第34卷 第7期

2020年7月

中文信息学报

JOURNAL OFCHINESEINFORMATION PROCESSING

Vol.34,No.7

July,2020

文章编号:1003-0077(2020)07-0001-18

神经机器翻译前沿综述

冯 洋1,2,邵晨泽1,2

(1.中国科学院 计算技术研究所 智能信息处理重点实验室,北京100190; 2.中国科学院大学,北京 100049)

摘 要:机器翻译是指通过计算机将源语言句子翻译到与之语义等价的目标语言句子的过程,是自然语言处理领域的一个重要研究方向。神经机器翻译仅需使用神经网络就能实现从源语言到目标语言的端到端翻译,目前已成为机器翻译研究的主流方向。该文选取了近期神经机器翻译的几个主要研究领域,包括同声传译、多模态机器翻译、非自回归模型、篇章翻译、领域自适应、多语言翻译和模型训练,并对这些领域的前沿研究进展做简要介绍。

关键词:神经机器翻译;模型训练;同声传译;多模态机器翻译;非自回归机器翻译;篇章翻译;领域自适应;多语言翻译

中图分类号:TP391 文献标识码:A

# FrontiersinNeuralMachineTranslation:ALiteratureReview

FENG Yang1,2 ,SHAO Chenze1,2 (1.KeyLaboratoryofIntelligentInformationProcessing,InstituteofComputing Technology,ChineseAcademyofSciences,Beijing100190,China; 2.UniversityofChineseAcademyofSciences,Beijing100049,China)

Abstract:Machinetranslationisatask whichtranslatesasourcelanguageintoatargetlanguageoftheequivalent meaningviaacomputer,whichhasbecomeanimportantresearchdirectioninthefieldofnaturallanguageprocess- ing.Neuralmachinetranslation models,asthe mainstreaminthereasearchcommunity,canperform end-to-end translationfromsourcelanguagetotargetlanguage.Inthispaper,weselectseveralmainresearchdirectionsofneu- ralmachinetranslation,including modeltraining,simultaneoustranslation,multi-modaltranslation,non-autore- gressivetranslation,document-leveltranslation,domainadaptation,multilingualtranslation,andbrieflyintroduce theresearchprogressesinthesedirections.

Keywords:neuralmachinetranslation;modeltraining;simultaneoustranslation;multi-modaltranslation;non-au- toregressivetranslation;document-leveltranslation;domainadaptation;multilingualtranslation

基于规则的方法是机器翻译研究的主流,这种方法对

# 神经机器翻译

机器翻译是指通过计算机将源语言句子翻译到与之语义等价的目标语言句子的过程,是自然语言处理领域的一个重要研究方向。1949 年,Warren Weaver在《翻译》中提出了使用机器进行翻译的思想,自此引发了该方向的研究热潮。机器翻译主要 可以分为三种方法:基于规则的机器翻译、基于统 计的机器翻译和基于神经网络的机器翻译。最初,

收稿日期:2020-01-06 定稿日期:2020-01-22

语法结构规范的句子有较好的翻译效果,但其也有规则编写复杂、难以处理非规范语言现象的缺点。20世纪90年代初,IBM 的PeterBrown等人发表了两篇重要论文[1-2],正式提出基于噪声信道模型的统计机器翻译模型。进入21世纪,深度学习等机器学习方法逐渐成熟,并开始被应用于自然语言处理领域。2013年,Kalchbrenner和Blunsom 提出利用神经网络进行 机 器 翻 译[3],随 后 一 两 年 内,Sutskever[4]、 Cho[5-6]、Bahdanau[7]等人提出了基于编码器—解码器

基金项目:国家重点研发计划政府间国际科技创新合作重点专项(2017YFE0192900)

结构的神经机器翻译模型,标志着机器翻译进入深度学习的时代。2016 年,JunczysDowmunt等人[8]在30多个语言对上对神经机器翻译和统计机器翻译进行对比,神经机器翻译在27个任务上超过了基于短语的统计机器翻译,这展现了神经机器翻译的强大能力。

