基于时空注意力机制的视频引导机器翻译方法

姜舟[1,2]，余正涛[1,2]，毛存礼[1,2]，郭军军[1,2]，高盛祥[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）

摘要：视频引导机器翻译是一种多模态机器翻译任务，其目标是通过视频和文本的结合产生高质量的文本翻译。但是之前的工作中，只基于视频中的时间结构选择相关片段引导机器翻译，所选片段中仍然存在大量与目标语言无关的信息。因此，在翻译过程中，视频中的时空结构依然没有得到充分利用，从而无法有效缓解机器翻译中细节缺失或翻译错误的问题。为了解决这一问题，本文提出了一种基于时空注意力（spatial-temporal attention, STA）的模型来充分利用视频中的时空信息引导机器翻译。本文提出注意力模型不但能选择与目标语言最相关的时空片段，而且能进一步聚焦片段中最相关的实体信息。所关注的实体信息能有效增强源语言和目标语言的语义对齐，从而使得源语言中的细节信息得到准确翻译。本文的方法基于 Vatex 公共数据集和构建的汉-越低资源数据集上进行实验，在 Vatex 与汉-越低资源数据集上 BLEU4 分别达到 32.66 和 18.46，相比于时间注意力基线方法的改进了 3.54 与 0.89 个 BLEU 值。

关键词：时空注意力；视频引导机器翻译；细节缺失；时间注意力；空间注意力

视频引导机器翻译是在给定一组视频和相关文档情况下，根据视频和语义的对应增强文档的翻译，通过视频线索解决歧义的问题。近年来，视频引导机器翻译在自然语言处理、计算机视觉等领域受到了很大的关注，因为它可以支撑更多的实际应用。例如，它能够克服互联网语言差异性较大的问题，帮助用户有效的理解视频；通过视频片段中展示的视觉场景和文本描述，为视障人士提供便利。

与图像引导机器翻译任务相比，视频引导机器翻译更具挑战性，因为视频是由连续的帧组成的，一个静态的视频剪辑持续 5 到 10 秒，包含 120 到 240 帧，其中包含了很多信息。以往的研究提出了多种不同模态的融合办法，但大多没有考虑到多种模态的相对重要性，在多模态任务中，不同模态通常不是同等重要的。例如在多模态机器翻译任务中，文本显然更重要一些，虽然视频中承载了更丰富的信息，但也包含了更多无关的内容，如果将视频特征直接进行编码，可能会引入大量噪声。因此，从提高翻译质量的角度看，提取出视频中与文本描述最相关的部分，可以更好的利用视频的时空语境对齐源语言和目标语言。因此，很多学者提出视觉注意机制[1-5]选择性地关注视频中的部分信息，但是大多基于时间注

基金项目：国家自然科学基金（61732005, 61761026, 61972186, 61672271, 61762056）；国家重点研发计划（Nos. 2019QY1802, 2019QY1801, 2019QY1800）；云南高科技人才项目（201606）；云南省重大科技专项（202002AD080001）；云南省基础研究计划（201901S070057，2018FB104）；昆明理工大学省级人培项目（KKSY201703005）

* **通信作者：**gaoshengxiang.yn@foxmail.com

意力机制的方法，在只使用粗略的帧级全局特征，造成的翻译效果不是很理想，且视频在单一帧中总是会出现多个突出的物体。尽管利用时间注意力对重要帧进行了选择性聚焦，但在每一帧中仍然很难注意到多个有意义的对象，这使得翻译的过程中的细节缺失的问题无法解决。以图 1 仅使用时间注意力方法为例，“horizontal bar”被错误的翻译为“水平酒吧”，实际是“单杠”的意思。

为了解决这项问题，本文首先利用 Faster R-CNN[6]检测并提取出局部特征，它可以根据物体的实际大小生成可变大小的包围盒，更准确的检测多个物体。其次，本文引入了一种时空注意力（STA）方法，不仅可以选择性地关注帧的特定子集，还可以关注该子集中的显著对象。最后通过最终得到视频特征作为视频输入对齐源语言和目标语言，从而增强翻译的质量。

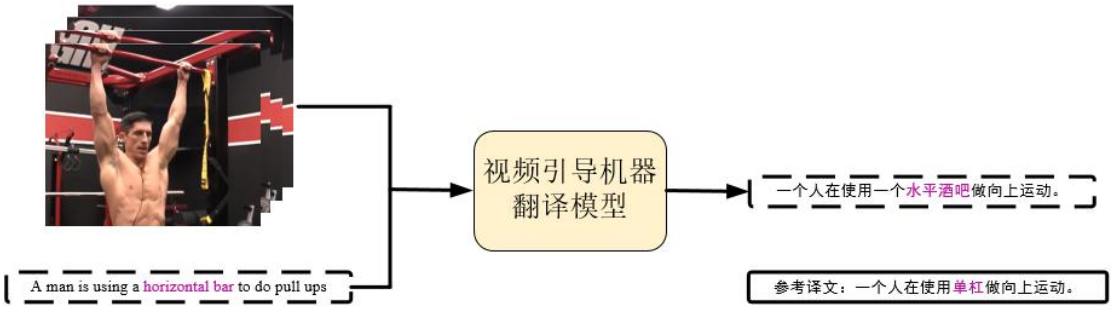

图 1 细节缺失的问题说明。

Fig. 1 Illustration of problems of detail missing.

