基于图文细粒度对齐语义引导的多模态神经机器翻译方法

叶俊杰[1,2]，郭军军[1,2,*]，谭凯文[1,2]，相艳[1,2]，余正涛[1,2]
1.昆明理工大学信息工程与自动化学院，云南昆明650500
2.昆明理工大学云南省人工智能重点实验室，云南昆明650500
junjieye.cdx@qq.com,guojjgb@163.com, kwtan0909@qq.com,
sharonxiang@126.com,ztyu@hotmail.com

摘要

多模态神经机器翻译旨在利用视觉信息来提高文本翻译质量。传统多模态机器翻译将图像的全局语义信息融入到翻译模型，而忽略了图像的细粒度信息对翻译质量的影响。对此，该文提出一种基于图文细粒度对齐语义引导的多模态神经机器翻译方法，该方法首先跨模态交互图文信息，以提取图文细粒度对齐语义信息，然后以图文细粒度对齐语义信息为枢纽，采用门控机制将多模态细粒度信息对齐到文本信息上，实现图文多模态特征融合。在多模态机器翻译基准数据集Multi30K 英语→德语、英语→法语以及英语→捷克语翻译任务上的实验结果表明，论文提出方法的有效性，并且优于大多数最先进的多模态机器翻译方法。

关键词： 多模态神经机器翻译；图文细粒度；语义交互；对齐语义；Multi30K

Based on Semantic Guidance of Fine-grained Alignment of Image-Text for Multi-modal Neural Machine Translation

Junjie Ye[1,2], Junjun Guo[1,2,*], Kaiwen Tan[1,2], Yan Xiang[1,2], Zhengtao Yu[1,2]
1.Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2.Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China
junjieye.cdx@qq.com,guojjgb@163.com, kwtan0909@qq.com,
sharonxiang@126.com,ztyu@hotmail.com

Abstract

Multi-modal neural machine translation aims to use visual information to improve the quality of only-text translation. Traditional multi-modal machine translation incorporates the global semantic information of images into the translation model, while ignoring the influence of fine-grained information of images on translation quality. In this regard, this paper proposes a multi-modal neural machine translation method based on fine-grained alignment of image-text with semantic guidance. The method firstly interacts image and text information across modalities to extract fine-grained aligned semantic information of image-text. Fine-grained alignment of semantic information is used as a pivot, and multi-modal fine-grained information is aligned to textual information using a gating mechanism to achieve multi-modal feature fusion of image and text. The experimental results on the multi-modal machine translation benchmark dataset Multi30K English → German, English → French and English → Czech translation tasks show that the proposed method is effective and outperforms large Most state-of-the-art multi-modal machine translation methods.

Keywords: Multi-modal neural machine translation , Fine-grained , Semantic interaction , Alignment semantics , Multi30K

第二十一届中国计算语言学大会论文集，第281页-第292页，南昌，中国，2022年10月14日至16日。

1 引言

多模态神经机器翻译（multi-modal neural machine translation, MNMT）(Caglayan et al., 2019; Yin et al., 2020; Li et al., 2021) 旨在利用额外模态信息（如图像、视频、声音）优化传统的文本机器翻译模型，通过融合图像等多模态特征提升机器翻译的性能(Caglayan et al., 2019; Yao and Wan, 2020; Ye and Guo, 2022)。近年来多模态神经机器翻译受到国内外研究者的广泛关注，相较于传统纯文本机器翻译系统，多模态神经机器翻译通过融合图像模态的信息，不仅可以提高翻译性能，还可以补全文本语义信息以及解决歧义词翻译问题(Huang et al., 2020; Li et al., 2021)。

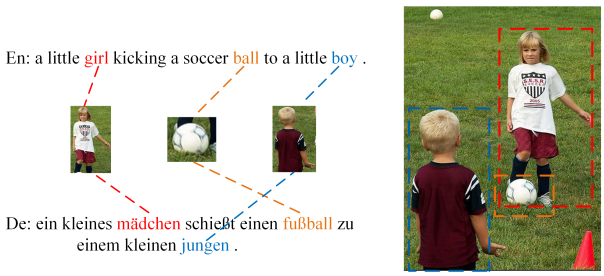

Figure 1: 图像模态和文本模态的语义表示

图像是一种语种无关的模态信息，可以同时被不同语言的人理解，因此可以利用视觉信息为枢轴跨越语言障碍，如上图1 中，不同语言对可以对齐在同一个图像区域，例如：女孩、男孩、足球等。然而，视觉和文本两种模态信息之间存在较大的语义鸿沟，跨模态表示学习及语义对齐通常较难。视觉和文本（源语言）的语义对齐存在两个层级，1）文本实体与图像全局语义信息对齐，图像全局特征可以提供有效的场景信息，例如：图1全局信息为草地上两个小孩子在踢球，2）文本实体与图像对象通过细粒度对齐，图像的细粒度特征信息可以增强文本中的重要信息，例如：图1 中的对象“女孩”、“足球”、“男孩”对齐到文本实体“girl”、“ball”、“boy”。

