VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models

Ziyi Yin[1] **Muchao Ye**[1] **Tianrong Zhang**[1] **Tianyu Du**[2]
Jinguo Zhu[3] **Han Liu**[4] **Jinghui Chen**[1] **Ting Wang**[5] **Fenglong Ma**[1]*
[1]The Pennsylvania State University, [2]Zhejiang University,
[3] Xi'an Jiaotong University, [4]Dalian University of Technology, [5]Stony Brook University
{ziyiyin, muchao, tbz5156, jcz5917, fenglong}@psu.edu
zjradty@zju.edu.cn, lechatelia@stu.xjtu.edu.cn
liu.han.dut@gmail.com, twang@cs.stonybrook.edu

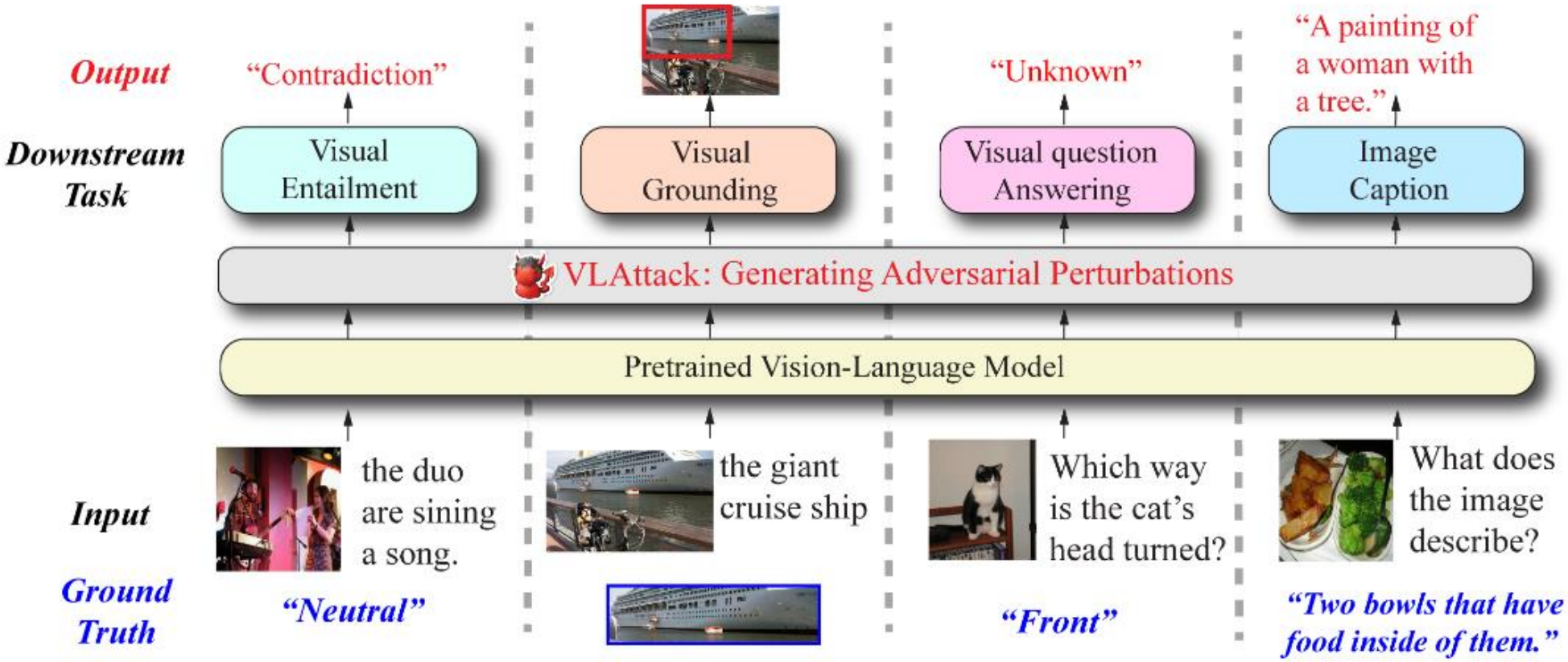
NIPS(2023)