尽管神经机器翻译已经表现出比统计机器翻译更加优异的翻译效果,但其仍具有巨 大的发展潜力。2016 年,Wu 等人[9]公 布了谷歌的神经机器翻译模型,该模型通过在层之间引入 残差连接解决了深度模型梯度消失的问题,将模 型层数堆叠到了8 层,使机器翻译的水平提升到 了一个新的台阶。随后,facebook的 Gehring等人[10]提出了基于卷积神经网络的编码器—解码器模型,在准确度上超越了谷歌的模型,并大幅提升 了翻译速度。2017 年,Vaswani等人[11]提出 了基于注意力机制的 Transformer模型,在模型的训练速度和翻译质量上都取得了大幅提升。2018 年,Hassan等人[12]将多种算法结合,并在翻译评价中引入人工评测,首次宣布模型在新闻领域的翻译 上达到了人类水平。2019 年以来,又有许多神经机器翻译的模型结构被人们提出[13-19],它 们在实验中展现出了高于基准 Transformer模 型的翻译质量。除此之外,反向翻译[20-21]、数据筛选[22-23]、预训练[24]等技术也对翻译效果的提升有显著作用。

尽管神经机器翻译在标准数据集上达到了相当高 的翻译质量,但在实际应用中,仍有许多问题需要解决。神经机器翻译模型存在训练和测试时行为不一致的问题,该问题被称为“曝光偏差”,引发了研究者的广泛关注;在同声传译的场景下,为了降低翻译的延迟, 模型需要在输入语句不完整的情况下输出译文,使用户能低延迟地收到高质量的翻译结果;除文本外,有时也存在图像、视频等其他模态的数据可供翻译模型使用,翻译系统可以融入这些信息以进一步地提高翻译质量;为了提升翻译速度,非自回归模型对翻译概率独立建模,因此能够并行解码出整句译文,但也会出现严重的漏译、过译现象;在对篇章文本进行翻译时,为了保证译文的一致性,模型在翻译时也需同时考虑上下

题,对神经机器翻译中同声传译、多模态机器翻译、非自回归模型、篇章翻译、领域自适应、多语言翻译和模型训练等方向的前沿研究进展做简要介绍。

# 主流框架

* 1. 基于循环神经网络的神经机器翻译模型

基于循环神经网络和编码器—解码器结构的神经机器翻译模型[4-7]在很长一段时间内都是神经机器翻译的主流模型。其中,Bahdanau等人[7]在编码器—解码器框架的基础上,提出了 RNNSearch 模型,该模型引入了注意力机制,使得生成每个目标端词语时,解码器可以将“注意力”集中到源端的几个相关的词语上,并从中获取有用的信息,从而获得更好的翻译表现。注意力机制使得翻译模型能够更好地处理长距离的依赖关系,解决了在循环神经网络中信息在长距离的传输中容易被丢失、遗忘的问题。RNNSearch模型被研究者广泛地用作基线模型,其注意力机制如图1所示,下面我们对其做具体介绍。