综上所述，本文做出了以下贡献：

•本文研究了在视频引导机器翻译任务中，引入局部特征，提高了视频帧中对多个小目标的识别和定位。

•本文提出了一种用于视频引导机器翻译的时空注意力（STA）方法。通过对每一帧上的空间特征和连续帧上的时间特征分配不同的权重，本文的方法能够捕获并保留视频中的全局信息，从而解决了翻译过程中细节缺失的问题。

•在视频引导机器翻译公共数据集 Vatex[5]上进行的大量实验表明，本文的方法通过适当的将空间注意力机制整合到时间注意机制中中获得了显著的收益。

1. 相关工作

在处理视觉信息的工作中，目前的最佳做法是使用一个多层叠加的卷积神经网络（CNN）训练系统完成相关的计算机视觉任务，并使用从训练网络中提取的潜在特征作为视觉表示。因此，大多数图像/视频引导机器翻译方法依赖于从最先进的 CNN 中提取的特征[7-9]，对于视频和图像的处理方式总的可分为根据全局特征和局部特征分类任务。

利用全局特征的视频/图像引导机器翻译：Elliot 等人[10]第一个尝试做图像引导机器翻译任务，他们在一个编码器-解码器框架内将问题表述为从源语言模型到目标语言模型的语义迁移，而没有使用注意力机制。他们使用预先训练的 VGG 特征初始化源语言模型和目标语言模型两者的隐状态，随后将初始化变量输入到

带有注意力机制的神经机器翻译的模型中。Calixto 等人[11]使用了循环解码器解决模型中的解码问题。Ma 等人[12]使用了 ResNet 特性初始化编码器和解码器。Madhyastha 等人[13]将后验概率向量作为一种视觉表征，而不是 CNN 倒数第二层的聚集特征。Huang 等人[14]将特征向量投影到源语言的空间中，将特征嵌入到序列的开头和结尾，以视觉信息丰富源语言的句子表示。然而，这些方法无法完全适用于视频引导机器翻译任务，因为这种表示完全忽略了视频帧的顺序，没有利用任何时间结构。为了解决这一问题，Wang 等人[5]考虑对视频进行预处理过程中使用的原始动力学训练数据集上预先训练的 I3D 模型提取特征，利用全局时间结构，对 I3D 提取的特征进行时间注意力机制处理得到最终视频特征作为视频输入，使得译码器在同一时间有选择地只关注一小部分帧。本文的 SAT 方法与 Wang 等人[5]都着重使用注意机制对视频特征进行选择性聚焦，本文的工作与其有几个重要的区别，本文的模型不仅有选择性地注意帧的特定子集，而且关注这个子集中的特定对象。本文认为在视频处理过程中对目标进行检测比较重要，如果没有对目标进行检测，会产生对目标错误的判断导致错误的引导翻译。其次，在视频引导机器翻译过程中，有效利用重要的局部特征对翻译质量影响较大。第三，只使用局部特征而不使用全局特征，忽略了上下文信息。

利用局部特征的视频/图像引导机器翻译：视频包含的特征类型比图像多，但是在特征处理方式上，很多视频引导机器翻译工作主要使用帧级外观特征。Xu 等人[15]探索了两种基于注意力的方法，它们能够关注图像中最相关的区域，利用这部分最相关的区域对齐目标词。Shetty 等人[16]利用了预先训练的 SVM 分类器，并将这些特征集成到全局特征中，但是他们只是采用简单的平均策略处理这些局部特征，这种方法有忽略每个框架的空间结构的风险。Yu 等人[4]利用光流对每帧的局部特征进行粗略检测和提取，并将所有的局部特征集中在一起。但是，粗糙的检测容易导致对目标的错误判断。此外，他们只使用了忽略上下文信息的补丁功能。因此，本文利用预先训练的 Faster R-CNN[19]模型，可以更准确地检测到对象，并根据对象的实际大小生成可变大小的包围框。综上所述，本文提出了一种使用全局特征和局部特征两种类型外观特征的方法，本文的方法可以在保持全局上下文信息的同时捕获到的细节特征，通过拥有全局视频特征和局部特征的视频特征有效的对齐源语言和目标语言，从而增强翻译的质量。

