计算机科学
COMPUTER SCIENCE

针对视频语义描述模型的稀疏对抗样本攻击

邱江兴, 汤学明, 王天美, 王成, 崔永泉, 骆婷

引用本文

邱江兴, 汤学明, 王天美, 王成, 崔永泉, 骆婷. 针对视频语义描述模型的稀疏对抗样本攻击[J]. 计算机科学, 2023, 50(12): 330-336.

QIU Jiangxing, TANG Xueming, WANG Tianmei, WANG Chen, CUI Yongquan, LUO Ting. Sparse Adversarial Examples Attacking on Video Captioning Model [J]. Computer Science, 2023, 50(12): 330-336.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100068

针对视频语义描述模型的稀疏对抗样本攻击

邱江兴　汤学明　王天美　王　成　崔永泉　骆　婷
分布式系统安全湖北省重点实验室,湖北省大数据安全工程技术研究中心,网络空间安全学院,华中科技大学 武汉 430074
(jiangxingqiu@hust.edu.cn)

摘　要　在多模态深度学习领域,尽管有很多研究表明图像语义描述模型容易受到对抗样本的攻击,但是视频语义描述模型的鲁棒性并没有得到很多的关注。主要原因有两点:一是与图像语义描述模型相比,视频语义描述模型的输入是一个图像流,而不是单一的图像,如果对视频的每一帧进行扰动,那么整体的计算量将会很大;二是与视频识别模型相比,视频语义描述模型的输出不是一个单词,而是更复杂的语义描述。为了解决上述问题以及研究视频描述模型的鲁棒性,提出了一种针对视频语义描述模型的稀疏对抗样本攻击方法。首先,基于图像识别领域的显著性分析的原理,提出了一种评估视频中不同帧对模型输出贡献度的方法。在此基础上,选择关键帧施加扰动。其次,针对视频语义描述模型,设计了基于 L_2 范数的优化目标函数。在数据集 MSR-VTT 上的实验结果表明,所提方法在定向攻击上的成功率为 96.4%,相比随机选择视频帧,查询次数减少了 45%以上。上述结果验证了所提方法的有效性并揭示了视频语义描述模型的脆弱性。

关键词:多模态模型;视频语义描述模型;对抗样本攻击;图像显著性;关键帧选择
中图法分类号　TP391.41

Sparse Adversarial Examples Attacking on Video Captioning Model

QIU Jiangxing,TANG Xueming,WANG Tianmei,WANG Chen,CUI Yongquan and LUO Ting
Hubei Key Laboratory of Distributed System Security,Hubei Engineering Research Center on Big Data Security,School of Cyber Science and Engineering,Huazhong University of Science and Technology,Wuhan 430074,China

Abstract　Despite the fact that multi-modal deep learning such as image captioning model has been proved to be vulnerable to adversarial examples,the adversarial susceptibility in video caption generation is under-examined. There are two main reasons for this. On the one hand,the video captioning model input is a stream of images rather than a single picture in contrast to image captioning systems. The calculation would be enormous if we perturb each frame of a video. On the other hand,compared with the video recognition model,the output of the model is not a single word,but a more complex semantic description. To solve the above problems and study the robustness of video captioning model,this paper proposes a sparse adversarial attack method. Firstly,a method is proposed based on the idea derived from saliency maps in image object recognition model to verify the contribution of different frames to the video captioning model output and a L_2 norm based optimistic objective function suited for video caption models is designed. With a high success rate of 96.4% for the targeted attack and a reduction in queries of more than 45% compared to randomly selecting video frames,the evaluation on the MSR-VTT dataset demonstrates the effectiveness of our strategy as well as reveals the vulnerability of the video caption model.

Keywords　Multi-model,Video caption,Adversarial example,Saliency map,Keyframe select

1　引言

现如今,多模态深度学习已经在各个领域取得了显著的进展,如视听语音识别(AVSR)[1]、视觉问答(VQA)[2]、图像和视频的语义描述模型[3-5]。视频语义描述模型的目标是用一句话或者一段话表达视频的内容[6]。相比描述图片,准确地描述清楚视频的内容更为困难,因为要想得出视频中各个元素之间的关系以及事情发生的先后关系,时间变量是一个很重要的参数。针对这一特点,大多数的研究工作都将此类模型设计为编码器-解码器的架构,将视频分类模型和自然语言处理模型二者有机地结合在一起。换句话说,就是采用 CNN + RNN 双模型架构,其中 CNN 用于视频/图像特征提取,RNN 用于时序分析和文本生成[6]。

然而,最近的研究表明多模态深度学习模型很容易受到对抗样本定向攻击的影响。定向攻击指攻击者对正常的输入施加轻微的扰动,而这种扰动是人类肉眼无法识别的,会使得训练有素的模型生成攻击者指定的输出内容。Xu 等[7]指出 VQA 模型容易受到对抗样本攻击。Chen 等[8]提出了基于

到稿日期:2022-11-09　返修日期:2023-03-26
通信作者:汤学明(xmtang@hust.edu.cn)

