力矩阵的误差记为 σ_{pre}。因此，我们有 $\sigma_{pre}(i,\ j) \overset{i.i.d}{\sim} N\left(0,\ \sigma_0^2\right)$ 以及 $\sigma(i,\ j) \overset{i.i.d}{\sim} N\left(0,\ \sigma_0^2\right)$。由于~存在注意力连接，我们得到输出 $\tilde{\boldsymbol{y}}$:

$$\begin{aligned} \tilde{\tilde{\boldsymbol{y}}}(j) = &\frac{1}{\tilde{\tilde{c}}_j}\sum_{i=1}^{N}\boldsymbol{x}(i)\left(\frac{1}{2}P\left(\theta^*, \boldsymbol{x}, i, j\right) + \right. \\ &\left.\frac{1}{2}P_{pre}\left(\theta^*, \boldsymbol{x}, i, j\right) + \frac{1}{2}\sigma(i, j) + \frac{1}{2}\sigma_{pre}(i, j)\right) \end{aligned} \tag{12}$$

若将 $\gamma(\theta^*,\ \boldsymbol{x},\ i,\ j))$ 表示为

$$\gamma\left(\theta^*, \boldsymbol{x}, i, j)\right) = P_{pre}\left(\theta^*, \boldsymbol{x}, i, j)\right) - P\left(\theta^*, \boldsymbol{x}, i, j\right) \tag{13}$$

则有

$$\begin{aligned} \tilde{\tilde{\boldsymbol{y}}}(j) = &\frac{1}{\tilde{\tilde{c}}_j}\sum_{i=1}^{N}\boldsymbol{x}(i)\left(P\left(\theta^*, \boldsymbol{x}, i, j\right) + \right. \\ &\left.\frac{1}{2}\sigma(i, j) + \frac{1}{2}\sigma_{pre}(i, j) + \frac{1}{2}\gamma\left(\theta^*, \boldsymbol{x}, i, j)\right)\right) \end{aligned} \tag{14}$$

~由于 P 中的误差相对较小，我们也有 $\tilde{c}_j \approx 1$。如果相邻Transformer层之间的真实值 P 差异相对较小，即 $\gamma(\theta^*,\ \boldsymbol{x},\ i,\ j))\ \approx 0$，那么我们有

$$\begin{aligned} &\tilde{\tilde{\Delta}}(\tilde{\tilde{\boldsymbol{y}}}(j)) = |\boldsymbol{y}(j) - \tilde{\boldsymbol{y}}(j)| \approx \\ &\left|\sum_{i=1}^{N} x(i)\left[\frac{1}{2}\sigma(i, j) + \frac{1}{2}\sigma_{pre}(i, j)\right]\right| \\ &\leq \frac{1}{2}\sum_{i=1}^{N}\left[\boldsymbol{x}^2(i) + \left(\frac{1}{2}\sigma(i, j) + \frac{1}{2}\sigma_{pre}(i, j)\right)^2\right] \end{aligned} \tag{15}$$

~对 $\tilde{\boldsymbol{y}}$ 中的元素取平均值，得

$$\begin{aligned} &\frac{1}{N}\sum_{j=1}^{N}\tilde{\tilde{\Delta}}(\tilde{\tilde{\boldsymbol{y}}}(j)) \leq \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{x}^2(i) + \\ &\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{2}\sigma^2(i, j) + \frac{1}{2}\sigma_{pre}(i, j)\right) \end{aligned} \tag{16}$$

e) 注意力连接的优越性：现在，我们分别比较式（11）和式（16）中 $\tilde{\boldsymbol{y}}$ 与 $\tilde{\boldsymbol{y}}$ 的误差~。注意，$\sigma_{pre}\ i\ j \overset{i.i.d}{\sim} N\left(0\ \sigma_0^2\right)$ $\frac{1}{2}\sigma\ i\ j\ \pm\ \frac{1}{2}\sigma_{pre}\ i\ j \overset{i.i.d}{\sim}$（，），以及 $N\left(0\ \frac{1}{2}\sigma_0^2\right)$（，）（，），。根据大数定律，对于式（11）中 $\tilde{\boldsymbol{y}}$ 的误差，有

$$\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}\sigma^2(i, j) \approx E\sum_{i=1}^{N}\sigma^2(i, j) = N\sigma_0^2 \tag{17}$$

