文章编号：1003-0077（2017）00-0000-00

融合图像注意力的多模机器翻译模型

李霞[1,2] 马骏腾[1] 覃世豪[1]

（1.广东外语外贸大学信息科学与技术学院，广东省 广州市 510006；
2.广州市非通用语种智能处理重点实验室，广东省 广州市 510006）

摘要： 多模机器翻译近年来成为研究热点之一。已有工作表明，融入图像视觉语义信息可以提升文本机器翻译模型的效果，已有工作多数将图片的整体视觉语义信息融入到翻译模型，而图片中可能包含不同的语义对象，并且这些不同的局部语义对象对解码端单词的预测具有不同程度的影响和作用。基于此，本文提出一种融合图像注意力的多模机器翻译模型，将图片中的全局语义和不同部分的局部语义信息与源语言文本的交互信息作为图像注意力融合到文本注意力权重中，从而进一步增强解码端隐含状态与源语言文本的对齐信息。通过在多模机器翻译数据集 Multi30k 上英语-德语翻译对以及人工标注的印尼语-汉语翻译对上的实验结果表明，本文提出的模型相比已有的基于循环神经网络的多模机器翻译模型具有较好的提升，证明了所提出模型的有效性。

关键词： 多模态机器翻译；图像注意力；图像全局语义；图像局部语义.

中图分类号： TP391 **文献标识码：** A

Image Attention Fusion for Multimodal Machine Translation

Xia Li[1,2], Junteng Ma[1], Shihao Qin[1]

(1. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou,510006, China. 2. Eastern Language Processing Center, Guangzhou, 510006, China)

Abstract: Multimodal machine translation has becoming one of the research hotspots in recent years. Previous works show that visual semantic information can improve the performance of machine translation. However, most of the existing work incorporate the overall visual semantic information of the image into the translation model, while the image may contain several different local semantic object features. These different local semantic object features can have different influences on the outputs of the decoder. To this end, an multimodal machine translation model fused image attention is proposed in this paper. We incorporate the interaction information between local and global image visual information with the words of source language as an image attention into the traditional textual attention, for better alignment from hidden states of the decoder to the source words. We carried several experiments on Multi30k dataset, the results on English-German and Indonesian-Chinese tasks (the latter is annotated by human manually) show that our model has a good improvement compared with the existing recurrent neural network based multimodal machine translation model.

Key Words: Multimodal machine translation; Image attention; Global visual semantic information; Local visual semantic information.

0 引言

多模态机器翻译[1]是指融合来自文本、语音、视频及图像等多种模态信息的机器翻译模型。相较于纯文本机器翻译模型，多模态机器翻译模型可借助文本模态之外的其他模态信息辅助和提升翻译结果，弥补单模态机器翻译的不足，提升机器翻译的准确性。本文主要面向融合文本和图像两个模态的多模机器翻译。直观来看，图像视觉语义信息可以一定程度上辅助和消解文本中较难处理的语义歧义现象。例如在纯文本模态下翻译单词“bank”时，需要根据其上下文信息来推断是翻译为“堤岸”还是“银行”两个不同的语义信

息，而当借助图像视觉信息时，即当图片中包含银行时，则可以较大概率确定单词“bank”的语义信息为“银行”而不是“堤岸”。

基于多模态机器翻译模型的优点，融合文本和图像视觉信息的多模态机器翻译研究近年来得到研究人员的关注。Vinyals 等[2]在图片描述生成任务工作中，参考机器翻译中端到端框架，将编码端对传统翻译架构中的源语言文本句子的编码改为经过预训练好的卷积神经网络输出的图片向量作为输入，送入解码端作为初始隐含状态向量表示，这样在图片描述句子生成的过程中，解码端可以更为充分地使用图片中的语义信息，提升图片描述生成任务的效果。Calixto[3]等则在基于编码解码的端到端机器翻译框架下，将图片的视觉信息分别整合到翻译框架的编码端和解码端，以增强图片对文本机器翻译的提升效果。在其另一个工作中[4]则使用了两个独立的注意力机制框架单独处理源语言中单词区域和图像区域，用于提升模型的翻译结果。

已有工作从不同角度融合了图像的视觉信息，在多模态机器翻译上取得了较好的效果。然而，这些工作将图片看成整体，从全局角度抽取图片的整体视觉语义信息作为图片的隐含表示融合到文本翻译模型中。而由于图片中可能包含多个不同的语义对象(如图 1 中包含了人、马和公牛三个不同的语义对象)，他们对辅助文本翻译模型的贡献程度应该是不同的。因此，抽取局部图片和获得不同的局部视觉语义信息可以从不同角度提升多模机器翻译结果。已有工作中，Huang等[5]的工作是为数不多的融合了图片局部视觉特征的多模机器翻译工作。他们分别抽取图片的局部区域和整图区域并投影到向量空间，将其看成是伪词加入模型的输入序列中，初步探索了融入局部视觉信息和基于注意力机制的端到端多模机器翻译。Caglayan[6]则将在抽取图片的局部区域表示以及全局图片表示后，以不同方式对这些全局图和局部图与文本进行不同的融合，以提升整体的机器翻译效果。

Huang[5]和 Caglayan 的工作[6]虽然融合了图片局部特征，但他们并没有充分考虑图片的不同部分对翻译文本的不同贡献，即图片不同部分和源语言文本单词之间的语义交互信息。我们认为，可以利用图片中包含的不同局部视觉对象与源语言文本单词之间的不同交互注意力信息增强神经机器翻译模型中的纯文本注意力[7]，辅助解码端与源语言文本的对齐，从而更好地提升文本翻译结果。基于这一动机，本文提出了一种融合图像注意力的端到端多模机器翻译模型，模型将图片中的不同局部图像视觉信息及图片全局视觉信息和源语言文本单词的交互信息作为图像注意力融合到文本注意力中，使得解码端在解码时除了可以看到源文本信息，还能看到源文本中那些与图片不同区域视觉信息的注意力信息，进而更好的解码输出翻译文本。本文工作的主要贡献如下：