L_2 范数的损失函数，也被称为 Show-and-Fool 算法，用于图像描述模型的定向攻击。Xu 等[9]研究了图像描述模型的部分定向攻击，他们形式化地将攻击问题转换为潜在变量的结构化输出学习问题。尽管如此，现有的工作大多集中在与图像相关的多模态深度学习上，较少涉及与视频相关的多模态模型的鲁棒性问题，原因之一就是针对视频描述模型生成对抗样本比较困难。视频描述模型的输入不再是单一的图像，而是一个图像序列。如果对视频的每一帧进行扰动，计算量会非常大。到目前为止，唯一的相关的工作[10]由于内存的限制，只选择了视频的前四帧进行扰动，这导致攻击具有局限性而且攻击效率不高。因此，视频描述模型对抗样本攻击的主要挑战在于如何选取对输出结果影响较大的关键帧以及如何设计目标优化函数来生成定向描述。

本文使用视频描述模型作为研究案例并评估其鲁棒性。为了解决上述问题，我们提出了一种新的稀疏对抗样本攻击算法——Select-and-Attack 算法。受图像识别领域的显著图(Saliency Map)[11]的启发，我们不需要对所有视频帧进行扰动，而是选取那些对结果影响比较大的关键帧，这也是方法名中“稀疏”的含义。生成对抗样本的过程分为两步。首先，根据 CNN 模型的评分函数(Score Function)计算每个像素点的权重，基于此计算每个视频帧的权重，并按照权重大小降序顺序选取视频帧。为了防止选取的视频帧过多或者过少，我们设定了一个边界值 b，被选取视频帧的权重总和不能超过该边界值，没有被选取的视频帧不会被施加扰动。然后，利用模型的反向传播，动态地调整损失函数，用较少的查询次数以及较小的扰动生成视频描述模型的对抗样本。在数据集 MSR-VTT[12]上的实验结果表明，对于定向攻击，本文方法的攻击成功率可以达到 96.4%，而基于 CNN 的攻击成功率仅为 23.9%。图 1 所示为两个使用本文提出的定向攻击方法生成的对抗样本示例。在视觉上人类很难分辨出施加扰动前后的图像，但对抗样本却能够轻易误导 S2VT 模型[13]生成攻击者指定的输出。整个对抗样本攻击的整体架构如图 2 所示。本文实验结果表明，与随机选取视频帧相比，基于显著图的方法去选取关键帧能够减少 45%的模型询问次数，同时验证了实验中语义视觉模型的弱点。最后还证明了本文方法在引入了注意力机制的模型上的攻击效果更好。

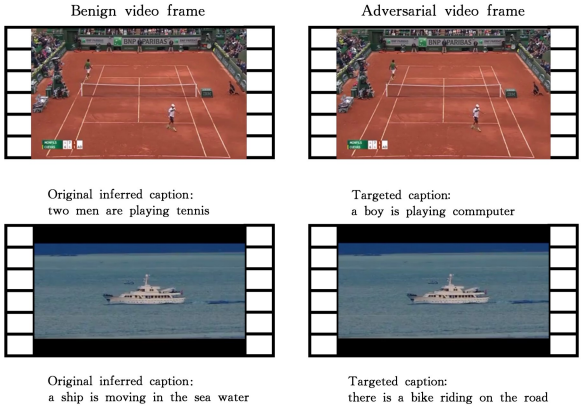

图 1 使用定向攻击生成的对抗样本

Fig. 1 Adversarial examples generated by targeted attack

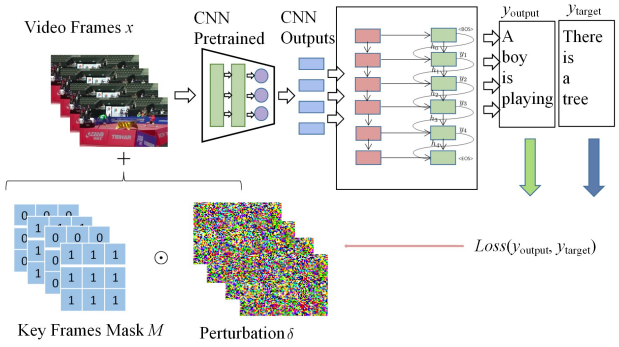

图 2 生成视频描述模型对抗样本的整体架构

Fig. 2 Overall architecture for generating adversarial examples for video captioning model

本文的主要贡献为以下 3 个方面：

(1)本团队是第一个针对视频描述模型创新性地提出稀疏对抗样本攻击方法的团队。基于显著图选取视频关键帧以及通过设计的损失函数，该方法实现了高成功率的攻击以及高效率的查询。最后在 MSR-VTT 数据集上的实验验证了该方法的有效性。

(2)实验证明了表现优异的视频描述模型很容易被基于 CNN+RNN 的攻击方法影响，但对仅仅基于 CNN 的攻击方法有一定的抵抗性。

(3)通过该攻击方法能够更深入地理解视频描述模型的内在机制。

2 相关工作

2.1 针对图像描述模型的对抗样本攻击

语义描述模型的框架主要由 CNN 和 RNN 组合而成。通常来说，图像分类模型仅仅需要 CNN，而描述模型生成语义描述还需要分析输出单词的时序特征。因此，在生成整个模型对抗样本时，计算其梯度是十分困难的。