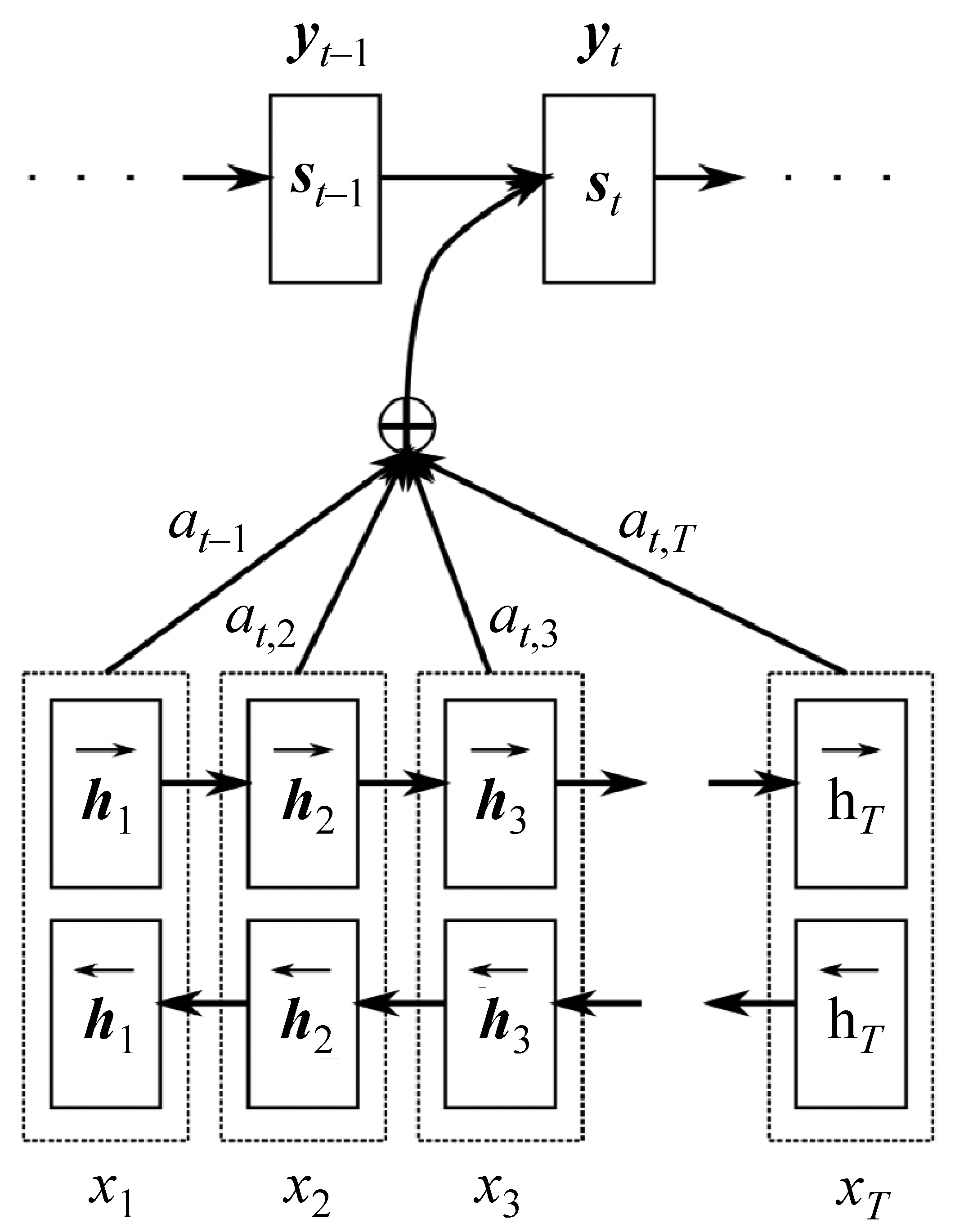


图1 循环神经网络中的注意力机制

(图片引自文献[7])

编码器 使用双向门控循环单元(GRU)[5]对源语句进行编码,以使每个位置的编码同时包含前、后文本的历史信息。双向 GRU 由前向 GRU 和后

向 GRU 组成,前向 GRU 从左向右读取源语句并计

文的信息;存在低资源的领域内数据和高资源的领域

外数据时,为了提升模型在领域内的翻译质量,需要对

算一系列前向隐状态(h⇀ ,…,h⇀n ),反向 GRU 从右

; 向左扫描源语句,计算一系列后向隐状态(h↼

1

,…,

领域外数据也进行合理利用 当需要在多个语言之间进行翻译时,训练多语言翻译模型可以大幅减少所需

1

h↼n )。假设输入序列的词嵌入为(x1,…,xn ),则编

的翻译模型数目,同时提升低资源语言的翻译质量。在接下来的内容中,我们将首先介绍神经机器

翻译主流框架的演变,随后针对本文上面提到的问

码器如式(1)所示。

h⇀i =GR⇀U(xi,h⇀i-1), h↼i =GR↼U(xi,h↼i+1),

hi =[h⇀i,h↼i]。 (1)

解码器 解码器是一个前向 GRU,逐词预测译文y。生成译文第j 个词yj 的概率为:

在卷积模块的编码后,解码器的第l 层对编码器的第u 层做注意力,首先对解码器的隐变量hl 做变

P(yj|y<j ,x,θ)=softmax(tj-1,cj,sj) (2)

其中,tj-1是词yj-1的词嵌入,sj 是解码器在第

i

换,如式(7)所示。

dl =Wlhl +bl +gi 7

()

i

d i

d

j 步时的隐状态,cj 为第j 步的注意力向量。状态 其中,gi 为目标端第i 个词的嵌入,W 、b 为线

sj 计算如式(3)所示。( ,

, )。

性变换的参数。求出对源端第u 层的注意力,如式

sj =GRUtj-1 sj-1 cj

(3)