2. 在视频引导机器翻译中利用时空注意力

注意机制在多模态任务中得到了广泛应用。在本节中，本文将对任务进行深入研究，并提出一种基于时空注意力的视频引导机器翻译方法。

2.1 总体框架

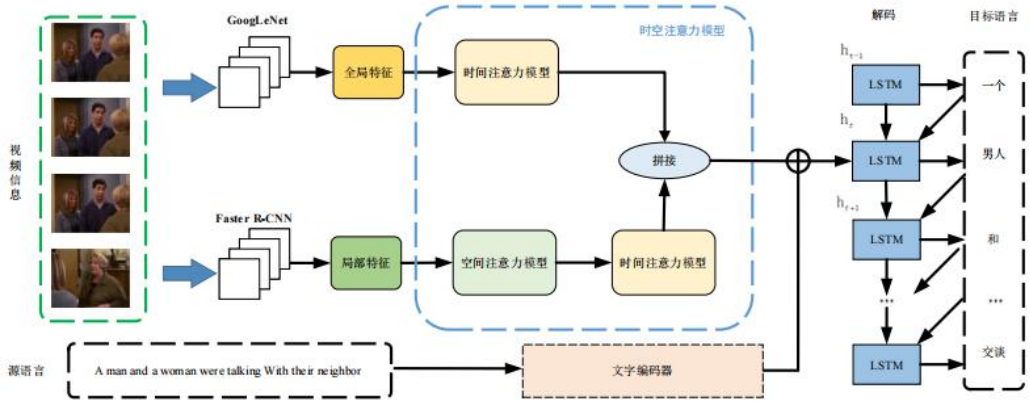

图 2 基于时空注意力(STA)的视频引导机器翻译模型图。

Fig. 2 Model diagram of video-guided machine translation with spatial-temporal attention (SAT).

本文基于 ConvNet +LSTM[3,15,17-25]架构构建了视频引导机器翻译框架，由源语言编码器，视频编码器以及目标语言解码器组成。如图 2 所示。编码器网络用于学习良好的视觉表示和源语言表示，解码器网络从编码器的输出生成相应的翻译。源语言编码器输入端为源语言单词序列 $S=\{s_1,s_2...,s_z\}$，视频编码器在编码器网络中，从视频帧中提取全局特征 f_g 和局部特征 f_l，所以一个视频就可以转换成 $F=\{f_1,f_2,...,f_k\}$，其中第 i 时刻的视频特征 f_i, $f_i=\{f_{gi},f_{li}\}$。在视频引导机器翻译中，本文将遵循 Wang 等人[5]的实现，在每个时间点，利用 soft-attention 模型从源语言句子中选择出关键词 $\varphi_t(S)$，利用提出的时间注意力模型从视频特征中选择出关键时空特征 $\varphi_t(V)$，并将二者输入到目标语言解码器中。其中，关键时空特征的选择过程我们将在 2.3 节中详细介绍。在解码器网络中，可根据拼接后的特征转换成目标语言序列 $Y=\{y_1,y_2,...,y_k\}$，公式如下：

$$h_t=o_t\odot c_t; \tag{1}$$

$$o_t=\sigma(W_oE[y_{t-1}]+U_oh_{t-1}+A_o\varphi_t(S)+B_o\varphi_t(V)+b_o); \tag{2}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot g_t; \tag{3}$$

$$i_t=\sigma(W_iE[y_{t-1}]+U_ih_{t-1}+A_i\varphi_t(S)+B_i\varphi_t(V)+b_i); \tag{4}$$

$$f_t=\sigma(W_fE[y_{t-1}]+U_fh_{t-1}+A_f\varphi_t(S)+B_f\varphi_t(V)+b_f); \tag{5}$$

$$g_t=\tanh(W_gE[y_{t-1}]+U_gh_{t-1}+A_g\varphi_t(S)+B_g\varphi_t(V)+b_g), \tag{6}$$

其中 h_{t-1} 为前一时刻的隐状态，σ为 *sigmoid* 激活函数，y_{t-1} 为前一个单词，E 是一个词嵌入矩阵，$E[y_{t-1}]$ 表示为前一个单词的词嵌入向量，

$W_o(W_i,W_f,W_g), U_o(U_i,U_f,U_g), A_o(A_i,A_f,A_g), B_o(B_i,B_f,B_g), b_o(b_i,b_f,b)$,依次为输入的权重矩阵、前一个隐状态、编码器的上下文和偏置。每当计算出新的隐藏状态，就可以使用单个隐藏层神经网络获得可能单词集合的概率分布：

$$p_t = soft\max(U_p \tanh(W_p[h_t,\varphi_t(S),\varphi_t(V),E[y_{t-1}]]+b_p)), \tag{7}$$