目前，已有多模态神经机器翻译模型主要通过设计合理的多模态融合模型，实现图文信息的有效整合，主要包含三种图文融合策略：1）跨模态注意力机制(Kwon et al., 2020; Song et al., 2021; Zhao et al., 2021; Li et al., 2021)，将图文模态映射到同一空间向量，然后采用注意力机制抽取与文本信息相关的视觉区域。2）多模态Transformer 融合方法，使用Transformer 分别编码文本特征和视觉特征(Takushima et al., 2019; Nishihara et al., 2020)，然后采用多头注意力机制或者拼接方法将它们融合作为编码器的输出(Yao and Wan, 2020; Gain et al., 2021; Li et al., 2021)。3）门控融合方法(Yin et al., 2020; Lin et al., 2020; Li et al., 2021)，基于多模态门控机制过滤图像中与文本关联性不强的信息，实现图文对齐融合。通过上述方式，文本语义信息和图像语义信息有一个简单的对齐。

然而图像和文本之间存在较大的语义鸿沟，仅仅采用上述图文融合策略，很难实现图像和文本语义的细粒度对齐和融合。为了提升图文对齐融合的能力，本文提出了一种基于图文细粒度对齐语义引导的多模态神经机器翻译方法，采用软对齐的方式实现图文特征的层级融合，图文细粒度对齐语义为枢纽，采用多模态门控机制比对图文两种模态信息，实现图文特征对齐融合，提升了多模态神经机器翻译的性能。与以前的工作相比，本文的主要贡献是两方面的：

- 提出一种图文细粒度对齐语义引导的多模态神经机器翻译方法，采用跨模态注意力机制，以图文细粒度对齐语义为引导，实现了融合图文信息的多模态神经机器翻译。

- 基于多模态机器翻译公共数据Multi30k的实验结果表明，本文提出的模型优于其它多模态机器翻译方法，并显著提高了英德、英法和英捷克语机器翻译的性能。

* 通讯作者：郭军军email地址：guojjgb@163.com
项目基金：国家重点研发计划(2020AAA0107904)；国家自然科学基金(61866020, 61762056)；云南省科技厅自然科学基金项目(2019FB082, 2019QY1801)

2 相关工作

近年来，多模态神经机器翻译受到了广泛的关注，特别是图文多模态融合方法在许多任务中都显示出巨大的潜力。国内外研究学者针对图文多模态机器翻译融合方法开展了许多研究，并取得了一定进展。目前，多模态神经机器翻译主要有基于循环神经网络（recurrent neural network, RNN）的机器翻译模型，与基于Transformer 的机器翻译模型。

2.1 基于RNN的多模态神经机器翻译模型

早期的多模态融合方法主要是基于循坏神经网络（RNN）的seq2seq 框架，利用全局视觉特征初始化RNN 编码器解码器的隐藏状态(Calixto et al., 2017b; Caglayan et al., 2017; Huang et al., 2016)，或利用视觉特征增强文本语义表征能力，提升机器翻译的性能(Huang et al., 2016)。尽管这些方法提高了机器翻译的性能，但视觉特征实际上并没有与文本特征对齐。为了更好地对齐视觉和文本语义特征，Caglayan et al. (2016b; Caglayan et al. (2016a)利用多模态注意力机制同时关注图像及其对应的文本，以对齐视觉和文本语义特征；Calixto et al. (2017a) 分别对源句子单词和图像采用了两种特定于模态的注意机制，以更好地对齐视觉和文本特征。Delbrouck and Dupont (2017) 提出了一种局部视觉注意机制，将局部视觉特征与相应的文本特征对齐。李志峰et al. (2020) 提出一种融合覆盖机制双注意力解码方法，借助覆盖机制分别作用于源语言和源图像，以解决模型过翻译及欠翻译问题。

2.2 基于Transformer的多模态神经机器翻译模型

随着机器翻译技术的发展，基于Transformer结构的多模态神经机器翻译方法被提出。我们将现有的多模态融合策略从三个方面总结如下：**1）**跨模态交互注意机制，Zhao et al. (2022) 利用对象检测特征和额外的区域相关注意机制来融合视觉区域特征和文本特征；Nishihara et al. (2020) 提出了一个有监督的跨模态注意模块，用于对齐文本特征和视觉特征；Song et al. (2021) 在每个Transformer 编码器层采用了一个共同注意图更新模块来对齐多模态特征。**2）**特征连接方法，Yao and Wan (2020) 使用多模态Transformer 来对齐视觉特征和文本特征；Takushima et al. (2019) 拼接视觉全局特征和文本特征作为多模态特征；Takushima et al. (2019) 直接连接文本表示和视觉表示作为多模态表示，以保留细粒度特征并避免模态特定特征的混淆。**3）** 门控融合方法，Yin et al. (2020) 提出了一种基于图的多模态神经机器翻译方法，通过文本图像门控注意力机制提取多模型特征；Lin et al. (2020) 采用门控机制来融合动态上下文引导胶囊网络提取的视觉特征；Li et al. (2021) 使用门控融合方法解决歧义词翻译问题；Zhao et al. (2021) 基于多模态Transformer，提出了一种词域对齐引导的方法来建立文本和视觉特征之间的语义相关性。