* + 1. 所示。

(l u )

注意力机制被用于提取与当前步预测高度相关

aj = expdi ×zj

, (8)

的源端信息,防止源端信息在长距离的解码中被遗

i m

exp(dl

zu )

忘一部分。在第j 步的解码中,与位置i 的源端信

其中,zu 为编码器第u 层第j 步的输出。用式

j∑=1

i × j

息的相关度如式(4)所示。 () j ,

eij =vaTtanh(Wasj-1 +Uahi) (4)

注意力向量为源端信息按相关度的加权和如

式(5)所示。

8 所得的注意力权重更新解码器的隐状态 如式

* + 1. 所示。

m

cl = ∑al (zu +e ) (9)

exp(e ) T

i ij j j j=1

αij = n , cj = ∑αijhi (5)

ij

在更新隐状态时,用 到的不仅有编码器输出

∑exp(ei'j) i=1

zu ,还有源端词嵌入e 。所得结果cl 将作为下一个

i'=1

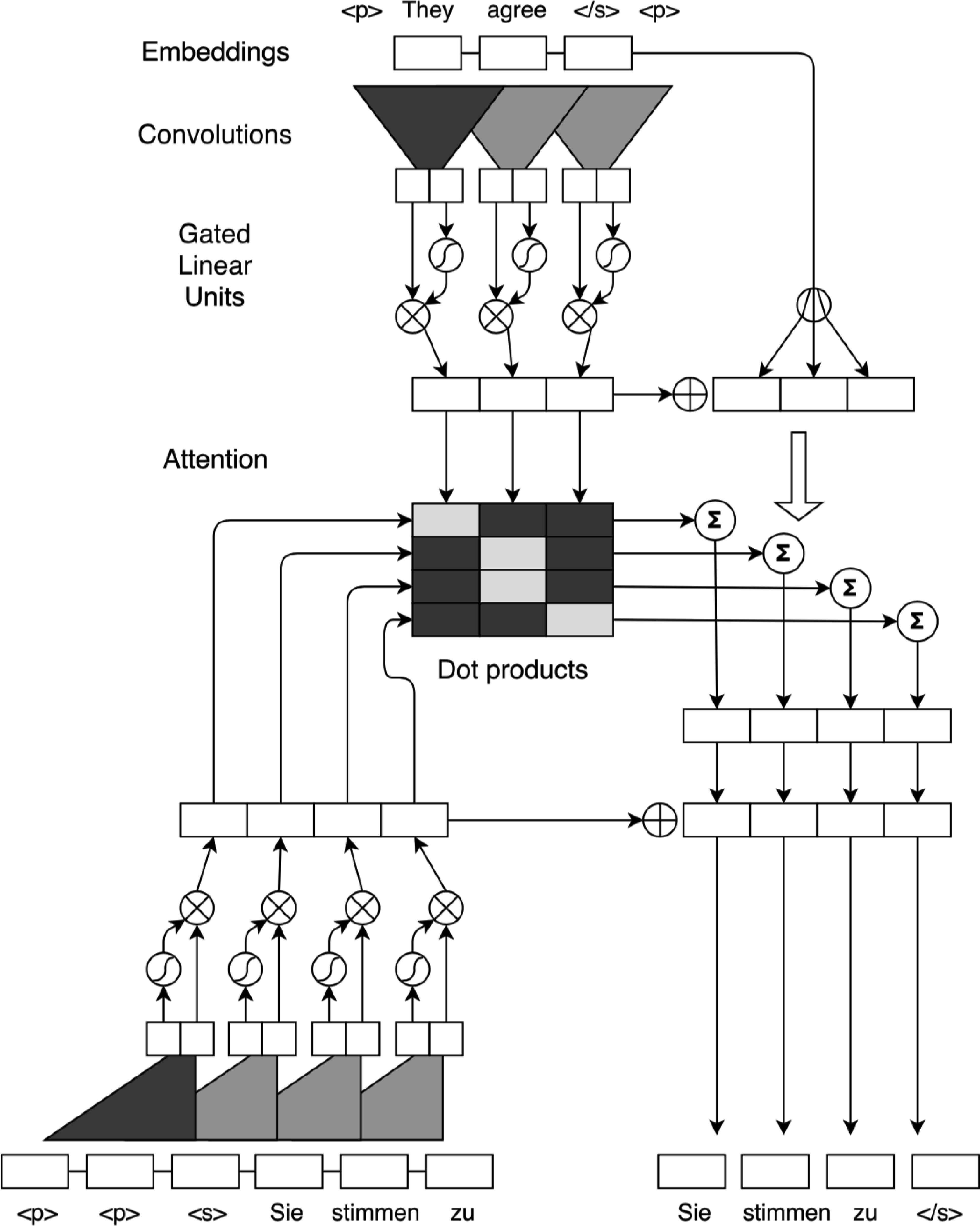
j j i

。

* 1. 基于卷积神经网络的神经翻译模型

在基于循环神经网络的神经机器翻译模型展现出强大的能力后,研究者们也开始探 索基于卷积神经网络的神经翻译模型。Meng等人[25]用卷积神经网络来编码源端语言,并将其集 成到统计机器翻译模型中。Gehring 等人[26]将 神经网络翻译模型的源端编码器替换成了基于卷积神经网络的结构,随 后 Gehring 等人[10]提 出了完全基于卷积神经网络的机器翻译模型。下面,我 们对文献[10]中的模型做简要介绍。