之前大多数针对多模态模型的对抗样本攻击集中在图像描述模型。Xu 等[7]第一个研究了视觉语言模型的对抗样本，攻击模型分别是 DenseCap[14]和 VQA 模型。对于 VQA 模型，为了最大化目标描述的输出概率以及最小化在原图像的扰动，他们设计了一个包含 3 个部分的优化目标函数。然而，他们所设计的目标函数相当于把整个输出语句当作一个标签，这样的攻击无法细粒度地控制句子中每一个单词如何生成。为了实施高效的定向攻击，Chen 等[8]进一步提出了一个损失函数。该函数的目的是尽可能使得句子中每一个位置上出现目标单词的概率最大。他们将这个算法称为 Show-and-Fool 算法，该算法在定向攻击上具有较高的成功率。除此之外，他们还提出了定向关键词攻击，只要输出语句中包括目标关键词就算攻击成功。Xu 等[9]提出了更为严格的定向关键词攻击，目标关键词必须出现在指定的位置。最近，Aafaq 等[15]针对图像描述模型提出了一个基于 GAN 的对抗样本攻击方法。以上这些方法都使用目标优化函数来生成对抗样本，但很少涉及关于视频描述模型的鲁棒性问题。

2.2 针对视频分类模型的对抗样本攻击

Li 等[16]是最早利用 GAN 对视频识别模型进行对抗

样本攻击的。Wei 等[17]证实了在对视频识别模型攻击时不需要对视频每一帧施加扰动，对一些视频帧的扰动会传播到后续的视频帧，他们针对该模型提出了一种基于 $L_{2,1}$ 范数的递归优化算法。Chen 等[18]提出在视频的结尾加上一些冗余的视频帧，然后仅对这些新增的视频帧施加扰动。Jiang 等[19]首次针对视频分类模型提出了黑盒攻击方法。为了减少对目标模型的询问次数，他们选择从图像分类模型计算得出施加的扰动。最近，Zhang 等[20]和 Wang 等[21]首先提出了利用强化学习选取关键帧，从而实施基于稀疏时序的黑盒攻击。然而，以上方法对模型的询问次数相对较多，而且提出的攻击模型都是单模态，模型输出仅仅是单词，而不是一个句子。

3 针对视频语义描述模型的白盒攻击方法

将一个输入视频形式化表示为 $x \in X \subset R^{N\times W\times H\times C}$，其中 X 表示视频集合，N，W，H，C 别表示帧总数、帧宽度、帧高度、帧通道。对于每一个视频 X 来说，目标输出表示为 $y=\{y_1, y_2, \cdots, y_t, \cdots, y_n \in Y$，其中 Y 是输出集合，y_t 表示句子中第 t 个单词在词汇表 V 中的索引，y_1 是开始的标志，y_n 是结束的标志，n 是输出语句的总长度。本文使用 $F(x): X \rightarrow Y$ 表示视频描述模型，共有两种类型的攻击：非定向攻击和定向攻击。非定向攻击可以表示为 $F(x_{adv}) \neq y$，定向攻击可以表示为 $F(x_{adv})=y$。

3.1 关键帧选取

Wei 等[17]证明了视频帧之间的时序关系使得对一个视频帧的扰动会传播到其他帧上，因此生成对抗样本的关键在于找出那些对攻击结果影响最大的视频帧。

在图像和视频识别领域，显著图经常被用于计算各种输入模块对输出结果的贡献度，它由像素点核心区域与周围区域的区别计算得到。基于显著图在图像识别模型上的成功应用，我们提出了一种判断不同帧贡献度的方法。

为了建立每个视频帧的显著图，我们采用了 Simonyan 等[22]提出的方法。伪代码如算法 1 所示，该方法的核心是计算每个像素点的类评分函数梯度。计算得到的梯度越大，对该像素点就可以施加越小的扰动，而这个扰动却能对最终结果产生很大的影响。基于这一原理，可以通过下面的公式计算视频帧中每一个像素点的权重。

$$w=\frac{\partial S_c}{\partial I}\Big|_{I_{(i,j)}} \tag{1}$$

其中，S_c 表示 CNN 模型的类评分函数，$\boldsymbol{I}$ 是视频第 t 帧的向量形式。因此，对于视频第 t 帧来说，该帧总权重的计算方法为：

$$w_t^I=\frac{1}{H\cdot W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\frac{\partial S_c}{\partial \boldsymbol{I}}\Big|_{I_{(i,j)}} \tag{2}$$

然后根据权重大小降序选取关键视频帧。当第 t 帧被选择时，关键帧掩码(*Mask*)M 中第 t 帧的数值会置为 0，其中 $M \in {}^{N\times W\times H\times C}$。权重总和限制在边界值 $b \in (0,1]$ 之内，以防止选取的视频帧过多或者过少，没有被选取的视频帧不会被添加扰动。

算法 1 关键帧选取算法

输入：视频 $x \in X \subset \mathbf{R}^{N\times W\times H\times C}$，权重边界值 $b \in (0,1]$

输出：关键帧掩码 $M \in {}^{N\times W\times H\times C}$

```
1. 初始化：M 置为 0；权重总和 B 置为 0
2. for t ← 1 to T-th frame in x do
3.    w_t ← getFrameWeight(t,x);
4.    weights.add({t,w_t});
5. end for
6. sortReverseByWeight(weights)
7. for i ← 1 to N in weights do
8.    if(B < b) then
9.       B+=weights[i].w;
10.   setFrameMask(M,i,1);
11.   end if
12. end for
13. return M.
```