~而对于式（16）中 $\tilde{\boldsymbol{y}}$ 的误差，有

$$\begin{aligned} &\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}\left(\frac{1}{2}\sigma^2(i, j) + \frac{1}{2}\sigma_{pre}^2(i, j)\right) \approx \\ &E\sum_{i=1}^{N}\left(\frac{1}{2}\sigma^2(i, j) + \frac{1}{2}\sigma_{pre}^2(i, j)\right) = \frac{1}{2}N\sigma_0^2 \end{aligned} \tag{18}$$

比较式（17）和式（18），我们发现，带有注意力连接的Transformer模型 $\tilde{\tilde{\boldsymbol{y}}}$ 的误差输出减少了 $\frac{1}{2}N\sigma_0^2$，这表明了注意力连接的优越性。

V. 实验

A. Set up

1) 数据集：我们在 IWSLT14 的六项任务（英语-德语、德语-英语、英语-法语、英语-意大利语、意大利语-英语和英语-罗马尼亚语）以及 CVIT PIB 数据集上的孟加拉语-古吉拉特语和古吉拉特语-泰米尔语翻译任务上进行了数值实验测试。

为了验证注意力链接结构在低资源条件下的有效性，我们首先从 IWSLT14 数据集中英语-德语、德语-英语、英语-法语、英语-意大利语、意大利语-英语和英语-罗马尼亚语机器翻译任务的完整双语训练划分语料库中随机抽取约 10K 和 30K 条平行语料库，模拟低资源场景。我们分别在 10K、30K 训练资源级别以及完整训练划分上训练模型，并在 IWSLT14 的英语-德语、德语-英语、英语-法语、英语-意大利语、意大利语-英语和英语-罗马尼亚语翻译任务的完整测试划分上进行评估。

然后，我们在 CVIT PIB 数据集的孟加拉语-古吉拉特语和古吉拉特语-泰米尔语翻译任务上进行实验。我们随机抽取约 5% 的数据

作为 CVIT PIB 数据集中孟加拉语-古吉拉特语和古吉拉特语-泰米尔语翻译任务的测试划分，其余数据作为训练划分。

同时，我们在 IWSLT14 数据集中英语-德语、德语-英语、英语-法语、英语-意大利语、意大利语-英语和英语-罗马尼亚语机器翻译任务的完整数据集上测试注意力链接结构。此外，我们将注意力链接架构与其他先进方法结合，并在 IWSLT14 英语-德语数据集上进行测试。

2) 模型超参数设置：我们主要比较两种结构：Transformer [18] 和带有注意力链接的 Transformer。我们采用标准的 6 层编码器和 6 层解码器结构。需要注意的是，注意力链接不会为模型引入额外参数。为确保公平，实验中 Transformer [18] 和带有注意力链接的 Transformer 模型使用相同的超参数。在实验中，我们设置 $d = 512, d_q = 128, d_k = 128, d_v = 128, d_{hidden} = 1024, h = 4$。我们的数值实验基于 fairseq[1] 代码。Transformer 和带有注意力链接的 Transformer 均采用 adam [39] 方法进行优化，并使用 fairseq$\beta_0 = 0.9, \beta_1 = 0.98, \epsilon = 10^{-8}$, $weightdecay = 10^{-4}$中的默认设置。此外，我们使用预热策略控制学习率的变化，并采用 fairseq$warmupstep = 4000$ 中的默认设置，在 IWSLT14 数据集的英语-德语、德语-英语、英语-法语、英语-意大利语、意大利语-英语和英语-罗马尼亚语机器翻译任务上进行实验，设置为 $lr = 5 \times 10^{-4}$。