3 方法

论文提出了一种图文细粒度对齐语义引导的多模态神经机器翻译模型，模型总体架构如下图2 所示。它主要包含四个网络：图文编码器、跨模态语义交互模块、多模态语义融合模块及解码器。

3.1 图文编码器

3.1.1 图像和文本模态信息表征

不失一般性，将$x_k = \{x_1^k, \ldots, x_n^k\}$ 和z_k 分别表示为源句输入及其对应图像，其中k 表示图文对的序号，n 是x_k 的序列长度。文本序列通过带有位置嵌入的传统嵌入层嵌入，图像特征由预训练的ResNet-101 模型(He et al., 2016) 提取。文本表征向量E_k^x 和视觉表征向量E_k^z 计算如下：

$$\mathrm{E}_k^x = \mathrm{Emb}_x(x_k) + \mathrm{PE}_x(x_k) \quad (1)$$

$$\mathrm{E}_k^z = \mathrm{Emb}_z(z_k) \quad (2)$$

其中，Emb_x 是文本序列词嵌入层，PE_x 是位置嵌入层，Emb_z 是基于ResNet-101 的视觉特征提取层，$\mathrm{E}_k^x \in R^{n \times d_1}$ 和$\mathrm{E}_k^z \in R^{7 \times 7 \times d_2}$，论文中$d_1$=128，$d_2$=2048。

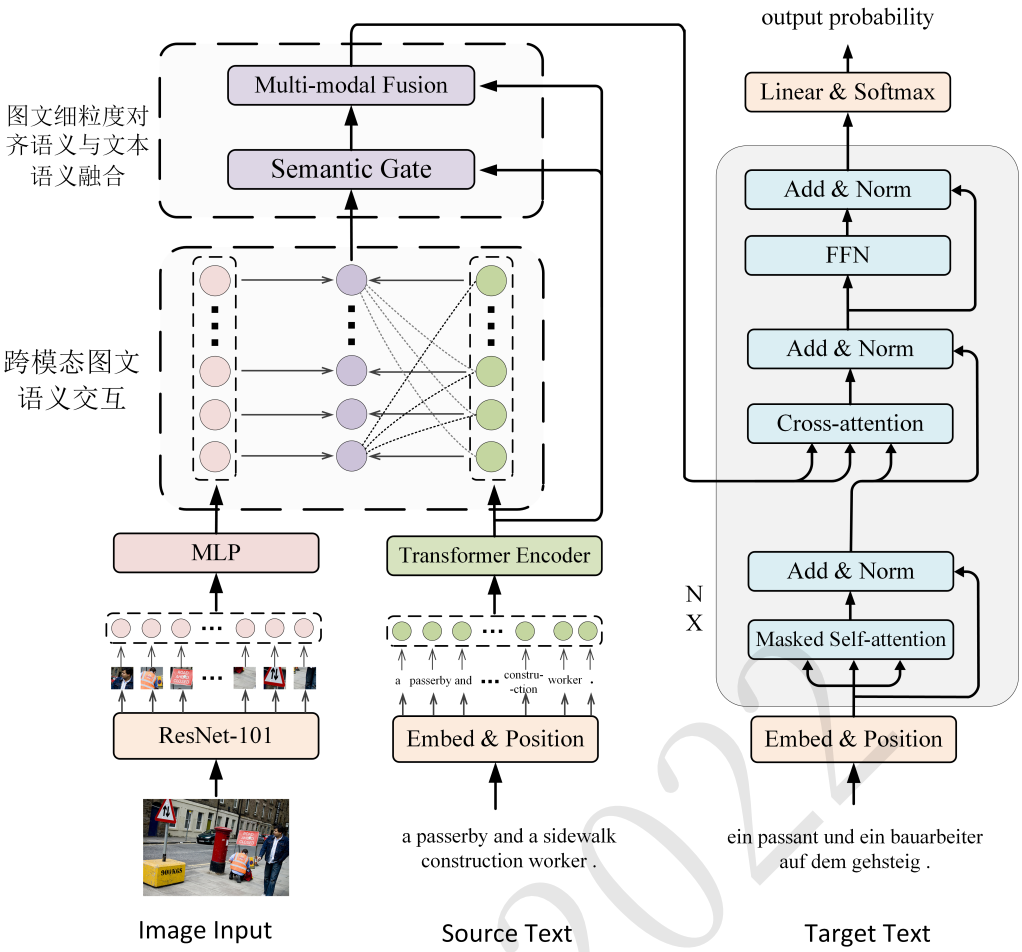

Figure 2: 基于图文细粒度对齐语义引导的多模态神经机器翻译模型

3.1.2 源语言文本编码器

编码器采用传统多头Transformer 编码器，具体框架如图3 所示，每个编码器层由两个子层组成：多头自注意力层和位置前馈网络（FFN）层。首先使用多头自注意力模块，将源文本表示作为查询/键/值矩阵来建立单词到单词的相互连接，可以表示为，

$$\mathrm{H}_{x_k}^{l} = \mathrm{Multihead}(\mathrm{E}_k^x, \mathrm{E}_k^x, \mathrm{E}_k^x) \tag{3}$$

$$= \mathrm{Concat}(head_k^1, \cdots, head_k^M) \tag{4}$$

其中，M 表示头数，Multihead$(\cdot)$ 是多头注意力层，$l = \{0, \cdots, 3\}$ 是Transformer 层索引。多头注意力的输出计算如下：

$$\mathrm{head}_k^{c\in[1,M]} = \sum_{j=1}^{n} \alpha_{ij}(\mathrm{E}_{k_j}^x \mathbf{W}_{k,c}^V) \tag{5}$$