1)提出一种对纯文本注意力增强的图像注意力融合机制，使得解码端在解码生成目标单词时能够更好地与源语言文本对齐，从而提升多模机器翻译结果，通过在多模机器翻译数据集 Multi30k 上开展的不同实验结果表明，所提出的图像注意力融合机制可以有效提升多模机器翻译效果；

2)为了测试所提出的模型在资源稀缺的小语种数据集上的翻译效果，我们对 Multi30k 数据集中测试集和校验集的双语句对分别进行了人工标注，将其中的英语-德语句对标注为印尼语-汉语翻译句对，并将所提出的模型在该数据集上进行了测试，发现提升效果高于英语-德语的实验结果。

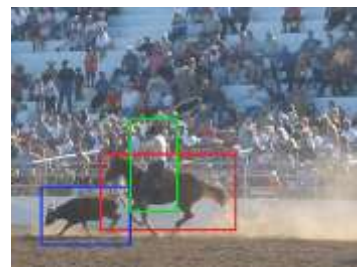


图 1. 包含多个不同语义对象的图片

1 融合图像注意力的多模机器翻译模型

经典神经机器翻译模型(Neural Machine

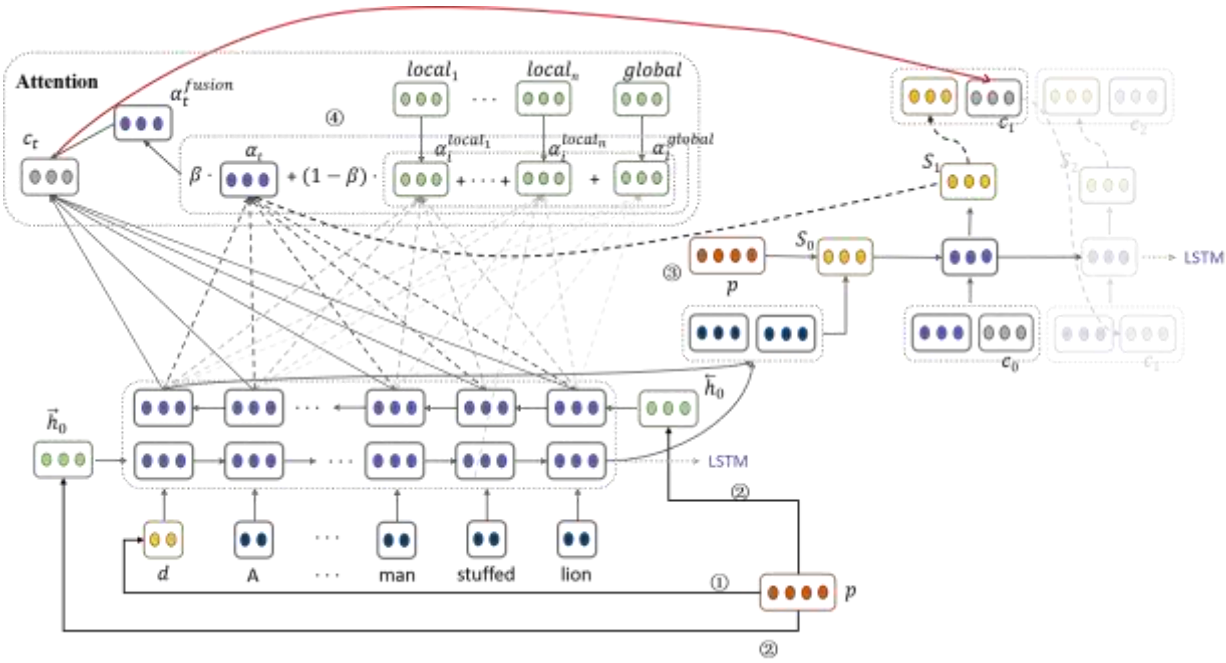


图 2. 融合图像注意力的多模机器翻译模型结构图

Translation，简称 NMT)采用的框架为基于编码和解码的序列到序列翻译模型，输入端为源语言单词序列$\mathrm{X}=(x_1,x_2,x_3,\ldots,x_M)$，输出端为目标语言单词序列$\mathrm{Y}=(y_1,y_2,y_3,\ldots,y_N)$，NMT 模型是希望学习得到 X 翻译为 Y 的概率$\mathrm{P}(Y\mid X)$最大的模型，从而学习得到训练集数据的条件概率分布。本文继续沿用基于编码解码的端到端机器翻译模型框架，在模型的编码端使用双向 LSTM 对源语言句子进行编码，在模型的解码端使用 LSTM 对目标语言句子进行解码，同时引入注意力机制使得解码端在解码预测下一个目标词时，可以获得该时刻隐含状态与源语言文本的对齐信息。所不同的是，本文提出了一种图像注意力融合机制，将图片的不同局部语义以及全局语义和源语言文本中单词的语义交互信息作为图像注意力融合到已有的文本注意力中，得到一种增强的融合注意力，使得解码端可以获取得到更为丰富的上下文信息，进一步提升模型的翻译效果，模型的整体结构图如图 2 所示。下面分别从源语言文本的编码、图像注意力融合机制、图像的全局和局部融合表示、图像视觉信息的不同融合方式以及模型训练等五个方面介绍本文模型。

1.1 源语言文本的编码

模型编码端使用双向 LSTM 对源语言句子进行编码，其中前向 LSTM 根据单词顺序，从左到右接收每个单词的词向量和上一个单元所输出的隐含向量，所得到的输出序列为$(\overrightarrow{\mathrm{h}_1},\overrightarrow{\mathrm{h}_2},\ldots,\overrightarrow{h_N})$，反向 LSTM 则从右到左接收单词序列的输入和上一个单元的隐含输出，得到输出序列$(\overleftarrow{h_1},\overleftarrow{h_2},\ldots,\overleftarrow{h_N})$。