3.2 视频描述模型攻击

攻击方式设定为两类：非定向攻击和定向攻击。其中定向攻击的思路主要受到 C&W 算法和 Show-and-Fool 算法的启发。

非定向攻击：对于给定的输入视频 X，非定向攻击问题可以形式化地表示为以下目标优化函数：

$$\text{argmin}\ \|M\odot\delta\|_2 - c\cdot Loss(y, x+M\odot\delta) \tag{3}$$

其中，$\odot$ 表示哈德曼积；δ 表示对输入视频施加的扰动；$\|M\cdot\delta\|_2$ 表示原视频和扰动后视频之间的 L_2-范数距离；$Loss(\cdot)$ 表示模型损失函数；常数 c 是一个预先定义的正规化常数，通常，c 越大，生成对抗样本的成功率越高，但扰动也会随之增大。非定向攻击的目的是使原输出与模型输出之间的损失值尽可能大，从而使得模型往偏离原输出的方向更新参数，因此式(3)中使用了减法。

定向攻击：定向攻击伪代码如算法 2 所示。

算法 2 稀疏对抗样本定向攻击

输入：视频 $x \in X \subset \mathbf{R}^{N\times W\times H\times C}$，视频描述模型 F，目标语句 y_{target}，关键帧掩码 M，迭代数 I，正规化常数 c

输出：对抗样本 x_{adv}

```
1.  初始化：初始扰动矩阵 δ
2.  for iteration=1,2,…,I do
3.     y←F(x+δ⊙M)
4.     cost=Loss(y,y_target)
5.     if y==y_target then
6.        return x+δ;
7.     end if
8.     distance=getL2Distance(x+δ,x)
9.     cost=cost + c * distance
10.       cost.backward()
11. end for
```

对于给定的输入视频 X，定向攻击问题可以形式化地表示为以下目标优化函数：

$$\text{argmin}\ \|M\odot\boldsymbol{\delta}\|_2 + c\cdot Loss(y, x+M\odot\boldsymbol{\delta}) \tag{4}$$

定向攻击的目的是使得目标输出与模型输出之间的损失值尽可能小，因此式(4)中使用了加法，从而使得模型往目标

输出的方向更新参数。首先需要通过引入新变量的方式将有约束问题转换为无约束问题：

$$\boldsymbol{\delta}=\frac{1}{2}(\tanh(w)+1)-x \tag{5}$$

其中，无论 w 值怎么变，输入仍旧在限定在区间[0,1]中。然后需要设计新的损失函数使得目标输出在所有可能的输出中出现概率最大。

$$\log P(Y|x+\boldsymbol{\delta})=\max_{Y'\in\Omega}\log P(Y'|x+\boldsymbol{\delta}) \tag{6}$$

其中，Ω 表示输出集合。每个单词的概率 $p(y_t'|x+\boldsymbol{\delta},y_1',\cdots,y_{t-1}')$ 通过一个 LSTM Cell 计算出，用 $f(\cdot)$ 表示，输入是上一个隐藏状态 h_{t-1} 以及上一个输出单词 y_{t-1}。

$$z_t^{(y_t)}=f(h_{t-1},y'_{t-1}) \tag{7}$$

$$p(y_{t'}|x+\boldsymbol{\delta},y_1',\cdots,y'_{t-1})=\mathrm{softmax}(z_t^{(y_t)}) \tag{8}$$

经过 f 输出的 $\boldsymbol{z}$ 是未正规化概率的向量形式，再经过 softmax 函数，即可得到正规化概率。$z_t^{(y_t)}$ 是输出语句中位置 t 上的单词 y_t 出现的概率。和直觉相反，这里目标单词出现的概率不需要尽可能增大，只需要让它大于词汇表中其余单词出现的概率即可。换句话说，大于剩下单词表中出现的最大概率 $\max_{k\neq y_t} z_t^{(k)}$ 就足够了。综上所述，最终损失函数和优化目标函数可以转换为：

$$Loss(y_{\text{target}},x+M\odot\boldsymbol{\delta})=\sum_{t=2}^{N-1}\max\{-\varepsilon,\max_{k\neq y_t}\{z_t^{(k)}\}-z_t^{(y_t)}\} \tag{9}$$

$$\arg\min_{\boldsymbol{\delta}}\|M\odot\boldsymbol{\delta}\|_2+c\cdot\sum_{t=2}^{N-1}\max\{-\varepsilon,\max_{k\neq y_t}\{z_t^{(k)}\}-z_t^{(y_t)}\} \tag{10}$$

其中，用于控制 $\max_{k\neq y_t} z_t^{(k)}$ 和 $z_t^{(y_t)}$ 之间的差值，将损失函数的输出控制在一定的范围内。当差值大于该常数时，损失函数的输出就是该常数，从而函数梯度为 0。因此模型会更加关注句子中其他位置上概率不是很大的目标单词，由此可以实施更有效率的定向攻击。