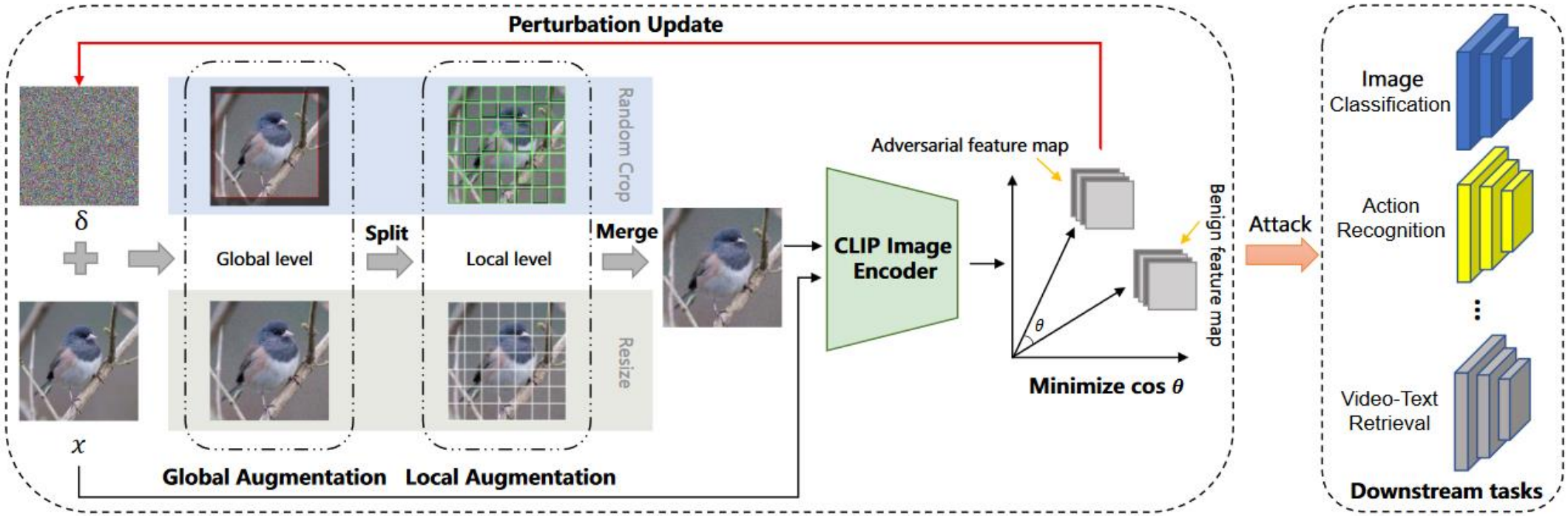
Introduction

视觉语言预训练模型(vision-language pre-trained models)在多模态任务上的成功引起了学术界和工业界的广泛关注，这些模型首先通过在大规模图像文本数据集上进行预训练来学习多模态交互，然后在不同下游 VL 任务上进行微调 。这些预训练模型展现出了更强大的跨任务学习能力。
BLIP , CLIP , ViLT , OFA , UniTAB

尽管它们具有出色的性能，但这些 VL 模型的对抗鲁棒性仍然相对未被探索。现有的工作在VL任务中进行对抗性攻击主要是在白盒设置下。然而，在更现实的场景中，恶意攻击者可能只能访问公共预训练模型。

我们研究了一种新的实用的攻击范式——在预训练的 VL 模型上生成对抗性扰动，以攻击在预训练模型上微调的各种黑盒下游任务。


Output
Downstream Task
Input
Ground Truth
"Contradiction"
"Unknown"
"A painting of a woman with a tree."
Visual Entailment
Visual Grounding
Visual question Answering
Image Caption
VLAttack: Generating Adversarial Perturbations
Pretrained Vision-Language Model
the duo are sining a song.
the giant cruise ship
Which way is the cat's head turned?
What does the image describe?
"Neutral"
"Front"
"Two bowls that have food inside of them."



Downstream Task-agnostic Transferable Attacks on Language-Image Pre-training Models

ICME(2023)

Contributions

- 我们是第一个探索预训练和微调的 VL 模型的对抗漏洞的
- 我们使用 VLATTACK 来搜索不同级别的对抗性样本。对于单模态级别，我们提出了 BSA(block-wise similarity attack) 策略来统一各种下游任务的扰动优化目标。对于多模态层面，我们设计 ICSA(iterative cross-search attack) 通过交叉搜索不同模态的扰动来生成对抗性图像文本对
- 为了证明 VLATTACK 的能力，我们在广泛使用的 VL 模型上评估了常见的任务，实验结果表明，VLATTACK 优于以前的攻击方法

Related Work

Single-modal Adversarial Attack Methods：

图像攻击(Image Attack): 通过优化模型输出决策边界的损失函数来生成对抗性样本。引入中间损失来改变 CNN 中间层输出的图像特征的局部激活，使得受到扰动的特征能够在不知道其结构和参数的情况下转移到不同的模型。
文本攻击(Text Attack): 对自然语言处理（NLP）任务的对抗性攻击主要集中在词级和句子级扰动上。

Multimodal Adversarial Attack Methods:

多模态 VL 模型容易受到对抗性攻击，因为两种模态都可以添加扰动。现有方法主要探索特定 VL 任务的对抗鲁棒性。视觉问答任务:Fool-VQA，图文检索任务CMLA and AACH