编码器 卷积神经网络在处理输入序列中的一个片段时,并不知道这个片段在句子中的具体位置。因此,在词嵌入中加入位置编码可以使模型获得更丰富的信息,增强模型的表示能力。假设w 为词嵌入,p为位置嵌入,则模型输入为w 与p 之和如式(6)所示。

w =(w1,…,wm ), p=(p1,…,pm ),

卷积模块的输入 图2展示了基于卷积神经网络的编码器-解码器结构及其注意力机制。

e=(w1 +p1,…,wm +pm )

模型的编码器就是多个卷积模块的叠加

(6)

通过

,

卷积操作对输入序列进行编码。令输入窗口大小为k,模型维数为d,则卷积模块的输入 X ∈Rkd ,卷积核的大小为W ∈R2d×kd ,将输入编码为2d 长度的向量,并通过 GLU 非线性变换[27]将其变换为d 维向量。随后,卷 积模块的输入通过残差连接[28] 与

图2 卷积神经网络中的注意力机制

(图片引自文献[10])

* 1. Transformer模型

GLU 输出相连,得到卷积模块的输出。

解码器 解码器与编码器结构基本相同

,也是

2017年,Vaswani等人提出了完全基于注意力

由多个卷积模块叠加而成,这里只列出其不同点。 机制的 Transformer模型[11],该模型创新性地使用

了自注意力机制来对序列进行编码,其编码器和解码器均由注意力模块和前向神经网络构成。Trans- former模型具有高度并行化的模型结构,因此在训练速度上远超循环神经网络,且在翻译质量上也有大幅提升。近期,Transformer已成为神经机器翻译研究中的主流模型,且在自然语言处理的其他领域中也有广泛应用。Transformer的模型结构如图 3所示,下面我们对其做具体介绍。