4 实验

4.1 实现细节

我们使用 ADAM 优化器[23]解决第 3 章中提出的问题。设置学习率(learning rate)为 5×10^{-4}，扰动向量初始值设置为 1×10^{-4}。模型最大询问数设置为 1000。对于给定的视频和 CNN 模型，其计算得到的稀疏度是一定的。在边界值 b 设定为 0.4 时，平均稀疏度(SR)为 24.3%。所有的实验都在 NVIDIA TITAN RTX GPU 上进行。我们从 MSR-VTT 的测试集和验证集中随机选取了 1200 个视频，最后使用 METEOR 来评估预测结果和目标描述之间的相似度。

4.2 数据集

本文用来评估攻击效果的数据集是 MSR-VTT(Microsoft Research-Video to Text)，这是一个广泛用于视频描述模型的数据集。该数据集包含 10000 个视频片段以及相对应的 200000 个视频描述，其中训练集中有 6531 个视频，测试集中有 2990 个，验证集中有 497 个。选择该数据集的主要原因在于选择的威胁模型是基于 MSR-VTT 训练出来的。

4.3 威胁模型

本文实验使用预先训练良好的视频描述模型 S2VT(Sequence to Sequence：Video to Text)作为威胁模型。该模型使用 ResNet[24]作为 CNN 的部分，用于提取视频帧的视觉特征。CNN 的输出作为 LSTMs 的输入。LSTMs 是一个编码器-解码器架构，其作为模型的 RNN 部分完成视频帧的语义描述。架构如图 2 所示。

4.4 评估指标

实验中使用的评估指标为：METEOR 分数、完全匹配率、准确率、询问数以及稀疏度。

METEOR 分数：一个很受欢迎的机器翻译评估指标，主要用于计算模型输出语句与目标语句之间的相似度，而且其被证明比 BLEU@N[25]，ROUGE_L[26]和 CIDEr[27]的评估准确率更高。因为两个完全不同的语句也可能表达相同意思，为此我们设定一个阈值 t，当计算得到的分数大于 t 时，则认为这是一次成功的定向攻击。t 默认设置为 0.4。

完全匹配率(EMR)：完全匹配是最严格的评估标准，要求输出语句与目标语句完全一致。计算公式为 $EMR=\frac{EM}{T}$，其中 EM 表示完全匹配的总数量，T 是测试样例的总数。

准确率(A)：表示攻击成功的次数在所有测试样例中的占比。计算公式为 $A=\frac{S}{T}$，其中 S 是攻击成功的数量，T 是测试样例的总数。当计算得到的分数大于 t 时，则认为这是一次成功的定向攻击。

询问数(Q)：表示成功生成一个对抗样本需要询问模型的次数的平均值。设置最大询问数为 1000。

稀疏度(SR)：表示视频中未添加扰动的视频帧占比。计算公式为 $S=\frac{C}{N}$，其中 C 表示未被扰动的视频帧数量，N 表示一个视频的帧总数。

4.5 定向攻击

本节实验主要是为了证明视频描述模型的脆弱性以及本文攻击方案的有效性。与图像描述模型相比，攻击视频描述模型更具有挑战性。对于视频模型来说，施加扰动的视频帧越多，计算代价就越大。因此，需要提出一个只关注关键帧的对抗样本攻击方案。

我们选择对 S2VT 以及 S2VT-Attend 两个模型发起定向攻击。这两个模型的区别在于内在架构不同，S2VT-Attend 在框架中引入了注意力机制；其次，目标语句的选择不是任意单词的组合，因为模型的输出结果和训练集有强关联性。换句话说，模型只能生成数据集中存在的单词和句式。因此，我们随机从 MSR-VTT 的验证集中选择目标语句。

对于 S2VT 模型：我们首先评估参数 c 对攻击效果的影响。如表 1 所列，c 越大，扰动就越大，但攻击成功率也随之越来越高。询问次数整体呈递减趋势，当 $c=1000$ 时，询问次数小幅度上涨。当 $c=100$ 时，本文方法取得了最高的成功率(96.4%)，而且询问次数也最少。我们检查大多数攻击不成功的例子发现，即使语句在句式和单词上不太一样，但很多输出语句与目标语句表达的意思相近，因此 68.7%的完全匹配度是一个较高的准确度。以上实验结果充分说明了本文提出

的方法的有效性，且证明了 S2VT 模型的低鲁棒性。

表 1 对模型 S2VT 定向攻击的结果

Table 1 Results of targeted attack on S2VT model

Method	Metric	c				
		0.1	1	10	10^2	10^3
Our	Q	567.37	482.23	479.34	**390.45**	410.84
	A/%	89.4	90.3	94.6	**96.4**	93.2
	$\|\delta\|_2$	**3.8361**	5.4913	7.8924	8.8361	9.8248
	I/%	60.1	64.2	66.9	68.7	**69.8**
RandS	I	880.44	823.13	830.27	730.38	786.34
	I/%	83.2	84.8	85.2	89.3	**89.8**
	$\|\delta\|_2$	**5.2532**	6.8278	9.2482	9.9214	10.5730
	EMR/%	49.2	51.4	54.8	55.2	**57.2**