B. 结果

在下方的表 II 中，我们展示了英语-德语任务中，在 1万训练资源级别下训练的模型的一些翻译示例。红色文字表示误译，绿色文字表示注意力链接架构相较于 Transformer+ 注意力链接架构所带来的改进。

在下方的表 III、IV 和 V 中，我们展示了模型在不同资源配置下的测试集 BLEU 分数。每项任务中，我们都在单块 P100 GPU 上训练模型，训练时间不超过 3 天。首先，我们发现，与单独使用 Transformer模型相比，采用注意力连接的 Transformer 在不同训练资源设置下都能取得更好的结果。其次，我们发现，通常资源量越少，提升幅度越大。平均而言，在 1万训练资源级别下，注意力链接架构可提升 1.0 BLEU 分数；在 3万训练资源级别下，可提升 0.8 BLEU 分数；在完整数据集上，可提升 0.3 BLEU 分数。第三，我们发现，注意力链接（AL）带来的提升通常会随语言而变化。例如，在 1万训练资源级别下，英语-德语任务的提升为 1.1；而在同一资源级别下，意大利语-英语任务的提升为 0.6。

在表 VI 中，我们展示了 CVIT PIB 数据集上孟加拉语-古吉拉特语和古吉拉特语-泰米尔语翻译任务的训练结果，表明该模型在真实的低资源场景中带来了显著提升。

[1]https://github.com/facebookresearch/fairseq

表 III 1万训练资源级别下 Transformer 与采用注意力链接训练的 Transformer 的 BLEU

Task	Transformer	Transformer+注意力链接
英语-德语	13.3	**14.4**
德语-英语	17.3	**18.0**
英语-法语	22.9	**23.9**
英语-意大利语	15.1	**16.4**
意大利语-英语	19.2	**19.8**
英语-罗马尼亚语	13.5	**14.7**
Avg	16.9	**17.9**

表IV Transformer 与采用注意力链接训练的 Transformer 在 3万训练资源上的 BLEU 水平

Task	Transformer	Transformer+注意力链接
英语-德语	20.7	**21.6**
德语-英语	25.3	**26.1**
英语-法语	31.6	**32.3**
英语-意大利语	23.4	**24.3**
意大利语-英语	26.5	**27.3**
英语-罗马尼亚语	20.9	**21.7**
Avg	24.7	**25.5**

表 V 在完整的 IWSLT14 训练数据集上训练的 Transformer 与采用注意力链接的 Transformer 的 BLEU

Task	Transformer	Transformer+注意力链接
英语-德语	28.6	**28.7**
德语-英语	34.4	**34.7**
英语-法语	40.4	**40.9**
英语-意大利语	30.8	**31.2**
意大利语-英语	34.7	**34.9**
英语-罗马尼亚语	28.3	**29.0**
Avg	32.9	**33.2**

表 VI 在 CVIT PIB 训练数据集上使用 Attention Link 训练的 Transformer 和带 Attention Link 的 Transformer 的 BLEU

Task	Transformer	Transformer+注意力链接
孟加拉语-古吉拉特语	9.1	**10.0**
古吉拉特语-泰米尔语	10.3	**10.4**
Avg	9.8	**10.2**

C. 与其他先进模型结合

在本节中，我们展示了所提出的注意力链接架构可以轻松地与当前先进的基于 Transformer 的神经机器翻译模型相结合，并带来性能提升。我们尝试将注意力链接与其他现有的先进基于 Transformer 的模型相结合：R-Drop [40] 和 Cutoff [41]。在与 Cutoff[41]模型结合后，我们在 IWSLT 德英任务上以 37.9 BLEU 分数达到了新的最先进水平。由于注意力链接不会引入新的参数和超参数，并且为确保公平性，我们在实验中保持结合后的超参数设置不变。实验结果如表 VII 所示。