它们的计算如式(1)~(2)所示，其中W_x为源语言单词x_i转换为词向量的词向量查找矩阵，最终编码器在每个时间戳上的输出是每个单词前向和后向隐含向量的拼接，即$h_i=[\vec{h}_i;\overleftarrow{h}_i]$，编码端源语言输入序列经过编码得到的输出为$\mathrm{h}=(h_1,h_2,\ldots,h_N)$。

$$\vec{h}_i=f_{enc}(W_x[x_i],\vec{h}_{i-1}) \tag{1}$$

$$\overleftarrow{h}_i=f_{enc}(W_x[x_i],\overleftarrow{h}_{i+1}) \tag{2}$$

1.2 图像注意力融合机制

类似于 Luong 等的工作[7]，基于纯文本注意力的神经机器翻译模型在模型的解码端通过使用注意力机制，在解码器翻译下一个目标单词时，可以更好地获取目标单词与源语言文本中某个单词的对齐关系。纯文本注意力机制的实现方法如式(3)~(6)所示：首先计算解码器 t 时刻的隐含

状态s_t和编码器隐含状态 h_i的对齐信息$e_{t,i}$，如式(3)所示，然后通过 Softmax 函数计算得到解码器 t 时刻隐含状态所对应编码器不同隐含状态 h_i 的权重$\alpha_{t,i}$，如式(4)所示。最后计算得到解码端 t 时刻隐含状态在编码器端的上下文向量c_t，如式(5)所示。最后，如式(6)所示，解码器 t 时刻的隐含状态s_t的结果由三部分计算得到：解码器的上一个隐含状态s_{t-1}；解码器 t-1 时刻预测输出的目标单词$\tilde{y}_{t-1}$(模型训练时$\tilde{y}_{t-1}$取与x_{t-1}对应的单词y_{t-1})，W_y为目标词的词向量查找矩阵；t 时刻解码器隐含状态的上下文向量c_t。

$$e_{t,i}$$

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{N} exp(e_{t,j})} \tag{4}$$

$$c_t = \sum_{i=1}^{N} \alpha_{t,i} h_i \tag{5}$$

$$s_t = f_{dec}\left(W_y[\tilde{y}_{t-1}], s_{t-1}, c_t\right) \tag{6}$$

本文将图片的视觉语义信息融入到翻译模型中，认为图片的不同区域的语义特征和源语言文本的语义交互信息可以辅助模型在解码端更好地与源语言单词进行对齐，为此提出将图像注意力融入到纯文本注意力中，得到一种增强的融合注意力。类似于文本注意力计算方法，图像注意力计算方法如式(7)~(12)所示：首先计算全局以及局部图片与编码器隐含状态 h_i 的对齐信息e_i^{global}和e_i^{localk}(k 表示第 k 个区域局部图片语义信息，$k = 1,2,\dots,L$)，如式(7)~(8)所示，其中 L 为图片所有局部区域图片的总个数，p_{global}，p_{localk}分别为全局及局部的图片特征，它们是在在预训练好的 CNN 网络中提取出来，具体的抽取方式图 3 所示。

$$e_i^{global}$$

$$e_i^{localk}$$

通过 Softmax 函数计算得到全局图片和不同局部区域图片分别和编码器隐含状态h_i所对应的影响力权重之和a_i^{img}，如式(9)所示，其中$k = 1,2,\dots,L$。我们注意到，本文提出的图像注意力a_i^{img}并没有和纯文本注意力一样与解码端 t 时刻相关，这是因为我们认为，在编码时源语言中和图片相关的单词或者出现在图片中的单词并不会随着时间的变化而变化，因此本文提出的图像注意力计算出来的对齐信息是与解码时刻t无关的。

$$a_i^{img} = \frac{exp\left(e_i^{global}\right)}{\sum_{j=1}^{N} exp\left(e_j^{global}\right)} + \frac{exp\left(e_i^{local1}\right)}{\sum_{j=1}^{N} exp\left(e_j^{local1}\right)} + \cdots \frac{exp\left(e_i^{localL}\right)}{\sum_{j=1}^{N} exp\left(e_j^{localL}\right)} \tag{9}$$

考虑到图片语义信息和解码器所要解码预测的下一个单词对源语言文本中单词的对齐关系的重要程度是基本一致的，因此本文通过对两者的加权和来获得图像注意力对文本注意力的增强效果。最终解码器 t 时刻隐含状态在编码器端的融合注意力权重$\alpha_{t,i}^{fusion}$为文本注意力权重$\alpha_{t,i}$和图像注意力权重a_i^{img}的加权求和，如式(10)所示，其中β为调节参数，$\beta \in (0,1)$。

$$\alpha_{t,i}^{fusion} = \beta \alpha_{t,i} + (1-\beta) a_i^{img} \tag{10}$$

$$c_t' = \sum_{i=1}^{N} \alpha_{t,i}^{fusion} h_i \tag{11}$$

$$s_t' = f_{dec}\left(W_y[\tilde{y}_{t-1}], s_{t-1}', c_t'\right) \tag{12}$$

本文模型在解码端 t 时刻的输出s_t'由三部分计算得到：解码器的上一个隐含状态向量s_{t-1}'；上一个隐含状态预测输出的目标词$\tilde{y}_{t-1}$，W_y为目标词的词向量矩阵；t 时刻解码器隐含状态的上下文向量c_t'，其中c_t'的计算公式如式(11)所示，整体s_t'的计算公式如式(12)所示。