式中$[h_t,\varphi_t(S),\varphi_t(V),E[y_{t-1}]]$表示这四个向量的连接。

2.2 目标检测和局部特征提取

目前，目标检测在计算视觉领域受到了越来越多的关注[22,26-31]，并且对多个局部特征的提取是本文 STA 方法训练和测试中的关健组成部分。为了检测和定位视频帧上多个目标，Yu 等人[4]利用光流沿盒边框下方粗略检测和提取了 n 个大小为 220 × 220 的图像块。Donahue 等人[32]和 Rohrbach 等人[33]设计了一种能够准确探测和定位多个目标的专用手动探测器。然而本文发现这两种方法的工程量非常大。受最近区域建议网络（RPN）[6]和基于区域的卷积神经网络（R-CNNs）[34]在目标检测方面的成功启发，本文将利用 Faster R-CNN 模型从输入视频帧中直接检测多个目标。

本文使用的 Faste R-CNN 模型以一幅图像作为输入并输出一组矩阵的对象建议，每个对象建议都有一个类置信度评分，分数越高，就越有可能存在某个类的对象（如图 3 所示）。为了减少不必要的计算复杂度，本文首先减少每帧包围框数，从 300 个减少到 100 个，其次本文选择 28 个等间距帧检测可能的目标，进一步降低了计算复杂度，使得 Faster R-CNN 模型在 MS COCO 检测数据集上进行了预训练，可以检测到 80 个目标。与 Jeffrey[32]和 Yu[4]等人相比，Faster R-CNN 模型不仅可以更准确地检测多个目标，而且大大减少了检测时间，更重要的是它能够生成可变大小的包围框，这对于目标检测处理更加灵活。

检测每个视频帧上的对象后，根据其类置信度评分$\{c_1,c_2,...,c_n\}$,选择 top-n 个对象表示重要的局部对象（如图 4 所示）。然后，本文将对每个对象表示为一个 4096 维的局部特征，从 Faster R-CNN 网络的 fc7 层提取。最后得到一组局部特征$f_{li}=\{f_{li1},f_{li2},...,f_{lin}\}$，其中每一帧$f_{lj} \in \mathrm{R}^{4096}$。

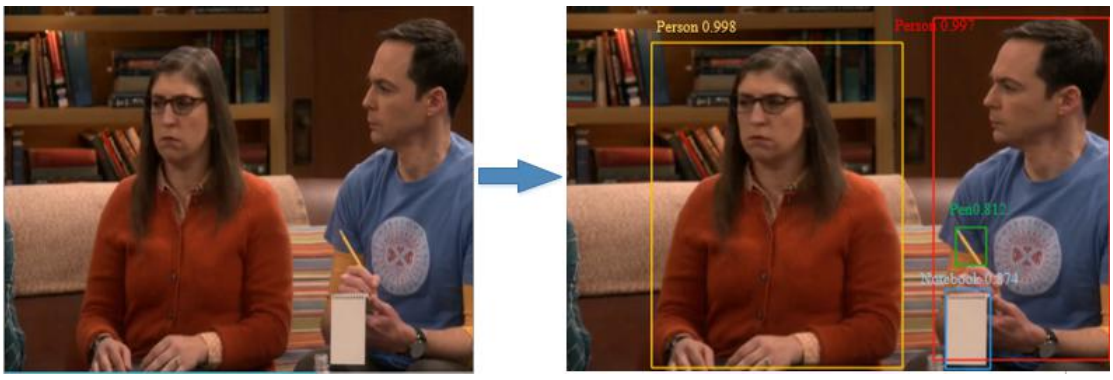

图 3 将输入的一帧视频使用 Faster R-CNN 进行目标检测示例。

Fig. 3 Enample detection using Faster R-CNN on a frame of a video.

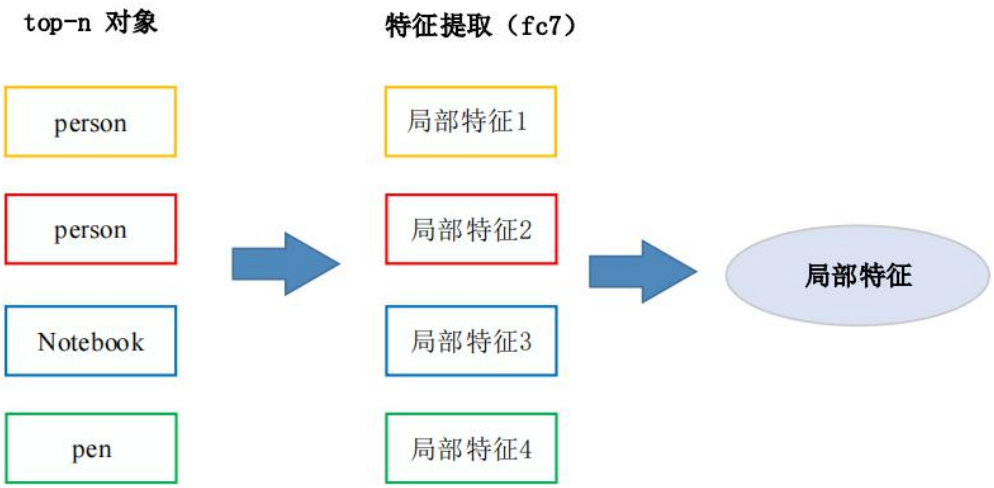

图 4 通过图 3 获得的目标检测的局部特征提取过程

Fig. 4 Local feature extraction process of target detection obtained from Fig.3.