其中n 表示源语言序列x_k的长度，α_{ij}为自注意力权重系数，且为：

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{(\mathrm{E}_{k_i}^x \mathbf{W}_{k,c}^Q)(\mathrm{E}_{k_j}^x \mathbf{W}_{k,c}^K)^{\mathrm{T}}}{\sqrt{d}}\right) \tag{6}$$

其中α_{ij} 是文本特征和文本特征的点积注意力矩阵，$\mathbf{W}_{k,c}^V$, $\mathbf{W}_{k,c}^Q$, $\mathbf{W}_{k,c}^K$ 是参数矩阵。

然后使用FFN 神经网络更新序列每个位置的状态，并得到F_{x_k}，如下所示：

$$\mathrm{F}_{x_k} = \mathrm{FFN}(\mathrm{H}_{x_k}^{l}) \tag{7}$$

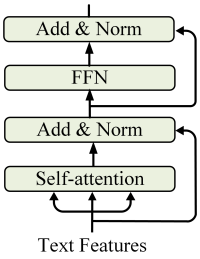

Figure 3: 传统多头Transformer 编码器模型

3.1.3 图像编码器层

图像由预训练的ResNet-101 模型抽取为$7 \times 7 \times 2048$ 维的特征矩阵，并把每个区域特征映射到与文本同一空间，将图像特征转化为一个49×128 维的特征矩阵。如下所示，

$$\mathrm{F}_{z_k} = \mathrm{MLP}(\mathrm{E}_k^z) \tag{8}$$

其中，MLP 是多层感知器。

3.2 跨模态语义交互模块

类似于Nishihara et al. (2020)，论文采用跨模态注意力机制实现文本语义和图像语义特征交互，将源文本语义表征作为查询矩阵，图像语义表征作为键/值矩阵，构建图文细粒度对齐语义表征向量，

$$\mathrm{H}_k = \text{Multihead-Im2te}(\mathrm{F}_{x_k}, \mathrm{F}_{z_k}, \mathrm{F}_{z_k}) \tag{9}$$

$$= \sum_{j=1}^{m} \widehat{\alpha}_{ij} (\mathrm{F}_{z_{k,j}} \mathbf{W}_1^V) \tag{10}$$

$$\widehat{\alpha}_{ij} = \mathrm{softmax}\left(\frac{(\mathrm{F}_{x_{k,i}} \mathbf{W}_2^Q)(\mathrm{F}_{z_{k,j}} \mathbf{W}_3^K)^{\mathrm{T}}}{\sqrt{d}} \right) \tag{11}$$

其中，Multihead-Im2te表示文本语义和图像语义的跨模态注意力机制，$\widehat{\alpha}_{ij}$是文本语义和图像语义的相似度权重，表示第i 个词与第j 个图像区域的相似度，$i \in (1, \cdots, n)$，m 是图像划分区域的个数，论文中m 为49。

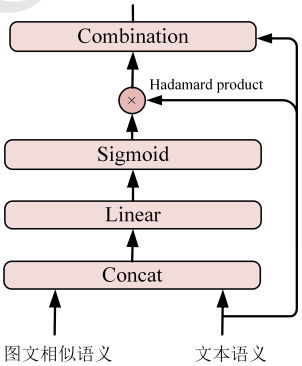

Figure 4: 多模态语义融合模块

3.3 多模态语义融合模块

为了充分的交互多模态语义信息，该文首先以图文细粒度对齐语义信息为枢纽，基于图文细粒度对齐语义信息来进一步交互文本模态和图像模态信息。为此，类似于Yin et al. (2020)，

本文采用门控策略来实现进一步的交互，如图4 所示，

$$\Omega = \text{Sigmoid}(\text{W}_\Omega(\text{H}_k \,||\, \text{F}_{x_k})) \tag{12}$$

$$\widehat{\text{H}}_k = \text{F}_{x_k} \otimes \Omega \tag{13}$$

其中，$\otimes$ 是哈达玛乘积（对应位置元素相乘），W_Ω是模型参数矩阵，|| 表示拼接操作，论文将图像和文本特征在最后一个维度进行拼接，$\widehat{\text{H}}_k$是有用的多模态信息。

然后采用相加的方式实现多模态特征的融合，具体为：

$$\text{O}_k = \widehat{\text{H}}_k + \text{F}_{x_k} \tag{14}$$

最终把编码器的输出O_k 输入到解码器进行解码。

3.4 目标语言解码器

类似地，$t_k = \{t_1^k, \ldots, t_s^k\}$ 表示源语言x_k 对应的目标句子序列，其中s 是t_k 的句子长度，目标句子表征为$\text{E}_k^t = \text{Emb}_t(t_k) + \text{PE}_t(t_k)$。如图2 右图所示，解码器采用传统的多头Transformer解码框架，每个解码器层由三个子层组成：1）掩码多头自注意层；2）跨语言多头注意层；3) 前馈网络层。