1.3 融合图片不同区域的视觉语义表示

多模机器翻译任务中，图片中的一些重要语义对象常常会出现在对应的文本句子当中，因此抽取图像的局部图片特征可以有效获取其与文本的交互对应关系。为此，我们分别抽取图片的全局特征表示和不同区域的局部特征表示，并分别送入全连接网络，得到压缩后的表示并进行拼接，得到融合完整图片语义信息和不同局部图片信息的图像视觉语义向量表示p，如式(13)，抽取的案例化图如图 3 所示。

$$\mathrm{p} = [\tanh(W_g p_{global}); \tanh(W_{l1} p_{local1}); \ldots; \tanh(W_{lN} p_{localN})] \quad (13)$$

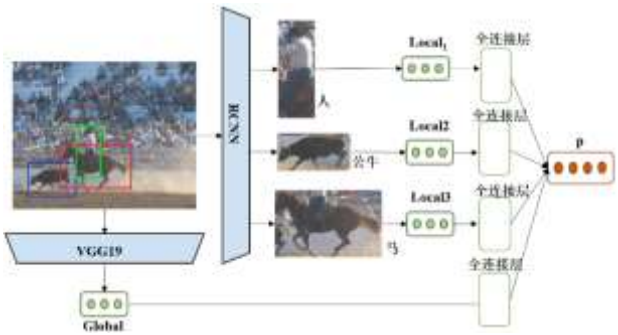


图 3. 全局图片和区域图片融合隐含表示

1.4 图视觉信息的不同融合方式

本文除了在解码端注意力机制方面提出了改进思路，同时在模型的编码和解码端也融入了图片的局部视觉语义和全局视觉语义信息，类似于 Calixto 等的工作[3]，我们采用相同的三种融合方式，所不同的是，我们不是将整个图片进行融合，而是将局部和全局图片视觉语义信息的融合隐含表示向量p(见 1.3 节)分别融合到模型的编码端和解码端，详细描述如下：

方式①为将图片向量p投影到和源语言单词相同的空间，将其当成一个伪词作为编码器的输入，我们把这种表示称为 **Local_Global$_{\mathbf{W}}$**。此时，$\mathrm{d} = W_I^2(W_I^1 \cdot p + b_I^1) + b_I^2$，其中$W_I^1$和$W_I^2$ 为可训练的转换矩阵，用于将图片转化到和单词相同的空间，b_I^1和b_I^2为偏置向量。我们把d看成是源语言单词序列的首词。

方式②为将图片视觉信息作为源语言句子编码端 LSTM 两个方向的初始化向量，分别为 $\overrightarrow{h}_0 = \tanh(W_f p + b_f)$ 和 $\overleftarrow{h}_0 = \tanh(W_b p + b_b)$，其中 W_f和W_b 是将图片向量p转换为编码器隐藏向量维度的参数矩阵，b_f和b_b为偏置向量。我们把这一融合称为 **Local_Global$_{\mathbf{E}}$**。

方式③是将图片视觉信息作为解码端的初始隐含向量作为额外输入，方法是将图片的全局和局部特征经过单层前馈神经网络后进行拼接得到向量p，通过参数矩阵W_{img}将p投影到和解码器隐含状态s_i维度相同的向量用于初始化s_0，如式(14)所示，将这种融合方式称为 **Local_Global$_{\mathbf{D}}$**。在后面的实验中，我们也用该方法初始化图像注意力融合机制模型解码器的隐含状态。

$$s_0 = \tanh(W_{di}[\overrightarrow{h}_N; \overleftarrow{h}_1] + W_{img} p + b_{di}) \quad (14)$$

1.5 模型训练

本文参考基于注意力机制的神经机器翻译模型[7]中的训练方法，使用最大似然条件概率作为损失函数对模型进行训练和优化，损失函数的计算公式如式(15)~(16)所示：

$$\min_\theta \frac{1}{N}\sum_{n=1}^{N} -log p_\theta(y_n \mid x_n) \quad (15)$$

$$\mathrm{p}(y_t|y_1, \ldots, y_{t-1}, \mathrm{x}) = \mathrm{g}(y_{t-1}, s_t, c_t) \quad (16)$$

其中θ为模型的参数，N 为训练集的样本个数，x_n代表源语言的输入序列，y_n代表目标语言输出序列，c_t为编码器第t时刻计算得到的隐含状态的上下文向量，y_t为目标单词，s_t为解码器t时刻的隐藏状态。

2 实验

2.1 实验数据

实验选取了多模态机器翻译使用的标准数据集 Multi30k[1]数据集[8]作为实验数据，该数据集是用于图像描述生成任务 Flickr30k[2]的拓展版本。数据集中每张图片由一句英文和一句由专业翻译者翻译的德语句子组成，训练集、校验集和测试集分别有 29,000 张、1,024 张和 1,000 张图片和其所对应的英德双语句对。数据预处理部分，我们针对德语和英语使用 Moses 统计机器翻译库[3]预处理脚本，包括分词、标点规范化和字母大小写正确处理等操作。实验使用 Multi30k 的训练集进行训练，并在校验集上对模型进行选择，选择最优的模型对测试集进行测试，并汇报测试集上的结果。实验采用的评价指标为机器翻译中常用的 BLEU4 指标值[9]。

2.2 实验设置

参考 Huang 等的工作[5]，在图像语义抽取方

[1] http://www.statmt.org/wmt16/multimodal-task.html
[2] http://shannon.cs.illinois.edu/DenotationGraph/
[3] https://github.com/moses-smt/mosesdecoder

面使用 ImageNet[10]上预训练好的 VGG19[11]中倒数第二个全连接层 4,096 维向量作为图像的全局特征表示。对于局部图像语义抽取，本文使用了由 Ren 等[12]提出的 RCNN 方法抽取每张图片所对应的局部区域，RCNN 中的区域提案网络在 VOC 2007[4]和 2012[5]数据集上事先预训练好。本文将区域提案网络抽取得到的局部区域再通过 VGG19 使用全局图片特征相同的抽取方式抽取得到所有局部区域的语义向量表示。