2.3 时空注意力

在视频引导机器翻译任务中，视频中含有大量的信息，人们在观察视频的时候，视频会给人们一种视觉反馈，它会选择性地将视觉皮层早期阶段地表征映射到一个更中心的非地形表征，它只包含场景中物体特定区域地特性。因此，本文利用人类的这种视觉反馈特性，设计了一种时空注意力（STA）的方法。本文提出的方法使解码器首先关注视频帧中的特定对象，研究对象之间随时间的交互作用，在翻译过程中，根据提取视频中的特征对齐源语言和目标语言。

STA 方法的输入由全局特征、局部特征信息组成，输出是一个动态视觉表示，然后将它送入 LSTM 解码器每次迭代。首先，将局部特征f_{li}通过空间注意力机制选择出语义上更相关的局部特征$\psi_i^t(VL)$，其次，本文利用时间注意力分别从全局特征f_{gi}和局部特征$\psi_i(VL)$分别生成全局时间表示$\varphi_t(VG)$和局部时间表示$\varphi_t[\psi(VL)]$，最后全局时间表示和局部时间表示连接成新的视频时间表示$\varphi_t(V)$，进入 LSTM 解码器进行每次的迭代。

空间注意力机制：利用空间注意力机制对前 top-n 个局部特征$f_{li}=\{f_{li1},...,f_{lin}\}$得到得每个帧转化为局部特征$\psi(VL)=\{\psi_1(VL),\psi_2(VL),...,\psi_k(VL)\}$，$\psi_i(VL)$通过空间注意机制对 n 个局部特征进行动态加权求和：

$$\psi_i^{(t)}(VL)=\sum_{j=1}^{n}\alpha_{ij}^{(t)}vl_{ij}, \tag{8}$$

其中，$\alpha_{ij}^{(t)}$为 t 时刻的空间注意力得权重。空间注意力权重它反映了输入视频中第 j 个局部特征的相关性。因此，本文设计一个函数，以 LSTM 解码器的前一个隐状态和第 j 个局部特征作为输入，并返回相关性分数 $e_{ij}^{(t)}$：

$$e_{ij}^{(t)} = w_l^T \tanh(W_e h_{t-1} + U_e vl_{ij} + z_e), \tag{9}$$

其中w_l^T, W_e, U_e, z_e是模型要学习的参数，并且在所有时间步长上，所有局部特征所共享的参数。

当通过局部特征计算出所有 $e_{ij}^{(t)}$后$(j = 1,...,n)$，用 *softmax* 函数对它们进行归一化，得到 $\alpha_{ij}^{(t)}$：

$$\alpha_{ij}^{(t)} = \frac{\exp\{e_{ij}^{(t)}\}}{\sum_{j'=1}^{n} \exp\{e_{ij'}^{(t)}\}} \tag{10}$$

综上所述，空间注意机制可以让解码器通过增加相应局部特征的注意权值选择性地关注更显著的目标。

时间注意力机制：本文对全局特征 $V[G] = \{v[g]_1, v[g]_2, ..., v[g]_k\}$ 和局部特征 $\psi(VL) = \{\psi_1(VL), \psi_2(VL), ..., \psi_k(VL)\}$ 进行编码，编码后成为一个句子长度的时间表征 $\varphi(V) = \{\varphi_1(V), \varphi_2(V), ..., \varphi_m(V)\}$。每个时刻的 $\varphi_t(V)$ 的表示为全局时间表征和局部时间表征的级联：

$$\varphi_t(V) = \{\varphi_t(VG), \varphi_t[\psi(VL)]\}, \tag{11}$$

其中$\varphi_t(VG)$是所有 k 个全局特征的动态加权和，$\varphi_t(VG)$是通过时间注意力机制的所有 k 个局部特征的动态加权和：

$$\varphi_t(VG) = \sum_{i=1}^{k} \beta_i^{(t)} v[g]_i; \tag{12}$$

$$\varphi_t[\psi(VL)] = \sum_{i=1}^{k} \gamma_i^{(t)} \psi_i(VL), \tag{13}$$

其中 $\sum_{i=1}^{k} \beta_i^{(t)} = 1$，$\sum_{i=1}^{k} \gamma_i^{(t)} = 1$，在 LSTM 解码器的每个时间步长 t 上，分别计算 $\beta_i^{(t)}$ 和