表 II 使用英德任务的 1万训练资源级别训练的模型翻译示例。

源句	这很难。
Transformer 结果	这是困难的。
Transformer+注意力链接 结果	这很困难。
目标句（参考答案）	这很困难。
源句	令人无比兴奋。
Transformer 结果	巨大。
Transformer+注意力链接结果	非常令人兴奋。
目标句（参考答案）	极其令人兴奋。
源句	告诉我关于这个世界的事。
Transformer 结果	我会¨告诉您关于¨它的事。
Transformer+注意力链接结果	告¨诉我 ¨关于这个世界。
目标句（参考答案）	给我讲讲这个世界。
源句	now，那是什么意思?
Transformer 结果	和那是什么意思?
Transformer+注意力链接 结果	nun，这是什么意思?
目标句（参考答案）	那么，这是什么意思?

[a]我们展示了在 ofen-de 任务的 1万训练资源级别上训练的模型的一些翻译示例。红色的词表示误译，绿色的词表示在 Transformer + 注意力链接架构中的改进。

表 VII：注意力连接与其他基于 Transformer 的方法相结合的蓝色

模型	BLEU
R-drop [40]	36.9±0.5
截断 [41]	37.7±0.5
R-drop [40]+注意力链接	37.3±0.5
截断 [41]+注意力链接	**37.9**±0.5

表 VIII 原始Transformer（不含注意力链接）的 BLEU、仅编码器包含注意力链接（编码器注意力链接）、仅解码器包含注意力链接（解码器注意力链接）、编码器和解码器均包含注意力链接（注意力链接）的 BLEU

Task	不含注意力链接	编码器注意力链接	注意力链接 AL	AL
英语-德语	13.3	13.8	14.2	**14.4**
德语-英语	17.3	**18.1**	17.7	18.0
英语-法语	22.9	**24.3**	23.5	23.9
英语-意大利语	15.1	15.7	15.5	**16.4**
意大利语-英语	19.2	19.7	19.2	**19.8**
英语-罗马尼亚语	13.5	14.1	14.3	**14.7**
Avg	16.9	17.6	17.4	**17.9**

VI. 消融研究

为进一步说明所提出的注意力连接的有效性，我们进行了消融实验，即仅将注意力连接添加到 Transformer 的编码器模块中，以及仅将注意力连接添加到 Transformer 的解码器模块中。我们在 IWSLT 的六个任务的 1万训练资源级别上进行了消融实验：英语-德语、德语-英语、英语-法语、英语-意大利语、意大利语-英语和英语-罗马尼亚语。

实验结果如表 VIII 所示。我们可以看到，仅在 Transformer 的编码器模块上添加注意力连接或仅在 Transformer 的解码器模块上添加注意力连接，通常都能提升模型性能，相较于原始 Transformer 模型而言。不过，这种提升通常不如在 Transformer 的编码器和解码器两端同时添加注意力连接明显。这进一步说明了注意力连接的有效性。此外，我们还可以看到，在编码器模块上添加注意力连接通常更有效。

VII. 结论

本文受快捷连接设计的启发，设计了一种名为注意力连接的新型高效架构。我们从理论上解释了注意力连接的优越性。此外，实验表明，采用注意力连接的 Transformer 模型性能大幅提升；通过结合注意力连接与其他先进方法，我们在 IWSLT14 德语-英语翻译任务上取得了新的最先进水平，BLEU 分数达到 37.9。

我们认为，注意力连接能够轻松捕获句子中更通用的语义信息，并且对噪声具有鲁棒性。此外，我们的架构可以轻松部署于基于 Transformer 的模型中，而无需增加额外参数。从某种意义上说，这是一种免费的午餐。我们希望本研究能为以低资源获取高效的神经机器翻译模型提供新思路。

关于局限性，本文主要研究低资源神经机器翻译模型。我们提出的注意力连接基于Transformer模型，因此也可能继承Transformer的缺点，例如难以选择合适的优化参数。