编码器端双向 LSTM 中的前向和后向 LSTM 隐含状态向量均为 512 维，源语言和目标语言的词向量维度为 620 维，解码器端单向 LSTM 隐含状态向量 512 维。模型采用 Adam 优化器[13]，初始学习率为 0.001，batch size 设置为 32，一共训练 26 轮，如果连续 4 轮模型在校验集上的困惑度值没有下降，则执行一次学习率衰减，衰减为原来的二分之一，如果这种情况出现了 5 次，则使用早期停止策略(early stop)结束训练。另外，本文的所有实验在融合图像注意力公式上的参数 β 均设定为 0.9。在测试集上，参考 Sutskever 等的工作[14]，使用集束搜索(beam search)，其中实验中的 beam size 设置为 10。

2.3 实验结果与分析

为了验证本文模型的有效性，我们分别选取了 Huang 等[5]、Calixto 等[3]和 Caglayan 等[6]提出的多模机器翻译模型作为我们的三个基准方法。同时，我们还分别比较了几种不同融合方式下使用文本注意力机制和使用本文提出的融合图像注意力机制的不同方法的实验结果对比。表 1 中的 $\text{Local\_Global}_{\text{w}}$、$\text{Local\_Global}_{\text{D}}$、$\text{Local\_Global}_{\text{W+D}}$ 和 $\text{Local\_Global}_{\text{W+E+D}}$ 为 1.4 节中介绍的方式用 1.3 节所抽取的图片特征分别当成伪词，初始化解码器，并分别采用两种方式融合或三种方式融合的方法。模型 Fusion_Attention_Global 和 Fusion_Attention_Global_Local 则是使用图像注意力融合机制(见 1.2 节)并分别使用图像特征初始化解码器的隐含状态，区别是前者只使用全局图片特征，而后者则是局部及全局的图片特征。

2.3.1 英语-德语翻译对上的实验结果

首先我们在 Multi30k 数据集上的英语-德语翻译对数据进行了实验，实验结果如表 1 所示。从表 1 可以看出，本文模型相比已有的采用局部图片融合方式的工作(Huang et al., 2016[5])、采用全局图片融合方式的工作(Calixto et al., [3])等不同图像融合方式的模型均有所提升。其中，本文的最优模型(Fusion_Attention_Global_Local)相比 Huang 等的工作[5]提升了 2.24 个 BLEU 值，相比 Calixto 等的工作 [3]提升了 1.44 个 BLEU 值，相比 Caglayan 等的工作[6]提升了 0.94 个 BLEU 值。在 1.4 节中我们提到，在我们融合图像注意力的模型的实验里使用了图片特征来初始化解码器的向量，从表 1 的实验结果上看，Fusion_Attention_Global 以及 Fusion_Attention_Global_Local 分别要比 Calixto 等的只融合全局图片进行初始化解码器隐含向量的工作[3]以及 $\text{Local\_Global}_{\text{D}}$ 都要有所提升，验证了图像注意力机制的提升效果。

表 1 英语-德语翻译对上的实验结果

	翻译模型	BLEU4
RNN-based	Huang et al., 2016[5]	36.5
	Calixto et al., 2017[3]	37.3
	Caglayan et al., 2017[6]	37.8
Our method	$\text{Local\_Global}_{\text{w}}$	38.51
	$\text{Local\_Global}_{\text{D}}$	38.17
	$\text{Local\_Global}_{\text{W+D}}$	38.67
	$\text{Local\_Global}_{\text{W+E+D}}$	38.21
	Fusion_Attention_Global	38.65
	Fusion_Attention_Global_Local	**38.74**

同时从表 1 的实验结果还可以看出，本文模型中使用的融合了图片的全局语义特征以及图片的不同区域的局部语义特征的融合方式，通过在模型的编码端和解码端分别融入（见 1.3 节和 1.4 节），整体在翻译效果上有所提升，这表明，本文方法虽然采用与 Calixto 等[3]相同的三种融合方式，但是 Calixto 等[3]只是采用了全局图片语义信息，而本文模型则是采用了不同局部语义和全局语义图像信息，这表明融合图片不同区域

[4] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/
[5] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

局部语义信息确实可以为文本机器翻译模型提供更多语义信息用于辅助提升翻译质量。

2.3.2 印尼语-汉语翻译对上的实验结果

本文还测试了在资源稀缺的小语种数据上的实验效果，为了进行测试，我们对 Multi30k 中校验集和测试集的所有英语-德语翻译对进行了人工标注，全部标注为一一对应的印尼语-中文翻译对数据。对于 Multi30k 面向印尼语-汉语的训练集数据，采用 GOOGLE 翻译[6]分别对英语-德语翻译句对翻译为相应的印尼语-汉语翻译句对，从而得到相应的训练集数据。

表 2 印尼语-中文翻译对上的实验结果

	翻译模型	**BLEU4**
基准模型	Text-only NMT	27.48
Our method	Local_Global_{W}	28.15
	Local_Global_{D}	27.83
	$\text{Local\_Global}_{W+D}$	27.80
	Fusion_Attention_Global	29.71
	Fusion_Attention_Global_Local	**29.75**

本文提出的模型以及不同的融合方式在印尼语-汉语翻译对上的实验结果如表 2 所示。其中表 2 的第一行所对应的模型 Text-only NMT 采用的是基于纯文本注意力的神经机器翻译，其与本文模型的区别是没有融入任何图像视觉信息，并且采用纯文本注意力机制。从表 2 的实验结果可以看出，本文模型相比纯文本注意力机制的神经机器翻译模型提升较为明显。本文模型的最优结果相比纯文本注意力模型在印尼语-汉语翻译对数据测试集上提升了 2.27 个 BLEU 值。而没有采用图像注意力，但是融合了图像视觉信息的模型（融合局部和全局信息的视觉语义信息，见 1.3 节），其结果相比没有融合任何图像视觉信息的纯文本注意力翻译模型也均有一定的提升。例如，将图片向量p投影到和源语言单词相同的空间，将其当成一个伪词作为编码器的输入**(Local_Global_{W})**模式相比没有融入图像视觉语义信息的纯文本注意力模型结果提升了 0.67 个 BLEU 值。