$\gamma_i^{(t)}$，并且将 $\beta_i^{(t)}$ 和 $\gamma_i^{(t)}$ 作为 t 时刻的时间注意力权值。

本文设计了两个时间注意函数计算非标准化相关性得分 $b_i^{(t)}$ 和 $c_i^{(t)}$，将前一个隐状态、第 i 个全局特征和第 i 个局部特征作为输入：

$$b_i^{(t)} = w_k^T \tanh(W_b h_{t-1} + U_b v[g]_i + z_b); \tag{14}$$

$$c_i^{(t)} = w_r^T \tanh(W_c h_{t-1} + U_c \psi_i(VL) + z_c), \tag{15}$$

其中 $w_k^T, W_b, U_b, z_b, w_r^T, W_c, U_c, z_c$ 是全局特征和局部特征的共享参数。然后，通过 *softmax* 函数对上式进行归一化：

$$\beta_i^{(t)} = \frac{\exp\{b_{i'}^{(t)}\}}{\sum_{i'=1}^{k} \exp\{b_{i'}^{(t)}\}} \tag{16}$$

$$\gamma_i^{(t)} = \frac{\exp\{c_i^{(t)}\}}{\sum_{i'=1}^{k} \exp\{c_{i'}^{(t)}\}} \tag{17}$$

因此时间注意力机制允许解码器通过增加相应的全局特征和局部特征的注意力权值选择性的关注帧的子集。

综上所述，本文提出的两种注意力机制被有序地整合到一个编码器-解码器的视频引导机器翻译模型中，能够更加关注如何更准确地预测重要的目标，同时关注语义上更相关地视频帧，将目标与视频帧的结合作为堆区源语言和目标语言的视频模态输入到视频引导机器翻译模型中。

3 实验

3.1 数据集

本文在公共数据集 Vatex 和实验室内部的汉-越视频翻译数据集上进行实验。Vatex 每个视频片段约 10 秒并每个视频提供了 5 个中文和 5 个英文的平行句子，掉质量较低的一部分样本后，发布版本数据集总共包含 206345 个翻译对。基于本文实验室的汉-越低资源的视频翻译数据集中，从汉越新闻网和 Youtube 微博等共收集了 10500 个视频片段，每个视频片段约为 10 秒左右并配有 5 个视频描述的汉越平行句对，其中测试集有 2000 个视频片段。

3.2 实验细节

特征提取：对于全局特征，本文采用卷积层为 1024 维 pool5/7×7_s1 层，表示为 $VG = \{vg_1, vg_2, ..., vg_k\}$。对于局部特征，本文将表示 $VL = \{vl_1, vl_2, .., vl_k\}$。这些

局部特征由 Faster R-CNN 提取，在实验中，为了减少计算量和内存消耗，每帧视频提取特征个数上限设为 5，因为每帧视频中包含对象个数通常小于 10 个。

模型和训练：本文的视频引导机器翻译模型如图 2 所示。本文使用单层 LSTM 单元，隐藏层大小为 1024。词嵌入大小设置为 512，学习率设置为 0.0001，在训练过程中，所有视频引导机器翻译模型会通过最小化负对数似然估计进行端到端训练。然后，使用 Adadelta 算法和反向传播算法计算梯度，它们都广泛用于优化注意模型的参数更新。最后通过最大化对数似然估计参数：

$$L(\theta)=\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{t_n}\log p(y_i^n \mid y_{<i}^n, x_s^n, x_v^n, \theta), \tag{18}$$

N 个源语言句子、视频、目标语言句子训练对 (x_s^n, x_v^n, y^n)，其中，x_s^n 代表输入源语言句子，x_v^n 代表对应的视频，并且每个描述 y^n 的单词长度是 t_n。选取 Bleu-4 作为大多数机器翻译实验评价指标，本文实验也将用它作为衡量实验的的参考标准。

3.3 实验结果与分析

为了验证模型的有效性，本文考虑了以下三个基线进行比较：(1)Base NMT 模型：本文只考虑机器翻译的文本信息，采用 LSTM 解码器模型。(2)带有全局视频特征和时间注意力的模型结构，无局部特征的方法(TA-NL)。（3）与带有时间注意力方法的全局视频特征和使用平均策略的局部特征模型方法(NTA)比较。

表 1　STA 模型对比实验

Tab.1 STA model contrast experiment

Model	英文→中文	中文→英文
Base NMT	27.67	24.14
TA-NL(G+Temproal Attention)	28.89	24.36
NTA(G+Temproal Attention+fc7+average)	29.62	24.71
TAT（G+fc7+Temproal Attention）	31.53	24.68
VMT	31.1	24.6
STA(summation)	29.37	24.53
STA(concat)	32.84	24.81