首先使用多头自注意力机制对目标句子特征进行提取，可以表示为，

$$\text{Q}_k = \text{Multihead}(\text{E}_k^t, \text{E}_k^t, \text{E}_k^t) \tag{15}$$

然后采用跨语言多头注意力机制实现图文多模态特征O_k 和目标序列特征Q_k 的交互，如下所示，

$$\text{Y}_k = \text{Cross-att}(\text{Q}_k, \text{O}_k, \text{O}_k) \tag{16}$$

然后使用FFN 神经网络更新序列每个位置的状态，并得到F_{d_k}，如下所示：

$$\text{F}_{d_k} = \text{FFN}(\text{Y}_k) \tag{17}$$

最后将解码器最后一层的输出作为softmax输入，通过softmax层预测目标句子的概率分布，可以表示为

$$\text{P} = \text{Softmax}(\text{W}_p\text{F}_{d_k} + b) \tag{18}$$

其中b和W_p是参数，F_{d_k}代表解码器最后一个隐藏状态的输出。

4 实验

数据集： 论文基于多模态神经机器翻译公共数据集Multi30K [-1] 基准数据集的英语→德语、英语→法语和英语→捷克语多模态翻译任务进行实验，其中训练、验证和测试集分别包含29k、1014 和1000 个文本图像对。每张图像都包含一个英文描述句子以及由专业翻译者翻译成的德语、法语和捷克语。论文采用四个测试集来评估提出的的多模态神经机器翻译模型，1) Test2016 测试集[0]，Multi30K 中划分的包含1,000 个示例图文句子对；2）Test2017测试集[1]，WMT2017中包含的1,000个测试图文句子对例子，包含更难翻译和理解的源句；3）我们还使用带有歧义COCO数据集作为域外测试数据，其中包含461个含歧义动词的图文句子对示例，并鼓励使用图像进行消歧；4）Test2018测试集[2]包含1,071个图文句子对实例，该测试集实体词多，低频词多。

数据预处理： 论文采用bpe分词对源语言、目标语言文本进行切分，bpe切割的粒度为6k，每个语言对的词表大小分别为英语→德语（En→De）的5,644→5,876，英语→法语（En→Fr）

[-1]https://github.com/multi30k/dataset
[0]https://www.statmt.org/wmt16/multimodal-task.html
[1]https://www.statmt.org/wmt17/multimodal-task.html
[2]https://www.statmt.org/wmt18/multimodal-task.html

的5,644→5,684，英语→捷克语（En→Cs）的5,644→5,972。采用Resnet-101模型对图像特征进行提取得到具有49 个局部空间区域特征的7x7x2048 维向量。
评估指标: 使用广泛用于评估机器翻译质量的BLEU和METEOR两个指标来评估翻译质量，1) 4-gram BLEU 指标(Papineni et al., 2002)，它在准确性和流畅度方面衡量翻译的质量，2) METEOR [3] 指标(Denkowski and Lavie, 2014)，它考虑了翻译质量的精度和召回率。

4.1 实验设置

论文基于Transformer (Vaswani et al., 2017) 搭建机器翻译框架，类似于Wu et al. (2021)，我们的模型堆叠4 层编码器-解码器。本文将编码器和解码器隐藏状态的维度设置为d_{model}=128，前馈网络的内层设置为d_{ffn}=256。学习率设置为0.005。max-tokens 设置为4096，warmup 更新步数设置为2000，标签平滑值设置为0.2。模型采用β_1, β_2 = (0.9, 0.98)的Adam 优化器。模型头数为4，并将dropout 设置为0.3 以避免过度拟合，beam size 设置为5。当BLEU 分数在验证集上的10 个epoch 内没有提高时，模型停止训练。我们采用单个GTX 3090 GPU 训练模型。

4.2 比较模型

为了直观验证本文提出的多模态神经机器翻译模型的优势，该文和以下最近最先进的多模态神经机器翻译模型进行比较，

- VAG-NMT (Zhou et al., 2018): 采用背景注意力机制来利用视觉信息增强模型翻译性能。
- DCCN (Lin et al., 2020): 提出了一种动态上下文引导胶囊网络（DCCN）来引导视觉特征提取以提高机器翻译性能。
- MNMT+SVA (Nishihara et al., 2020): 一种有监督的视觉注意机制，用于捕获与文本相关的视觉区域以进行机器翻译。
- OVC+L_v (Wang and Xiong, 2021): 构建了一个对象级的视觉上下文语义框架，以有效地探索和捕获视觉信息以指导机器翻译。
- WRA-guided (Zhao et al., 2021): 基于多模态Transformer，提出了一种词域对齐引导的方法来建立文本和视觉特征之间的语义相关性。
- IO-MMT (Song et al., 2021): 搭建了一个关系感知图编码器，以充分利用图像和源语句内部的关系，并在目标端提出一个有效的多模态奖励函数，以提高翻译视觉一致性。
- DLMulMix (Ye and Guo, 2022): 提出了一种新型双级交互式多模态混合编码器(DLMulMix)，提取有用的视觉特征来增强文本级机器翻译。