同时我们发现，相同的模型在印尼语-汉语小语种数据集上的整体 BLEU4 值低于英语-德语翻译对上的结果，一方面可能是因为训练集为 GOOGLE 翻译后的数据存在较大的噪音，另一方面也可能是因为印尼语和汉语属于两个不同的语系，两者数据分布的差异性更大。

2.4 案例分析

为了更好地展示本文模型的效果，我们对数据集的翻译结果进行了案例分析，本文选取了印尼语-汉语翻译句对上的翻译结果进行案例分析。分别选取了 2 副图像以及基于纯文本注意力的模型和参考人工翻译句子，以及本文所提出的模型的翻译结果对比。其中图 4 和图 5 分别是两个案例所对应的图片，表 3 和表 4 则是不同模型在这两幅图片上的翻译结果。

图 4. 案例 1 图片

图 5. 案例 2 图片

以案例 1 所对应的图 4 为例，不同模型的实验结果如表 3 所示。从实验结果中可以看出，图片的原始印尼语为“wanita yang berjubah merahmenari dengan pria berjas .”，所对应的人工翻译文本为“身穿红色连衣裙的女士正与穿着西装的男子**共舞**。”，纯文本注意力模型（没有融合图像视觉语义信息）其翻译的结果为“穿着红色长袍的女士和穿着西装的男人。”，我们可以看到翻译结果并没有翻译出共舞这个动作，而本文提出的模型均能很好的翻译。

[6] https://translate.google.com/

表 3 不同模型在案例 1 上的机器翻译效果

翻译模型	句子
印尼语原文	wanita yang berjubah merah menari dengan pria berjas .
人工中文翻译	身 穿 红色 连 衣裙 的 女士 正 与 穿 着 西装 的 男子 **共舞** 。
Text-only NMT	穿 着 红色 长袍 的 女士 和 穿 着 西装 的 男人 。
Local_Global$_W$	穿 着 红色 长袍 的 女人 与 穿 着 西装 的 男人 **一起 跳舞** 。
Fusion_attention_global	穿 着 红色 长袍 的 女人 与 穿 着 西装 的 男人 **共 舞** 。
Fusion_attention_global_local	红色 长袍 的 妇女 **跳舞** 与 穿 着 西装 的 男人 。

表 4 不同模型在案例 2 上的机器翻译效果

翻译模型	句子
印尼语原文	polisi anti huru hara berdiri di belakang sementara seorang pria muda dengan syal merah menutupi wajahnya berjalan .
人工中文翻译	防暴警察站在后面，一名戴着红色围巾的年轻男子**捂着脸**走路。
Text-only NMT	防暴 警察 站 在 后面 ， 而 一 名 戴 着 红色 围巾 的 年轻人 在 她 的 脸上 <unk> 。
Local_Global$_W$	警察 防暴 警察 站 在 后面 ， 而 一个 年轻人 戴 着 红色 围巾 <unk> 走 。
Fusion_attention_global	防暴 警察 站 在 后面 ， 而 一个 戴 着 红色 围巾 的 年轻人 正在 走路 。
Fusion_attention_global_local	防暴 警察 站 在 后面 ， 而 一个 年轻人 用 一 条 红色 的 围巾 **遮住** 了 走 去 。

针对案例 2 所对应的图 5，不同模型的实验结果如表 4 所示。从实验结果可以看出，本文提出的模型则能很好的翻译出该句印尼语，而纯文本模型翻译的结果为“防暴 警察 站 在 后面 ，而 一 名 戴 着 红色 围巾 的 年轻人 在 她 的 脸上 <unk> 。”效果相对不好。实验结果中也发现，本文提出的模型整体上融合局部和全局图片信息的结果在 BLEU4 指标上略高于融合全局图片信息的结果。但部分翻译效果后者优于前者。

2.5 可视化分析

2.5.1 图像和源语言单词的对齐可视化

为了观察图像信息是否能够更好的对齐源语言的单词，我们单独计算了图片对每个源语言单词h_i所计算的权重a_i^{img},如图 6 所示。

图 6. 图片对源语言句子单词权重a_i^{img}最高的单词

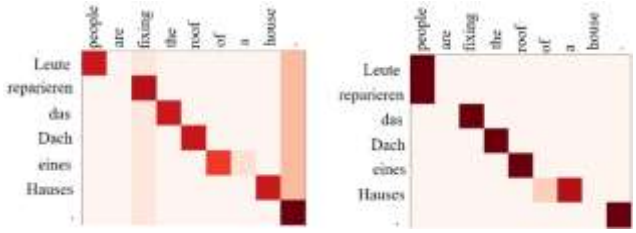
我们用红色加粗标出权重最高的几个单词。可以看出在这 4 个例子中，权重较高的单词基上都出现在图片当中。例如，图 6 中的(1)对齐到了“swimming”和“pool”；(2)对齐到了“uniform”“palms”等；(3)对齐到了“rides”，“bull”等；(4)对齐到了“playing”和“soccer”等。同时我们也发现，有些关键单词如“girl”、“boy”等这些在图中比较显眼的局部对象并没有很好的捕捉到，同时有部分句子所计算的权重的效果比较一般，我们认为这与抽取的局部图片特征有一定关系，因为在抽取时我们所使用的预训练模型是在较小的数据集上预训练的，可能导致识别的物体的精确度和数量不够准确。