此外，为了验证关键帧提取方法的有效性，我们对每一组实验会另外添加一组对照组。对照组随机选取视频帧进行扰动，但两组视频的稀疏度保持一致，攻击算法也一致。从表 1 中可以看出，对于随机选择的视频帧，当 c 设置为 1000 时，成功率也高达 89.8%，这也印证了 Wei 等的实验结论。他们证明了随机对某个视频帧添加的扰动可以通过传播的方式影响到其他视频帧，且利用这一原理在视频识别模型上得到了较好的实验效果。但是由表 1 还可以看出，引入关键帧选取机制之后，本文方法提高了约 8%的成功率，减少了约 45%的查询次数。原因是我们选择扰动的视频帧的权重都相对较大，这意味着这些高权重视频帧对模型输出结果起着更为重要的作用。以上结果验证了本文方法的有效性。

对于 S2VT-Attend 模型：我们设置 c 为 100，根据上面实验得出的结论，c 为 100 时攻击效果较好，因此我们没有设置不同的 c。如表 2 所列，与攻击 S2VT 相比，模型引入注意力机制后，所有指标表现都更好。一个可能的原因是在每次对输入添加扰动后，S2VT-Attend 先提取其视觉特征，在对特征进行解码时，因为注意力机制放大了扰动对输出产生的影响，所以在 S2VT-Attend 上的攻击表现效果更好。

表 2 对模型 S2VT-Attend 定向攻击的结果

Table 2 Results of targeted attack on S2VT-Attend model

Method	S/%	Q	A/%	$\|\delta\|_2$	EMR/%
Our	24.3	432.6	96.8	5.4351	70.7
RandS	24.3	830.4	89.9	8.9214	46.5

对攻击失败例子的分析：可以将分析结果分为 3 类，图 3 所示为 3 类输出结果的样例。

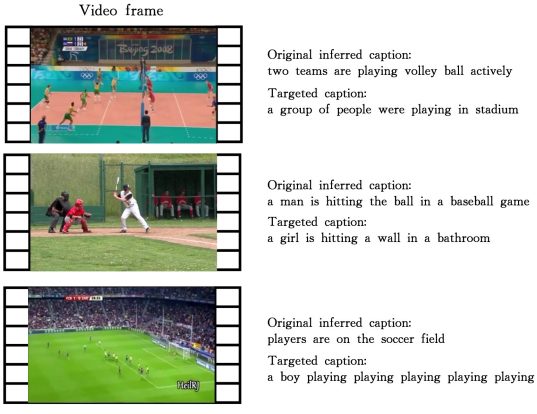

图 3 攻击失败样例

Fig. 3 Example of attack failure

第一类输出语句尽管 METEOR 评分低于设定的边界值，但表达的语义和目标语句高度相关并且有正确的语法结构。我们做了另一组实验，用其他评分算法（BELU-N 和 ROUGE）为这一类语句评分。实验结果如表 3 所列，阈值 t 设置为 0.4 时，大部分评分算法都能判定这是一次成功的定向攻击。因此，这一类结果是由评分算法的局限性导致的误判。第二类输出语句在语义上和目标语句有差距，其有正确的语法结构，但只包含目标语句的 1 个或多个关键词。第三类输出语句在语义上和目标语句不相同，而且不包含目标语句的关键词，并且语法结构不正确。第二和第三类结果是由攻击算法和视频描述模型的内在机制所导致的。本文提出的攻击算法中，关键帧是基于类评分函数的梯度选取的。然而视频的语义是时序和图像特征综合的结果。因此，关键帧必然不能关注到所有要素。关于视频描述的内在机制将在 4.8 节中进行更详细的讨论。

表 3 不同评估算法的评分结果

Table 3 Results of different evaluation methods

c	BELU-1	BELU-2	BELU-3	BELU-4	ROUGE	METEOR
10	0.46	0.59	0.37	0.29	0.48	0.26
10^2	0.72	0.55	0.48	0.38	0.65	0.35
10^3	0.56	0.67	0.52	0.40	0.58	0.38

4.6 性能对比

目前关于视频描述模型鲁棒性的研究较少，本团队是第一个对视频描述模型施加稀疏对抗样本攻击的。因此在本节中，我们将图像描述模型的对抗样本攻击施加到视频描述模型上。这里我们使用 Xu 等提出的方法 Densecap Attack (DA) 和 Xu 等提出的方法 Latent SSVM (LS) 作为对比方法。但由于这两种方法所设计的目标优化函数不一样，很多参数对结果都会产生影响，因此表中的数据并不能强有力地说明哪一种方法更好。单从表 4 的数据来看，本文方法的成功率要高 3%～5%左右，而 LS 的成功率比 DA 要高约 2%。这是因为 DA 方法的目标函数相当于把整个输出语句当作一个标签。这样的攻击无法细粒度地控制句子中每一个单词的生成。而 LS 方法是对某个位置上的单词出现概率做细粒度的控制，该方法使目标语句和同一位置的所有其他可能的部分语句之间的对数边际似然率最大化。这可以通过具有潜在变量的结构支持向量机 (SVM) 来优化。因此可以看出，对输出句子的控制粒度越高，攻击效果就越好。但即使损失不对输出单词进行细粒度概率控制，也能在很大程度上影响模型的输出结果，再一次证明了模型的低鲁棒性，也反应了攻击视频描述模型不是简单地将图像描述中的研究成果迁移到视频描述模型中。