参考文献

[1] R. Rajan、R. Sivan、R. Ravindran 和 K. Soman，“从英语到马拉雅拉姆语的基于规则的机器翻译”，载于《2009年计算、控制与电信技术进展国际会议》。IEEE，2009年，第439–441页。[2] A. Hurskainen、J. Tiedemann 等，“从英语到芬兰语的基于规则的机器翻译”，载于《第二届机器翻译会议论文集（WMT2017）》。计算语言学协会，2017年。[3] M. A. Sghaier 和 M. Zrigui，“从突尼斯方言到现代标准阿拉伯语的基于规则的机器翻译”，Procedia Computer Science，第176卷，第310–319页，2020年。[4] Y. Shiwen 和 B. Xiaojing，“基于规则的机器翻译”，载于《劳特里奇翻译技术百科全书》。劳特里奇，2014年，第224–238页。[5] M. L. Forcada、M. Ginest´ı-Rosell、J. Nordfalk、J. O’Regan、S. Ortiz-Rojas、J. A. P´erez-Ortiz、F. S´anchez-Mart´ınez、G. Ram´ırez-S´anchez 和 F. M. Tyers，“Apertium：用于基于规则的机器翻译的自由/开源平台”，《机器翻译》，第127–144页，2011年。[6] M. Artetxe、G. Labaka 和 E. Agirre，“无监督统计机器翻译”，arXiv预印本 arXiv:1809.01272，2018年。[7] F. J. Och，“统计机器翻译中的最小错误率训练”，载于《计算语言学协会第41届年会论文集》，2003年，第160–167页。[8] F. J. Och、C. Tillmann 和 H. Ney，“改进统计机器翻译的对齐模型”，载于《1999年联合SIGDAT自然语言处理实证方法与超大规模语料库会议》，1999年。[9] R. Zens 、 F. J. Och 和 H. Ney，“ 基于短语的统计机器翻译 ”，载于《KI 2002：人工智能进展：德国人工智能第25届年会，KI 2002，德国亚琛，2002年9月16–20日，论文集25》。施普林格，2002年，第18–32页。[10] N. Kalchbrenner 和 P. Blunsom，“循环连续翻译模型”，载于《2013年自然语言处理实证方法会议论文集》，2013年，第1700–1709页。[11] I. Sutskever、O. Vinyals 和 Q. V. Le，“利用神经网络进行序列到序列学习”，《神经信息处理系统进展》，第27卷，2014年。[12] K. Cho、B. Van Merri¨enboer、C. Gulcehre、D. Bahdanau、F. Bougares、H. Schwenk 和 Y. Bengio，“使用RNN编码器-解码器学习用于统计机器翻译的短语表示”，arXiv预印本 arXiv:1406.1078，2014年。[13] D. Datta 、 P. E. David 、 D. Mittal 和 A. Jain，“ 使用循环神经网络进行神经机器翻译”，《国际工程与先进技术期刊》，第9卷，第4期，第1395–1400页，2020年。[14] A. Hermanto、T. B. Adji 和 N. A. Setiawan，“用于英印尼机器翻译的循环神经网络语言模型：实验研究”，载于《2015年信息技术科学国际会议（ICSITech）》。IEEE，2015年，第132–136页。[15] L. Jian、H. Xiang 和 G. Le，“用于英语机器翻译的基于LSTM的注意力嵌入”，《科学编程》，第2022卷，2022年。[16] H. Xu、Q. Liu、J. van Genabith、D. Xiong 和 M. Zhang，“用于神经机器翻译的多头高度并行化LSTM解码器”，载于《计算语言学协会第59届年会与第11届自然语言处理国际联合会议论文集（第1卷：长篇论文）》，2021年，第273–282页。[17] D. Bahdanau、K. Cho 和 Y. Bengio，“通过联合学习对齐与翻译实现神经机器翻译”，arXiv预印本 arXiv:1409 0473，2014年。。[18] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力就是一切”，《神经信息处理系统进展》，第30卷，2017年。[19] I. Beltagy、M. E. Peters 和 A. Cohan，“Longformer：长文档 Transformer”，arXiv预印本 arXiv:2004.05150，2020年。[20] C. Zhu、W. Ping、C. Xiao、M. Shoeybi、T. Goldstein、A. Anandkumar 和 B. Catanzaro，“长短Transformer：适用于语言与视觉的高效Transformer”，《神经信息处理系统进展》，第34卷，2021年。