2.5.2 图像注意力的可视化效果

为了展示所提出的图像注意力对文本注意力的增强效果，我们对部分例子进行了可视化，展示使用文本注意力和使用融合图像注意力增强后的对齐可视化效果对比。

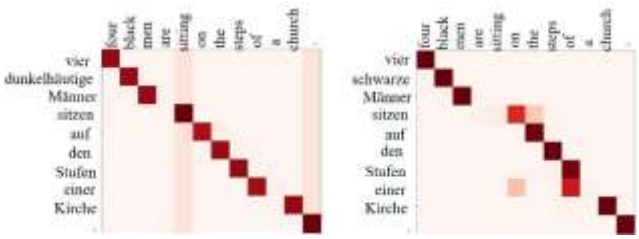
如图 7 所示，可视化分析案例 1 中为英语-德语翻译句对，其中源语言句为“people are fixing the roof of a house.”，目标语言句为“Leute reparieren das Dach eines Hauses.”我们可以看到，目标语言单词“Leute”在文本注意力和本文提出的融合图像的注意力上都对齐到了源语言单词“people”，而单词“reparieren”在文本注意力上错误对齐到源语言单词“people”，而本文提出的融合图像注意力模型中，单词“reparieren”正确对齐到了源语言单词“fixing”。同样的还有单词：“dach”在文本注意力模型中错误对齐到单词“the”，而在本文的模型中正确对齐到单词“roof”。

同样在英语-德语句对“four black men are sitting on the steps of a church.”和“vier schwarze Männer sitzen auf den Stufen einer Kirche.”案例中（见图 8 可视化案例分析 2），目标语言中的单词“sitzen”和“Stufen”均在文本注意力中错误对齐单词，而在我们的模型中则分别正确对齐单词“sitting”和“steps”，表明我们方法的有效性。



(a) Image fusion attention (b) Text-only attention

图 7. 可视化分析案例 1



(a) Image fusion attention (b) Text-only attention

图 8. 可视化分析案例 2

3 相关工作

机器翻译从统计机器翻译方法[15,16,17,18]开始，相关技术发展迅速，如基于短语的统计机器翻译方法和融合不同结构的统计机器翻译方法[19,20,21]等。近年来，机器翻译研究从基于统计的方法转为基于深度学习的方法，在基于深度学习的机器翻译模型或方法中，主要研究侧重于使用端到端神经网络方法来实现翻译建模。英国牛津大学Kalchbrenner和Blunsom[22]于2013年在分布式连续表示的基础上提出了神经机器翻译，该模型采用神经网络以端到端的方式进行翻译建模，是机器翻译领域较早提出神经机器翻译概念和模型的研究。加拿大蒙特利尔大学Cho等[23]和谷歌公司Sutskever等[14]于2014年分别对该方法进行了完善，更好的推动了基于神经网络模型的机器翻译。在后续的一些研究中，研究人员提出了将注意力机制应用于机器翻译研究当中，取得了很好的效果。Bahdanau等[24]在Cho等[23]工作的基础上引入注意力机制，显著提高神经机器翻译模型的翻译性能。Luong等[7]在Bahdanau等[24]提出的注意力机制上进行改进，提出了全局注意力机制和局部注意力机制。同时，Zhou等[25]提出一个多源的机器翻译架构，将NMT与SMT的输出作为其输入。除了用RNN实现seq2seq的机器翻译之外，有研究者开始用CNN来构建序列模型[26,27]。Gehring等[28]提出了一种完全基于卷积神经网络（CNN）来实现机器翻译中的“编码器-解码器”结构。

近年来，融合多个模态信息的机器翻译研究成为机器翻译研究的一个热点。如 Vinyals 等[2]提出了一种 IDG 模型，该模型将图像编码为一个向量作为基于 seq2seq 框架的输入。Calixto 等[4]使用两个独立的注意力机制，能够单独地处理源语言单词和图像的区域。Huang 等[5]提出了用 VGG19 和 RCNN 抽取全局以及局部的图片特征，将他们投影到向量空间，加入模型输入的源语言单词序列当中。Calixto[3]等还提出仅使用图片的全局特征，以单词输入、初始化解码器、初始化编码器三种不同的方式融合进神经机器翻译模型当中。Caglayan[6]则是将图片信息以多种渠道融合：一种是计算区域图片特征与目标单词的上

下文向量与纯文本的上下文向量进行拼接，另外一种则是将图片特征与编码器或解码器的输做点乘计算，或用于初始化编码器及解码器。

已有基于递归神经网络模式的多模态机器翻译工作主要体现在两个方面：一方面是图片视觉信息融合在模型的哪些位置以及融合方式等，另一方面则是如何改进解码端注意力方面的工作。本文正是从这两个方面对已有相关工作(Huang et al., 2016[5]、Calixto et al., 2017[3]、Caglayan et al., 2017[6])进行的扩展，一方面提出了融合图像注意力机制，用于增强文本注意力，更好的辅助解码端预测目标词时对齐源语言文本的单词，同时本文还充分使用了图片的全局图片视觉语义信息和不同区域的局部视觉语义信息用于对模型进行融合编码和对解码端的初始化，本文的实验结果也验证了本文所提出模型的可行性。

4 结论

多模机器翻译模型由于融合了不同模态的多方面的信息，使得机器翻译结果得到了较好的提升，因此近年来多模机器翻译研究成为新的研究热点之一。本文面向融合图片视觉语义信息的多模机器翻译任务，提出了一种融合图像注意力的多模机器翻译模型，通过将图像中不同部分的视觉语义信息及全局视觉语义信息和编码器端不同隐含状态（源语言文本）的交互信息作为图像注意力融入到纯文本注意力中，得到一种增强的图像注意力多模机器翻译模型。

模型分别在 Multi30k 英语-德语翻译对和印尼语-汉语翻译对数据集上进行了实验测试，结果表明本文提出的模型相比已有基线系统（不同角度融合图像视觉语义信息到翻译模型中）均具有较好的提升，尤其在印尼语-汉语这种资源稀缺的小语种数据集翻译对上，结果提升更为明显，验证了本文所提出的模型的有效性。

本文开展了基于 RNN 框架下的融合图像视觉语义信息的多模态机器翻译模型，并对提出的模型在实验数据集上进行了验证。后续工作将进一步探索融合图像视觉语义信息在基于 transformer 框架下的多模机器翻译模型的工作。

参考文献

[1] BALTRUSAITIS T, AHUJA C, and MORENCY L P. Multimodal Machine Learning: A Survey and Taxonomy. [C/OL] //IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. http://arxiv.org/abs/1705.09406.