说明：G 为 GoogLeNet,fc7 为 Faster R-CNN fc7 层提取特征，Average 为每 10 帧提取一帧的平均策略。

表 1 为 Vatex 数据集的验证集对比实验，在英文到中文的语料中，本文通过大量实验得出了 STA 算法获得了实验中最高的 BLEU 值，在英文到中文的语料中，得出了在与 Base NMT 模型相比，本文 STA 方法有了大幅度提高。与 TA-NL 相比本文的方法获得了 3.95 个 BLEU 值的提升，通过结果表明，本文将局部特征融入到全局特征中确实提高了视频帧中多个小目标的识别和定位。相比较于 NTA 方法，STA 方法获得了 3.22 个 BLEU 值的提升。通过两组实验结果表明，模型增加局部特征，是可以为结果带来改善。与 VMT 方法相比，本文的 STA 方法获得了 1.74 个 BLEU 值得提升。与 TAT 的方法相比本文的方法获得了 1.31 个 BLEU 值得提升，通过结果表明时间注意力难以区分视频帧上的小对象。因此，空间注意力是视频引导机器翻译方法的重要组成部分。本文也通过全局时间表征和局部时间表征的两个特征进行求和与拼接的方式进行了实验，发现，拼接后的效果明显好于求和后的效果。本文观察到，利用空间和时间信息带来得改善是互补的，当空间注意力机制和时间注意力机制同时使用时效果最好。在中文翻译到英文得实验中，同样获得了提升，但是在增加局部特征后效果没有英文翻译到中文实验中那样明显，本文猜测，可能与局部特征在 MSCOCO 数据集上预训练有关。

表 2 Vatex 公共测试集实验

Tab.2 Vatex public test set experiment

Model	英文→中文	中文→英文
VMT	29.12	26.42
STA	32.66	-

表 3 汉越语料对比实验

Tab.3 Chinese-Vietnamese data det contrast experiment

Model	汉→越
Base NMT	17.32
VMT	17.57
STA	18.46

表 2 是在 Vatex 公共测试集上获得了 32.66 个 BLEU 得成绩，说明了本文得实验有着很好得有效性和可靠性。表 3 是基于本文实验室汉-越低资源数据集中进行的实验，在 10500 个视频于 52500 对汉越平行句的训练下，效果也有着明显的提升。说明了本文的模型在低资源环境下，仍有着不错的成绩。

4 结论

本文从细节缺失的角度对视频引导机器翻译中存在的问题进行了深入的研究。在机器翻译中将视频作为引导是多模态机器翻译中的一种重要的翻译方法，任务根据识别每一帧上具有空间结构和连续帧上的时间结构的特点，本文提出了

一种新的视频引导机器翻译的方法，该方法基于时空注意力机制，将局部目标信息集成到全局信息中。与现有的方法相比，本文的方法可以关注多个突出的对象，从而产生详细准确的翻译描述。在 Vatex 公共数据集上进行了大量实验表明，时空注意力机制在视频引导机器翻译中是有效的，同样在资源稀缺的小语种数据集实验中，也表明了时空注意力在对资源较少视频引导机器翻译的任务中效果仍有提升。未来，本文将探索不同词性的语言与视频中对象的关系，准备在单词和连续的帧上搭建一个桥梁。

参考文献：

[1] HORI C, HORI T, LEE T Y, et al. Attention-Based Multimodal Fusion for Video Description. arXiv preprint,2017,arXiv:1701.03126.

[2] PAN P B, XU Z W, YANG Y, et al. Hierarchical recurrent neural encoder for video representation with application to captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 1029–1038.

[3] YAO L, TORABI A, CHO K, BALLAS N, et al. Describing videos by exploiting temporal structure. In Proceedings of the IEEE international conference on computer vision, 2015, 4507–4515.

[4] YU H N, WANG J, HUANG Z H, et al. Video paragraph captioning using hierarchical recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 4584–4593.

[5] WANG X, WU J, CHEN J, et al. VATEX: a large-scale, high-quality multilingual dataset for video-and-language research. In ICCV, 2019,4581-4591.

[6] REN S Q, He K M, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. 2015, 91–99.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd international conference on learning representations (ICLR), Banff, 2015.

[8] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on machine learning (ICML), Lille,2015, pp 448–456.

[9] HE K, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, 2016, pp 770–778.

[10] ELLIOTT D, FRANK S, HASLER E, et al. Multi-language image description with neural sequence models. Computing research repository 2015,arXiv :1510.04709.

[11] CALIXTO I, ELLIOTT D, FRANK S, et al. Dcu-uva multimodal mt system report. In: Proceedings of the 1st conference on machine translation (WMT), Association for Computational Linguistics (ACL), Berlin,2015, pp 634–638.

[12] MA M, LI D, ZHAO K,et al. OSU multimodal machine translation system report. In: Proceedings of the 2nd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Copenhagen, 2017, pp 465–469.

[13] MADHYASTHA P S, WANG J, SPECIA L, et al. Sheffield MultiMT: using object posterior predictions for multimodal machine translation. In: Proceedings of the 2nd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Copenhagen, 2017, pp 470–476.