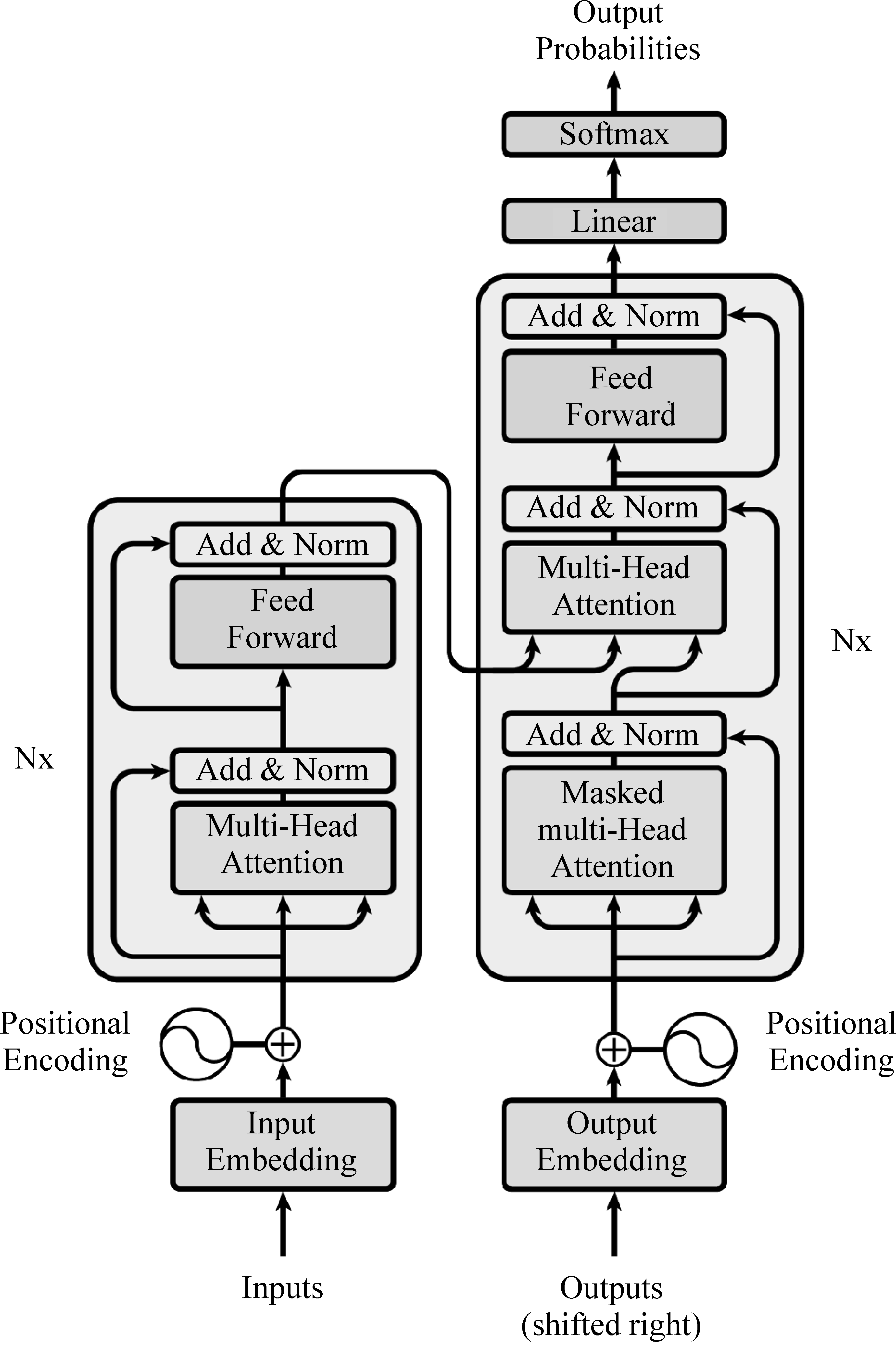


图3 Transformer模型

(图片引自文献[11])

编码器 由于 Transformer模型中没有对序列顺序的显式利用,因此模型需要对词进行位置编码, 以表示序列中不同词的位置关系,如式(10)所示。

PE(pos,2i)=sin(pos/100002i/dmodel), PE(pos,2i+1)=cos(pos/100002i/dmodel)(10)其中,pos表示词在句子中的位置,i 表示维数。

将位置编码和原本的词嵌入相加后,输入到模型的

编码器中。编码器由n 个结构相同的层组成,每层

其中,dk 为键K 的维度。在编码器中,使用的是自注意力模块,即将模型输入同时作为Q、K、V,让输入序列对自身进行注意力计算。除此之外,作者还提出了多头的机制,即在计算注意力时,把输入平均分成多个部分,每个部分独立地计算注意力,最后把得到的注意力结果做拼接,作为最终结果。在计算完自注意力后,用如下的前馈神经网络对输入做变换,如式(12)所示。

FFN(x)=max(0,xW1 +b1)W2 +b2 (12)

除上述自注意力模块和前馈神经网络模块,残差连接[28]、层归一化机制[29]也被应用到了 Trans- former模型中。

解码器 Transformer的解码器与编码器结构基本相同,这里只列出其不同点。在解码时,译文是单向生成的,前面的解码步骤不应看到后面步骤的翻译结果。因此,在训练时,解码器到自注意力模块里会引入一个单向的 mask 矩阵,使从前往后的注意力结果被固定为0。另外,在解码器的自注意力模块与前馈神经网络模块之间,还有一个编码器-解码器注意力模块,将解码器的“注意力”集中到源端相关的词语上。其中,解码器的输入为式(11)中Q,编码器的输出同时为K、V。