表 4 DensecapAttack 和本文方法对比

Table 4 Comparison between DA and our method

Method	S/%	Q	A/%	$\|\delta\|_2$	EMR/%
Our	24.3	395.45	96.4	8.8361	68.7
DA	24.3	722.50	88.5	9.3872	65.3
LS	24.3	594.24	90.2	5.2824	66.5

4.7 仅基于 CNN 的定向攻击

以上实验验证了视频描述模型很容易受到基于 CNN+

RNN 的对抗样本攻击。本节主要是为了验证模型在仅基于 CNN 的攻击下的稳定性。本文方法通过设计目标优化函数，考虑到了模型中 RNN 的部分。而对于图像/视频分类任务来说，定向攻击只需要增加目标标签（一个单词）在整个标签集中出现的概率，因此仅仅考虑模型中 CNN 的部分，攻击多模态模型的效果不会非常理想。为了验证这一想法，将所提方法和 I-FGSM[28]，C&W[29]，L-BFGS[30] 和 Deepfool[31] 攻击算法进行对比。这几种算法被广泛应用于分类模型的对抗样本研究工作中[32-36]。我们随机从数据集中选取 500 个视频，目标描述也是随机选择的。为了消除关键帧选择方法对实验结果的影响，对于相同视频会使用相同的关键帧掩码。同时，为了使得 CNN 攻击方法的成功率更高，我们将 METEOR 评分的阈值设置为更低的水平 0.25。然而从表 5 的实验结果可以看出，仅基于 CNN 的攻击效果要远远差于本文方法。I-FGSM 的成功率只有 20.4%，平均查询次数为 970.9，C&W 的成功率为 27.3%，平均查询次数为 890.3。实验结果表明攻击多模态模型不仅要考虑特征提取模型即 CNN 部分，还要设计合适的损失函数针对语言处理模型即 RNN 部分。

表 5 仅基于 CNN 定向攻击的结果

Table 5 Results of targeted attack only based on CNN

Method	S/%	Q	A/%	$\|\delta\|_2$	EMR/%
I-FGSM	24.3	970.77	20.4	14.8361	8.7
C&W	24.3	890.30	27.3	3.5681	10.6
L-BFGS	24.3	920.53	22.4	8.4728	8.3
Deepfool	24.3	947.92	24.3	6.8872	9.2

4.8 视频描述模型的内在机制分析

本节基于对抗样本攻击的结果对视频描述模型的内在机制进行分析与讨论。从整体上来看，视频描述模型的架构如图 2 所示。首先用 CNN 模型逐一对所有视频帧进行编码，提取其向量特征，然后将输出结果作为 RNN 模型的输入，逐个解码单词的生成句子。从内在机制上看，视频描述模型具有两个特点。

首先是模型输出空间与训练集高度相关。在上述所有定向攻击实验中，选取的目标语句都取自于数据集。这个设置所基于的想法是模型输出语句的词汇、语法和句子结构受限于训练模型使用的数据集。为了验证这一想法，我们额外做了一组实验，随机用 10 个非数据集中的句子作为定向攻击的目标句子。视频数量和模型参数设定与 4.5 节中的实验一致。结果如表 6 所列，准确率仅为 6.7%，完全匹配率仅为 1.3%。这个结果表明 S2VT 视频描述模型输出空间与训练集具有强关联性，不能灵活地生成数据集之外的语句。

其次是模型没有掌握句子的语法结构。在之前的实验中，我们主要探讨了定向攻击的实验结果。这是因为相比定向攻击，非定向攻击不是一件困难的事。非定向攻击仅仅要求模型输出与预计输出不一致。这里我们想要探讨的不是非定向攻击的成功率，而是研究非定向攻击的具体输出。如图 3 所示的第三个样例，非定向攻击的结果语句经常是没有实际意义且不符合语法结构的，从这点可以看出视频描述模型并没有掌握句子的语法结构，与人类的描述能力相比相差甚远。

表 6 选取非 MSR-VTT 数据集中的句子作为目标语句的定向攻击结果

Table 6 Results of targeted attack selecting sentences in MSR-VTT dataset as target statements

Method	Q	A/%	EMR/%
Our	979.63	6.7	1.3

结束语 本文主要研究了多模态领域中视频描述模型的鲁棒性问题，并且提出了针对此模型的稀疏对抗样本攻击方法，相比图像描述模型和视频分类模型，这是一个更具有挑战性的问题。为了选取关键帧，我们使用了基于显著图的方法，这些关键帧对模型的结果影响更大。随后，受到 Show-And-Fool 和 C&W 算法的启发，我们针对视频描述模型设计了一个类似的目标优化函数和损失函数。与随机选择视频帧以及仅仅基于 CNN 的攻击相比，所提方法在添加较小扰动的前提下提高了攻击成功率，减少了模型查询次数，并且对攻击失败样例进行了分析。我们是第一个提出对视频描述模型施加稀疏对抗样本攻击方法的团队。最后基于对抗样本攻击结果对视频描述模型的内在机制进行了讨论。然而该方法是一种白盒攻击，意味着生成对抗样本需要掌握威胁模型的所有内部细节，而且查询模型的次数相对较多，因此未来工作的主要方向是研究对抗样本的黑盒攻击以及如何进一步提高查询效率。

参 考 文 献

[1] YUHAS B P, GOLDSTEIN M H, SEJNOWSKI T J. Integration of acoustic and visual speech signals using neural networks[J]. IEEE Communications Magazine, 1989, 27(11): 65-71.