[2] VINYALS O, TOSHEV A, BENGIO S, et al. Show and Tell: A Neural Image Caption Generator. [C/OL] // IEEE Conference on Computer Vision and Pattern Recognition, 2015. https://arxiv.org/pdf/1411.4555.pdf

[3] CALIXTO I, LIU Q, CAMPBELL N. Incorporating Global Visual Features into Attention-Based Neural Machine Translation. [C/OL] //Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017. https://arxiv.org/pdf/1701.06521.pdf.

[4] CALIXTO I, LIU Q and CAMPBELL N. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. [C/OL] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1913-1924. https://arxiv.org/pdf/1702.01287.pdf.

[5] HUANG P Y, LIU F, SHIANG S R, et al. Attention-based Multimodal Neural Machine Translation. [C/OL] //Proceedings of the First Conference on Machine Translation, 2016. https://www.aclweb.org/anthology/W16-2360.

[6] CAGLAYAN O, ARANSA W, BARDET A et al. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. [C/OL] // Proceedings of the Conference on Machine Translation, 2017. https://www.aclweb.org/anthology/W17-4746.

[7] LUONG M T, PHAM H, MANNING C D. . Effective Approaches to Attention-based Neural Machine Translation. [C/OL] // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015. https://aclweb.org/anthology/D15-1166

[8] ELLIOTT D, FRANK S, SIMA'AN K. Multi30K:

Multilingual English-German Image Descriptions. [OL] https://arxiv.org/pdf/1605.00459.pdf

[9] PAPINENI K, ROUKOS S, WARD T, et al. BLEU:a Method for Automatic Evaluation of Machine Translation. [C] // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

[10] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge. [J/OL]. International Journal of Computer Vision, 2015, 115(3): 211-252. https://arxiv.org/pdf/1409.0575.pdf [11] K. Simonyan and A. Zisserman, “very deep convolutional networks for large-scale image recognition,” *ICLR*, pp. 1–14, 2015.

[11] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. [C/OL] // International Conference on Learning Representations, 2015. https://arxiv.org/pdf/1409.1556.pdf

[12] Ren S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. [C] // In Advances in Neural Information Processing Systems (NIPS), 2015.

[13] KINGMA D P and BA J L. Adam: A Method for Stochastic Optimization. [C/OL] // International Conference on Learning Representations, 2015. https://arxiv.org/pdf/1412.6980.pdf.

[14] SUTSKEVER I, VINYALS O, and LE Q V. . Sequence to Sequence Learning with Neural Networks. [C/OL] // Advances in Neural Information Processing Systems, 2014. https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

[15] 刘群. 统计机器翻译综述[J]. 中文信息学报，2003, 17(4): 2-13.

[16] BROWN P F. , COCKE J , PIETRA S A. D , et al. A Statistical Approach to Machine Translation [J]. Computational Linguistics ,1990.

[17] BROWN P F. , PIETRA S A. D, PIETRA V J. D, et al. The Mathematics of Statistical Machine Translation : Parameter Estimation [J]. Computational Linguistics , 19, (2), 1993.

[18] OCH F J, NEY H, Discriminative Training and Maximum Entropy Models for Statistical Machine Translation [C] // Association for Computational Linguistics, 2012.

[19] LIU Y, WANG K, ZONG C, et al. A unified framework and models for integrating translation memory into phrase-based statistical machine translation[J]. Computer Speech & Language, 2019, 54: 176-206.

[20] ZHANG J, ZONG C . Learning a Phrase-based Translation Model from Monolingual Data with Application to Domain Adaptation. [C] // The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), 2013.

[21] TU Z P, LIU Y, HWANG Y S, et al. Dependency Forest for Statistical Machine Translation. [C] // Proceedings of the 23rd International Conference on Computational Linguistics, 2010.

[22] KALCHBRENNER N, BLUNSOM P. Recurrent Continuous Translation Models. [C/OL] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013. https://www.aclweb.org/anthology/D13-1176.

[23] CHO K, BAHDANAU D, BOUGARES F, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. [C/OL] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. https://www.aclweb.org/anthology/D14-1179

[24] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate. [C/OL] // International Conference on Learning Representations, 2015. https://arxiv.org/pdf/1409.0473.pdf

[25] ZHOU L, HU W, ZHANG J, et al. Neural system combination for machine translation. [C] // Association for the Advancement of Artificial Intelligence, 2017.

[26] BRADBURY J, MERITY S, XIONG C et al.

Quasi-Recurrent Neural Networks. [C/OL] // Association for the Advancement of Artificial Intelligence, 2017. https://arxiv.org/pdf/1611.01576.pdf

[27] KALCHBRENNER N, ESPEHOLT L, SIMONYAN K, et al. Neural Machine Translation in Linear Time. [OL] 2016. https://arxiv.org/pdf/1610.10099.pdf.

[28] GEHRING J, AULI M, GRANGIER D, et al. Convolutional Sequence to Sequence Learning. [C/OL] // Proceeding ICML'17 Proceedings of the 34th International Conference on Machine Learning, 2017. https://arxiv.org/pdf/1705.03122.pdf