* 1. 性能比较

对于这三种基本模型,在 WMT2014 英语到德语的翻译任务上,按 照文献 [9-11]的 配置,采 用 BLEU-4作为评价指标,性能如表1所示。

表1 主流框架性能比较

|  |  |
| --- | --- |
| 模型 | En->De |
| RNNSearch[9] | 25.82 |
| ConS2S[10] | 25.16 |
| Transformer(base)[11] | 27.3 |
| Transformer(big)[11] | 28.4 |

# 同声传译

在同声传译的场景下,为了降低翻译的延迟,模

主要包含两个模块:注意力和前馈神经网络。注意力模块以点乘注意力为基础,对输入的请求Q、键K 和值V 做如下操作,如式(11)所示。

dk





型需要在输入语句不完整的情况下输出译文,使用户能在低延迟内收到高质量的翻译结果。翻译质量和延迟一般是互斥的,在翻译前等待的时间越长,得

( , , )

QKT  ( )

到的源端信息越完整,翻译质量通常就越高,但延迟

AttentionQ K V

=softmax

V 11

也就越高。因此,模型需要在翻译质量和时间延迟

上进行权衡,找出质量和延迟之间的平衡点。在翻译质量上,一 般还是用机器翻译中常用的 BLEU 值[30]作为评价指标。在延迟的评价上,研究者提出了 average proportion(AP)[31]、consecutive wait (CW )[32]、averagelagging(AL)[33]、differentiable averagelagging(DAL)[34]等指标,目前还没有达成一致。在评价同传翻译模型的效果时,通常会结合翻译质量和时间延迟两者,画出延迟-BLEU 值的曲线图,认为曲线整体在上方的模型效果更好。

在同声传译中,一个主要的难点就是让模型决定是否在当前位置输出翻译。如果翻译过早,系统 读取的输入不够,会导致翻译质量较低。如果翻译 过晚,虽然能获得足够的信息,但是会造成翻译的延迟较高。因此,读写策略的制定也成了研究者重点关注的内容。Cho和 Esipova[31]首次提出了基于神经机器翻译的同声传译,并提出了一种基于模型翻译概率的变化来制定读写策略的方法。Gu等人[32]则是通过将翻译质量 BLEU 和延迟指标 AP、CW 设定为奖赏值,运用强化学习算法来让模型自动地学习读写策略。相比于文献[31]的静态策略,文献 [32]提出的方法能通过调节翻译质量和延迟奖赏的比值来调节翻译延迟,因此可以适应不同场景下的同传需求。上面两种方法共同的缺点是:它们均是 直接使用整句的翻译模型来翻译不完整的输入句, 导致训练和测试的模型行为不匹配。文献[33,35]提出了基于固定延迟的读写策略,使得译文总是落后原文固定数目个单词。这种固定的读写策略可以在给定的延迟下训练和测试模型,但也使得模型无

法针对特定输入适当地加快或放慢速度。Zheng等人[36-37]提出了两种对读写策略的监督学习方法,文献[36]中提出的方法使得模型能够尽量输出延迟范围内的动作序列,文献[37]从整句翻译模型的概率变化中分析出动作序列,以此为监督来训练读写策略模块。Arivazhagan等人[34]使用单调的硬注意力机制来得到读写策略,并在训练中使用软注意力来进行模拟,使得模型的翻译部分和读写策略能同时进行训练。为了使得延迟可控,文献[34]引入了可导的延迟指标 DAL,令模型在训练时也对延迟进行优化。Ma等人[38]随后将该单调注意力方法扩展到了 Transformer模型的多头注意力机制中。

除读写策略外,Alinejad等人[39]在文献[32]的基础上加入了Predict操作,使用语言模型来预测当前输入的下一个单词,使模型能获得更完整的输入。束搜索解码算法通常难以应用在同传场景中, Zheng等人[40]对束搜索算法进行改进,使之能在有限步内进行推测,得到更准确的翻译结果。