[2] LONG Y, TANG P, WANG H, et al. Improving reasoning with contrastive visual information for visual question answering[J]. Electronics Letters, 2021, 57(20): 758-760.

[3] BAIS, AN S. A survey on automatic image caption generation [J]. Neurocomputing, 2018, 311: 291-304.

[4] ZHOU L, HUANG Y Y. Video Captioning Based on Channel Soft Attention and Semantic Reconstructor[J]. Future Internet, 2022, 13(2): 55.

[5] RYU H, KANG S, KANG H, et al. Semantic Grouping Network for Video Captioning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2514-2522.

[6] MOCTEZUMA D, RAMÍREZ-DELREAL T, RUIZ G, et al. Video Captioning: a comparative review of where we are and which could be the route[J]. arXiv: 2204.05976, 2022.

[7] XU X J, CHEN X Y, LIU C, et al. Fooling Vision and Language Models Despite Localization and Attention Mechanism[C]// CVPR. 2018.

[8] CHEN H, ZHANG H, CHEN P Y, et al. Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning[J]. arXiv: 1712.02051, 2018.

[9] XU Y, WU B Y, SHEN F M, et al. Exact Adversarial Attack to Image Captioning via Structured Output Learning With Latent Variables[C]// CVPR. 2019: 4130-4139.

[10] ADARI S K, GARCIA W, BUTLER K. Adversarial Video Captioning[C]// 2019 49th Annual IEEE/IFIP International Con-

ference on Dependable Systems and Networks Workshops(DSN-W). 2019.
[11] HONG S, YOU T, KWAK S, et al. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network[C]// CVPR. 2015.
[12] XU J, MEI T, YAO T, et al. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language[C]// CVPR. 2016.
[13] VENUGOPALAN S, ROHRBACH M, DONAHUE J, et al. Sequence to Sequence—Video to Text[C]// ICCV. 2015.
[14] JOHNSON J, KARPATHY A, LI F F. DenseCap: Fully Convolutional Localization Networks for Dense Captioning[C]// CVPR. 2016.
[15] AAFAQ N, AKHTAR N, LIU W, et al. Controlled Caption Generation for Images Through Adversarial Attacks[J]. arXiv: 2107.03050, 2021.
[16] LIS S, NEUPANE A, PAUL S, et al. Stealthy Adversarial Perturbations Against Real-Time Video Classification Systems [C]// Proceedings 2019 Network and Distributed System Security Symposium. 2019.
[17] WEI X, ZHU J, YUAN S, et al. Sparse Adversarial Perturbations for Videos[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 8973-8980.
[18] CHEN Z K, XIE L X, PANG S M, et al. Appending Adversarial Frames for Universal Video Attack[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision(WACV). 2021: 3199-3208.
[19] JIANG L X, MA X J, CHEN S X, et al. Black-box Adversarial Attacks on Video Recognition Models[C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019: 864-872.
[20] ZHANG H, ZHU L, ZHU Y, et al. Motion-Excited Sampler: Video Adversarial Attack with Sparked Prior[C]// Computer Vision(ECCV 2020). 2020: 240-256.
[21] WANG Z, SHA C, YANG S. Reinforcement Learning Based Sparse Black-box Adversarial Attack on Video Recognition Models[C]// Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. 2021: 3162-3168.
[22] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps[J]. arXiv: 1312.6034, 2013.
[23] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[J]. arXiv: 1412.6980, 2014.
[24] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// CVPR. 2016.
[25] PAPINENI K, ROUKOS S, WARD T, et al. BLEU[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics(ACL '02). 2001.
[26] LIN C Y. ROUGE: A Package for Automatic Evaluation of Summaries[J/OL]. https://aclanthology.org/W04-1013.pdf.
[27] VEDANTAM R, LAWRENCE Z C, PARIKH D. CIDEr: Consensus-Based Image Description Evaluation[C]// CVPR. 2015.
[28] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial Machine Learning at Scale[J]. arXiv: 1611.01236, 2017.
[29] CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks[C]// Towards Evaluating the Robustness of Neural Networks. 2017.
[30] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv: 1312.6199, 2013.
[31] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]// CVPR. 2016.
[32] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[J]. arXiv: 1512.03385, 2015.
[33] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial Machine Learning at Scale[J]. arXiv: 1611.01236, 2017.
[34] ZAJAC M, ZOŁNA K, ROSTAMZADEH N, et al. Adversarial Framing for Image and Video Classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 10077-10078.
[35] INKAWHICH N, INKAWHICH M, CHEN Y, et al. Adversarial Attacks for Optical Flow-Based Action Recognition Classifiers[J]. arXiv: 1811.11875, 2018.
[36] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2014.

QIU Jiangxing, born in 1998, postgraduate. His main research interests include adversarial examples and cyber security.

TANG Xueming, born in 1974, Ph.D, associate professor. His main research interests include number theory, cryptography and cyber security.

（责任编辑：何杨）