[21] H. Zhang、Y. Gong、Y. Shen、W. Li、J. Lv、N. Duan 和 W. Chen，“Poolingformer：利用池化注意力进行长文档建模”，载于《国际机器学习会议》。PMLR，2021年，第12 437–12 446页。[22] J. W. Rae, A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap, “用于长程序列建模的压缩 Transformer”，arXiv 预印本 arXiv:1911.05507，2019。[23] N. Kitaev, Ł. Kaiser, and A. Levskaya, “Reformer：高效的 Transformer”，arXiv 预印本 arXiv:2001.04451，2020。[24] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer：具有线性复杂度的自注意力”，arXiv 预印本 arXiv:2006 04768，2020。[25] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, “学习用于机器翻译的深层 Transformer 模型”，arXiv 预印本 arXiv:1906.01787，2019。[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “向量空间中词表示的高效估计”，arXiv 预印本 arXiv:1301 3781，2013。[27] R. Sennrich, B. Haddow, and A. Birch, “利用单语数据改进神经机器翻译模型”，arXiv 预印本 arXiv:1511 06709，2015。[28] S. Edunov, M. Ott, M. Auli, and D. Grangier, “大规模理解回译”，arXiv 预印本 arXiv:1808.09381，2018。[29] H. Li, J. Sha, and C. Shi, “重新审视中文与越南语低资源机器翻译中的回译”，IEEE Access，第 8 卷，第 119 931–119 939 页，2020。[30] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, “预训练词嵌入何时以及为何有助于神经机器翻译？”，arXiv 预印本 arXiv:1804.06323，2018。[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT：用于语言理解的深层双向 Transformer 预训练”，arXiv 预印本 arXiv:1810.04805，2018。[32] V. C. D. Hoang, P. Koehn, G. Haffari, and T. Cohn, “神经机器翻译的迭代回译”，载于《第二届神经机器翻译与生成研讨会论文集》，2018 年，第 18–24 页。[33] A. Araabi and C. Monz, “为低资源神经机器翻译优化 Transformer”，arXiv 预印本 arXiv:2011.02266，2020。[34] E. Van Biljon, A. Pretorius, and J. Kreutzer, “关于低资源语言翻译的最优 Transformer 深度”，arXiv 预印本 arXiv:2004.04418，2020。[35] W. You, S. Sun, and M. Iyyer, “用于神经机器翻译的硬编码高斯注意力”，arXiv 预印本 arXiv:2005.00742，2020。[36] A. Raganato, Y. Scherrer, and J. Tiedemann, “基于 Transformer 的机器翻译中的固定编码器自注意力模式”，arXiv 预印本 arXiv:2002.10260，2020。[37] K. He, X. Zhang, S. Ren, and J. Sun, “用于图像识别的深度残差学习”，载于《IEEE 计算机视觉与模式识别会议论文集》，2016 年，第 770–778 页。[38] C. Yun, S. Bhojanapalli, A. S. Rawat, S. J. Reddi, and S. Kumar, “Transformer 是序列到序列函数的通用逼近器吗？”，arXiv 预印本 arXiv:1912.10077，2019。[39] D. P. Kingma and J. Ba, “Adam：一种随机优化方法”，arXiv 预印本 arXiv:1412.6980，2014。[40] L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu 等，“R-drop：用于神经网络的正则化随机失活”，《神经信息处理系统进展》，第 34 卷，2021。[41] D. Shen, M. Zheng, Y. Shen, Y. Qu, and W. Chen, “一种简单却难以超越的自然语言理解与生成数据增强方法”，arXiv 预印本 arXiv:2009.13818，2020。