进一步的，为了更公平地证明本文提出的模型的优越性和有效性，在相同的参数设置和训练设备的基础上本文复现了三个最受欢迎的多模态融合方法，

- Gated Fusion MNMT (Li et al., 2021): 一种有效的多模态融合方法，通过增强文本中的重要信息来提升机器翻译性能。该方法广泛应用于多模态神经机器翻译和自然语言处理领域的其他多模态任务中。
- Multimodal self-att (Yao and Wan, 2020): 提出了一种图像感知多模态Transformer 模型来提取有用图像信息以提高机器翻译性能。该方法主要是将文本特征和视觉特征连接起来进行多模态交叉注意。
- Doubly-ATT (Arslan et al., 2018): 在解码器的源-目标交叉注意子层和自注意子层之间使用了一个额外的视觉注意子层，视觉诱发的注意力权重和源语言的注意力权重相加作为双重注意力权重。

[3]http://www.cs.cmu.edu/ alavie/METEOR/

Model	Multi30K En→De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
已有的多模态机器翻译模型						
VAG-NMT (Zhou et al., 2018)	-	-	31.6	52.2	28.3	48.0
DCCN (Lin et al., 2020)	39.7	56.8	31.0	49.9	26.7	45.7
MNMT+SVA (Nishihara et al., 2020)	39.9	58.1	-	-	-	-
OVC+L_v (Wang and Xiong, 2021)	-	-	32.4	52.3	28.6	48.0
WRA-guided (Zhao et al., 2021)	39.3	58.3	32.3	52.8	28.5	48.5
IO-MMT (Song et al., 2021)	41.3	59.2	33.5	52.8	-	-
DLMulMix (Ye and Guo, 2022)	41.77	58.93	33.07	51.85	29.90	49.09
基于Fairseq复现的翻译模型						
Transformer (NMT) (Vaswani et al., 2017)	40.96	58.35	32.59	51.21	29.16	48.37
Doubly-ATT (Arslan et al., 2018) †	41.44	59.08	33.15	52.34	29.22	48.41
Multimodal self-att (Yao and Wan, 2020) †	41.50	58.52	32.51	51.33	29.10	48.48
Gated Fusion MNMT (Li et al., 2021) †	41.58	58.88	33.01	51.90	30.04	48.95
Our model	**42.37**	**59.67**	**34.78**	**54.06**	**31.02**	**50.64**

Table 1: Multi30k 英语→德语（En→De）翻译任务在BLEU 和METEOR 指标上的比较结果。† 表示基于我们的Transformer 模型复现以前的多模态融合方法。最佳结果以粗体突出显示。Transformer (NMT) 表示使用纯文本数据进行机器翻译。

4.3 在英语→德语多模态翻译任务上的实验结果

英语→德语多模态翻译任务的实验结果如下表1 所示。本文从三个方面对现有模型进行总结和比较:

1) 与现有的多模态机器翻译模型比较: 实验结果表明，本文提出的模型优于现有的多模态翻译模型，并且在大多数测试集上BLEU 和METEOR 评估指标提高了1∼2 个点。并且，该文提出的模型只需少量参数即可获得出色的结果。根本原因是本文提出的方法可以有效地交互细粒度的多模态语义信息，而现有模型在进行多模态融合时只是简单的整合多模态信息。

2) 与纯文本机器翻译比较: 本文提出的多模态翻译模型在BLEU 和METEOR 指标上显著优于纯文本机器翻译基线，并在所有测试集上提高了大约2 个点。这表明本文提出的多模态机器翻译模型可以有效利用图像信息来增强机器翻译。

3) 与复现的多模态方法比较: 为了更公平地比较本文提出的模型的有效性，基于相同的训练环境，本文复现了最近的三种多模态融合方法。结果被展现在表1，本文的方法在所有评估指标上都比最近的多模态融合方法有了显着改进，这表明深度的交互视觉细粒度语义信息有助于提高翻译性能。

4.4 在英语→法语多模态翻译任务上的实验结果

为了探索所提出模型的稳健性，该文在英语→法语多模态翻译任务上进行实验，结果如表2所示。与英语→德语任务类似，该文提出的模型在英语→法语任务上与现有的多模态翻译模型、纯文本翻译模型和复现的多模态融合方法进行比较，得出以下有趣的结论:

首先，与现有模型相比，本文提出的模型在两个评价指标上仍然取得了显著的提高，这与英语→德语翻译任务的结果是一致的。另外，与纯文本机器翻译基线模型相比，具有图像信息的多模态机器翻译模型取得了优异的结果，这表明本文提出的多模态翻译模型可以有效的与视觉信息交互以增强机器翻译。

其次，在英语→法语任务上复现近期有竞争的多模态融合方法，结果表明本文提出的方法优于复现的多模态融合方法。相比较于现有的多模态翻译模型，本文的模型取得了较好的翻译结果。英语→法语翻译任务的结果再次证明了所提出方法的有效性和普遍性。