[14] HUANG P Y, LIU F, SHIANG S R, et al. Attention-based multimodal neural machine translation. In: Proceedings of the 1st conference on machine translation, Association for Computational Linguistics (ACL), Berlin, vol 2,2016 pp 639–645.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.. In ICML, Vol, 14,2016 77–81.

[16] SHETTY R and JORMA L. Video captioning with recurrent networks based on frame-and video-level features and visual content classification. arXiv preprint,2015, arXiv:1512.02949.

[17] LI L H, TANG S, DENG L X, et al. Image Caption with Global-Local Attention.2017 In AAAI.

[18] DAKSH V and SRINIVASARAGHAVAN G, et al. Human Trajectory Prediction using Spatially aware Deep Attention Models.2017, arXiv preprint arXiv:1705.09436.

[19] YOU Q Z, JIN H L,WANG Z W, et al. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016,4651–4659.

[20] YAN Y C, NI B B, YANG X K, et al. 2017. Predicting Human Interaction via Relative Attention Model. arXiv preprint,2017 arXiv:1705.09467.

[21] ZHANG X S, GAO K, ZHANG Y D, et al., Jintao Li, and Qi Tian. Task-Driven Dynamic Fusion: Reducing Ambiguity in Video Description. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

[22] YAN C G,XIE H T, YANG D B, et al. 2017. Supervised hash coding with deep neural network for environment perception of intelligent vehicles. IEEE Transactions on Intelligent Transportation Systems 2017.

[23] PAN Y W, MEI T, YAO T, et al. Jointly modeling embedding and translation to bridge video and language. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2016, 4594–4602.

[24] VIGNESH R, TANG K, MORI G, et al. Learning temporal embeddings for complex video analysis. In Proceedings of the IEEE International Conference on Computer Vision.2015, 4471–4479.

[25] SHU X B, TANG J H, QI G J, et al. Concurrence-Aware Long Short-Term Sub-Memories for Person-Person Action Recognition.2017, arXiv preprint arXiv:1706.00931.

[26] ZHANG D W, HAN J W, JIANG L, et al. Revealing event saliency in unconstrained video collection. IEEE Transactions on Image Processing 26, 4 ,2017, 1746–1758.

[27] ZHANG D W, HAN J W, LI C, et al. Detection of co-salient objects by looking deep and wide. International Journal of Computer Vision 2, 120 2016, 215–232.

[28] ZHANG X S, ZHANG H W, ZHANG Y D, et al. Deep fusion of multiple semantic cues for complex event recognition. IEEE Transactions on Image Processing 25, 3 2016, 1033–1046.

[29] YAN C G, XIE H T, LIU S,et al. Effective Uyghur language text detection in complex background images for traffic prompt identification. IEEE Transactions on Intelligent

Transportation Systems. 2017.

[30] YAN C G, ZHANG Y D, XU J Z, et al. A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. IEEE Signal Processing Letters 21, 5 2014, 573–576.

[31] YAN C G, ZHANG Y D, XU J Z, et al. Efficient parallel framework for HEVC motion estimation on many-core processors. IEEE Transactions on Circuits and Systems for Video Technology 24, 12 2014, 2077–2089.

[32] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition.2015,2625–2634.

[33] ROHRBACH A, ROHRBACH M, QIU W, et al. Coherent multi-sentence video description with variable level of detail. In German Conference on Pattern Recognition. Springer,2014,184–195.

[34] GIRSHICK R, DONAHUE J, DARREL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, 580–587.

Video-guided Machine Translation by Spatial-Temporal Attention

Abstract：Video-guided Machine Translation as one of multimodal neural machine translation tasks,which target is generating high-quality text translation by tangibly engaging both video and text. But most of the existing methods , only relevant segments were selected to guide the machine translation based on the temporal structure of video, which only have still a lot of information in the selected fragment that is not relevant to the target language. So in the process of translation, the spatial-temporal structure of the video is still underutilized to mitigate the lack of detail or translation errors in machine translation. In order to solve this problem, we propose a spatial-temporal attention (SAT) method to address such problems. We proposed the attention model can not only select the most relevant segment of time and space with the target language,but also further focus the mostrelevant entity information in the segment and automatically focus on the most relevant spatial-temporal segments given the sentence context, the entity information of concern can effectively enhance the semantic alignment between the source language and the target language, so that the details in the source language can be translated accurately. The method in this paper is based on the Vatex public dataset and the Chinese-Vietnamese low resource dataset in the laboratory. The BLEU4 values on Vatex and Chinese-Vietnamese resource datasets were 32.66 and 18.46,respecticely, which improved 3.54 and 0.89 BLEU values compared to the temporal attention baseline method.

Keywords：Spatial-Temporal Attention;Video-guided Machine Translation;Detail Missing;Temporal Attention;Spatial Attention