# 多模态机器翻译

除文本外,有时也存在图像、视频等其他模态的信息可供使用,多模态翻译系统同时将源端文本和其他模态的信息作为模型输入,翻译系统在其他模态信息的辅助下进行翻译,如图4 所示。WMT 在2016到2018年连续三次将多模态机器翻译作为共享任务[41-43],这 项 任 务 也 受 到 了 研 究 者 的 广 泛关注[44-45]。

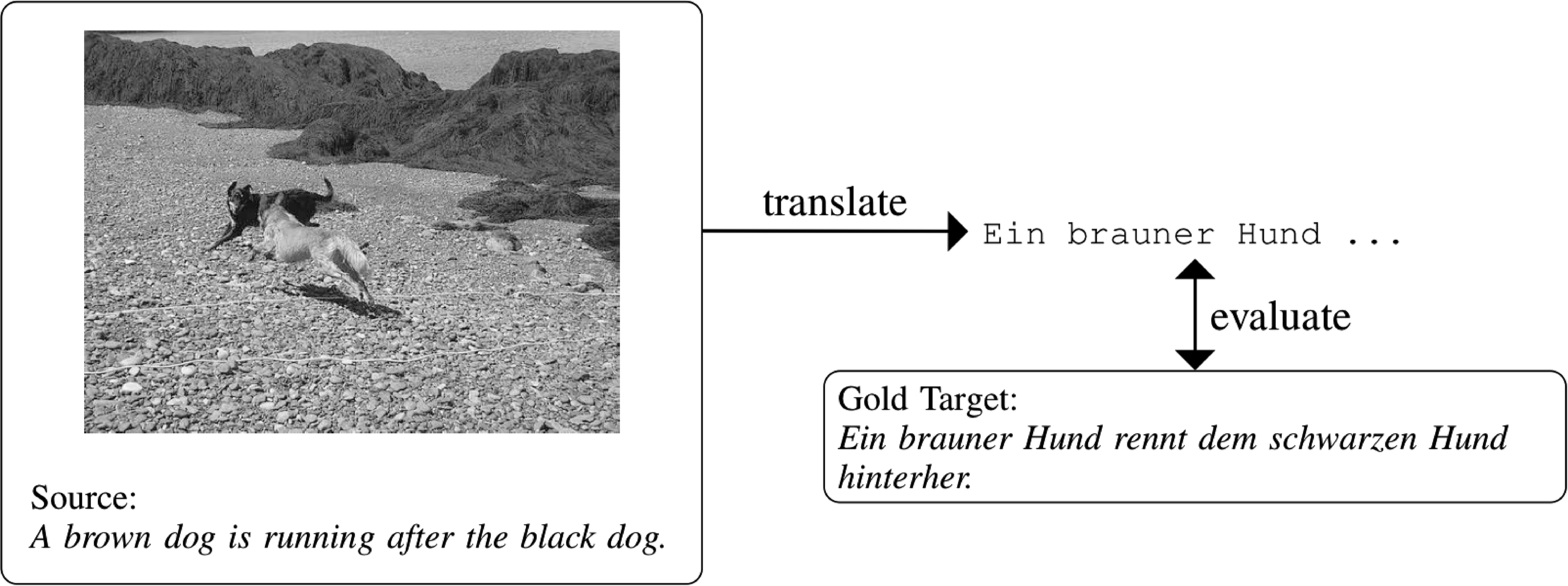


图4 多模态翻译任务,以图像和源端文本为输入,输出目标端译文

(图片引自文献[41])

在多模态机器翻译中,一般是以图像为额外的模态信息,辅助翻译系统进行翻译。在对图像进行表示时,通常是用从大型视觉数据集上训练的卷积神经网络(CNN)来提取图像的多种深层特征。为了在翻译模型中应用图像信息,早期的工作通常将图像信息作为输入语句的一部分[46-47],或用其对编码器、解码器

的状态做初始化[46,48-49]。随后,人们开始尝试通过注意力机制来挖掘视觉信息的作用。Caglayan 等人[50]提出了多模态的注意力机制,让模型在解码时同时对文本表示和图像表示做注意力,更充分地对图像信息进行利用。Calixto等人[51]提出了一种类似的双注意力解码器,通过对文本和图像分别去求注意力,并通