4.5 消融实验

为了进一步验证本文提出方法的有效性，我们移除了模型的不同组件，以进行消融研究，结果被报道在表3。分析英语→德语和英语→法语翻译结果可以总结为两点：1）移除图文细粒

Model	Multi30K En→Fr					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
已有的多模态机器翻译模型						
VAG-NMT (Zhou et al., 2018)	-	-	53.8	70.3	45.0	64.7
DCCN (Lin et al., 2020)	61.2	76.4	54.3	70.3	45.4	65.0
OVC+L_v (Wang and Xiong, 2021)	-	-	54.2	70.5	45.2	64.6
WRA-guided (Zhao et al., 2021)	61.8	76.3	54.1	70.6	43.4	63.8
IO-MMT (Song et al., 2021)	62.5	76.9	54.9	71.7	-	-
DLMulMix (Ye and Guo, 2022)	62.23	76.85	55.18	**73.37**	44.42	66.41
基于Fairseq复现的翻译模型						
Transformer (NMT) (Vaswani et al., 2017)	60.33	75.64	53.45	71.57	43.61	65.72
Doubly-ATT (Arslan et al., 2018) †	60.94	75.99	53.63	71.56	44.78	65.35
Multimodal self-att (Yao and Wan, 2020) †	61.44	75.77	54.56	71.62	44.59	65.08
Gated Fusion MNMT (Li et al., 2021) †	61.24	76.26	54.15	71.77	44.29	64.91
Our model	**62.73**	**77.34**	**55.56**	73.14	**46.59**	**67.68**

Table 2: Multi30k 数据集上英语→法语（En→Fr）翻译任务的比较结果。

Model	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
En→De 多模态翻译任务						
MNMT_*gat*	41.32	59.16	34.22	53.89	29.69	49.91
MNMT_*att*	41.65	59.36	34.27	53.74	30.64	50.21
MNMT	42.37	59.67	34.78	54.06	31.02	50.64
En→Fr 多模态翻译任务						
MNMT_*gat*	62.46	77.12	55.48	73.04	45.65	67.06
MNMT_*att*	61.98	76.86	54.86	72.57	45.59	66.83
MNMT	62.73	77.34	55.56	73.14	46.59	67.68

Table 3: 模型不同组件的消融实验，MNMT_*att*是移除跨模态语义交互模块的模型，MNMT_*gat*是移除图文细粒度对齐语义与文本语义融合模块的模型，MNMT是本文完整的模型。

度对齐语义与文本语义融合模块，相比于完整的模型，在三个测试集上的两个评估指标都有所下降，这表明细粒度的图文语义交互可以为文本机器翻译提供更精确的信息，特别在MSCOCO测试集上模型的性能有重大的衰退，这表明该模块能有效的利用细粒度图像信息来解决翻译中的歧义性单词。2）移除跨模态图文语义交互模块，相比于完整的模型，模型整体翻译效果都有所下降，特别是在英语→法语多模态翻译任务上BLEU 和METEOR 评估指标都下降超过0.5个分数，这表明，跨模态交互图文语义建立图文细粒度对齐语义信息有助于指导图文多模态对齐融合，提升机器翻译的性能。通过以上的消融实验对比分析，验证了本文模型不同组件的有效性。

4.6 在英语→捷克语多模态翻译任务上的实验结果

为了进一步验证本文方法的有效性和鲁棒性，我们在英语→捷克语多模态翻译任务上评估模型。表4 给出了复现的多模态融合方法和本文的多模态融合方法的BLEU值和METEOR值。可以看到，本文的方法取得了最好的效果，在基线模型（NMT）的基础上取得了+2.01、+3.38的BLEU值提升及+1.11、+2.05 的METEOR值提升。相比于复现的方法，在两个评估指标上本文的方法取得了超过+1 点的BLEU值和METEOR值提升，翻译结果显著提升。这证明了本文方法对于不同语言对是有效且通用的。

4.7 翻译实例分析

为了验证本文方法在翻译过程中确实有效地指导了目标序列的生成，我们通过一些具体

En→Cs				
	Test2016		Test2018	
Model	BLEU	METEOR	BLEU	METEOR
Transformer (NMT)	32.70	32.34	27.62	29.03
Doubly-ATT (Arslan et al., 2018) †	33.25	32.28	29.12	29.87
Multimodal self-att (Yao and Wan, 2020) †	33.12	32.01	28.75	29.51
Gated Fusion MNMT (Li et al., 2021) †	33.77	32.24	29.43	29.41
Our model	**34.71**	**33.45**	**31.00**	**31.08**

Table 4: 实验结果在英语→捷克语（En→Cs）多模态翻译任务。


Src： a man *urinating* on a street corner .
MNMT_att： ein mann *urcht* an einer straßenecke .
MNMT_gat： ein mann *uriert* an einer straßenecke .
MNMT： ein mann *uriniert* an einer straßenecke .
Tgt： ein mann *uriniert* an einer straßenecke .


Src： a *bicycle rider* is going down *a long stair* way .
MNMT_att： ein *radfahrer* fährt *eine lange treppe* hinunter .
MNMT_gat： ein *fahrradfahrer* fährt *einen langen treppenweg* hinunter .
MNMT： ein *fahrradfahrer* fährt *eine lange treppe* hinunter .
Tgt： ein *fahrradfahrer* fährt *eine lange treppe* hinunter .


Src： two *men dressed in green* are preparing food in a restaurant .
MNMT_att： zwei *männer in grüner kleidung* bereiten in einem restaurant essen zu .
MNMT_gat： zwei *männer in grün* bereiten in einem restaurant essen zu .
MNMT： zwei *grün gekleidete männer* bereiten in einem restaurant essen zu .
Tgt： zwei *grün gekleidete männer* bereiten in einem restaurant essen zu .


Src： a bride and groom stand together with a *bouquet* in the sunlight .
MNMT_att： eine braut und ein bräutigam stehen zusammen mit einem *strauß* im sonnenlicht .
MNMT_gat： eine braut und ein bräutigam stehen zusammen mit einem *blumenstrauß* im sonnenlicht .
MNMT： eine braut und ein bräutigam stehen zusammen mit einem *bukett* im sonnenlicht .
Tgt： eine braut und ein bräutigam stehen zusammen mit einem *bukett* im sonnenlicht .

Figure 5: En→De 测试集翻译实例，下划线标记表示提升的翻译.

的翻译实例对本文方法的有效性进行验证。图5 是本文选取的一些英语→德语测试集翻译实例。从图中可以看出，本文提出的完整翻译模型（MNMT）翻译效果最好、翻译结果与参考答案基本对齐，有效的提升了预测句子的质量。例如，第一个翻译实例中，本文方法成功的翻译“urinating”到“uriniert”，而缺乏细粒度语义交互的两个模型翻译错误。第二个翻译实例中，MNMT模型准确翻译源语言句，MNMT_*att* 模型翻译文本实体“bicycle rider”失败，验证了跨模态图文语义交互帮助对齐文本实体与图像对象，MNMT_*gat* 模型翻译“a long stair”有轻微错误，验证了图文细粒度对齐语义与文本语义交互帮助对齐图文细粒度信息。第三个翻译实例中，MNMT 模型翻译结果与参考答案相同，MNMT_*att* 和MNMT_*gat* 模型正确翻译单词“men”、“green”，但翻译语序有误，没有对齐目标序列。第四个翻译实例中，MNMT 模型正确翻译“bouquet”为“bukett”，而没有充分语义交互的MNMT_*att* 和MNMT_*gat*模型翻译错误该词。上述实例分析表明，本文方法通过交互文本语义和图像语义细粒度信息，可以显著提高翻译的译文质量。

5 总结

本文提出了一种新颖的基于图文细粒度对齐语义引导的多模态神经机器翻译方法，该方法首先跨模态交互图文信息，以提取图文细粒度对齐语义信息，然后以图文细粒度对齐语义信息为枢纽，采用门控机制将多模态细粒度信息对齐到文本信息上，实现图文多模态特征融合。三个基准多模态翻译任务的实验结果证明本文提出的方法的有效性和优越性，并在三个基准任务上取得了强有竞争力的结果。进一步的消融实验分析表明，本文提出的方法可以深度交互图文

多模态语义信息，提取有用的模态细粒度信息以提高机器翻译的性能。在未来的工作中，我们会探索从多模态神经机器翻译模型的解码器方面进行改进，进一步提升多模态神经机器翻译的性能。

参考文献

Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. 2018. Doubly attentive transformer machine translation. *arXiv:1807.11605*.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost Van de Weijer. 2016a. Does multimodality help human and machine for translation and image captioning? *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 627–633, Association for Computational Linguistics*.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016b. Multimodal attention for neural machine translation. *arXiv:1609.03976, http://arxiv.org/abs/1609.03976*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. 2017. Lium-cvc submissions for wmt17 multimodal translation task. *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7–8, 2017, pp. 432–439. doi:10.18653/v1/w17-4746*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017a. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada, July. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017b. Incorporating global visual features into attention-based neural machine translation. *In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp. 992–1003. doi:10.18653/v1/d17-1105*.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark, September. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online, July. Association for Computational Linguistics.

Soonmo Kwon, Byung-Hyun Go, and Jong-Hyeok Lee. 2020. A text-based visual context modulation neural model for multimodal machine translation. *Pattern Recognition Letters*, 136:212–218.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. *arXiv preprint arXiv:2203.09173*.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4304–4314.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Enhancing neural machine translation with dual-side multimodal awareness. *IEEE Transactions on Multimedia*.

Hiroki Takushima, Akihiro Tamura, Takashi Ninomiya, and Hideki Nakayama. 2019. Multimodal neural machine translation using cnn and transformer encoder. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 2–9.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online, August. Association for Computational Linguistics.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Junjie Ye and Junjun Guo. 2022. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Applied Intelligence*, pages 1–10.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3025–3035*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Region-attentive multimodal neural machine translation. *Neurocomputing*.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium, October-November. Association for Computational Linguistics.

李志峰, 张家硕, 洪宇, 尉桢楷, and 姚建民. 2020. 融合覆盖机制的多模态神经机器翻译. 中文信息学报, 34(3):12.