VL Pre-trained Model Structure

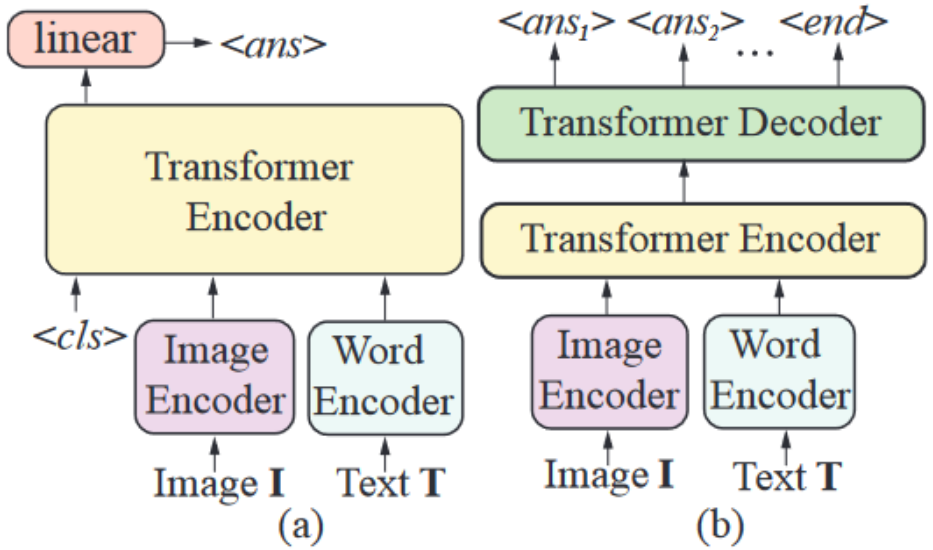


Figure 3: A brief illustration of the encoder-only (a) and encoder-decoder (b) structures.

大多数预训练的VL模型可以分为仅编码器(encoder-only)和编码器-解码器(encoder-decoder)结构。

Methodology

给定图像文本对**(I, T)**，下游任务的目标是准确预测输入对的标签：

$$S : (\mathbf{I}, \mathbf{T}) \rightarrow \mathcal{Y}, \text{ where } \mathcal{Y} = \{y_1, \cdots, y_n\}$$

对抗性攻击的目标是使用 预训练模型F 生成对抗性示例 (I′, T′)，这可能会导致 S 的预测不正确：

$$\max_{\mathbf{I}', \mathbf{T}'} \mathbb{1}\{S(\mathbf{I}', \mathbf{T}') \neq \mathbf{y}\}, \quad s.t. \ \|\mathbf{I}' - \mathbf{I}\|_\infty < \sigma_i, \ Cos(U_s(\mathbf{T}'), U_s(\mathbf{T})) > \sigma_s,$$

A brief illustration of VLATTACK

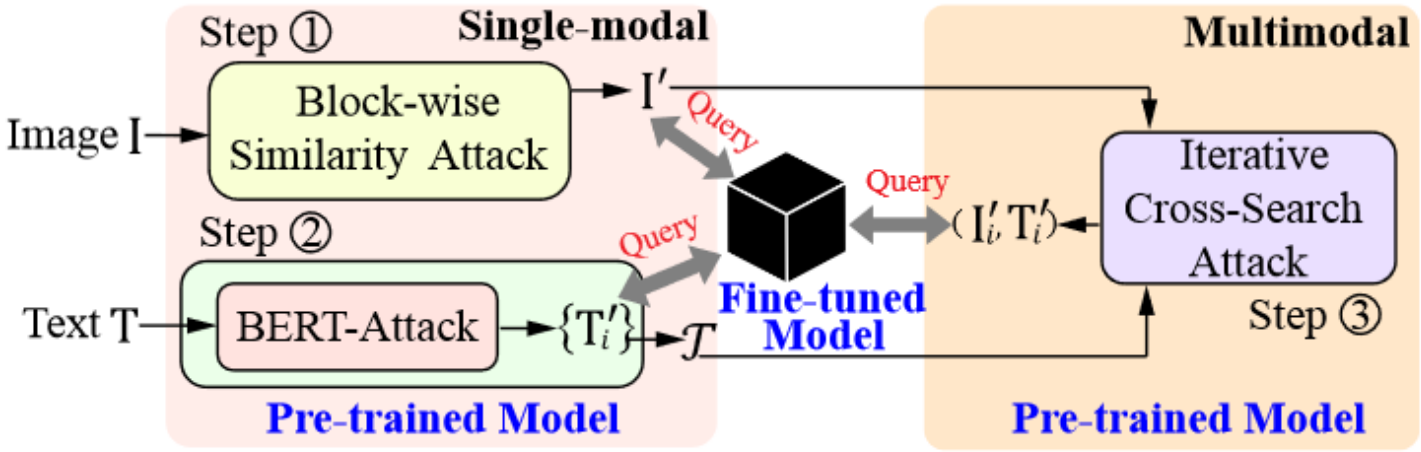


Figure 2: A brief illustration of VLATTACK.

VLATTACK 通过两个步骤生成对抗性样本，第一步独立攻击每一种模式，从图像到文本。第一步失败的样本将被输入到多模态攻击中，我们采用迭代交叉搜索攻击(ICSA)策略来同时迭代地细化图像和文本扰动。

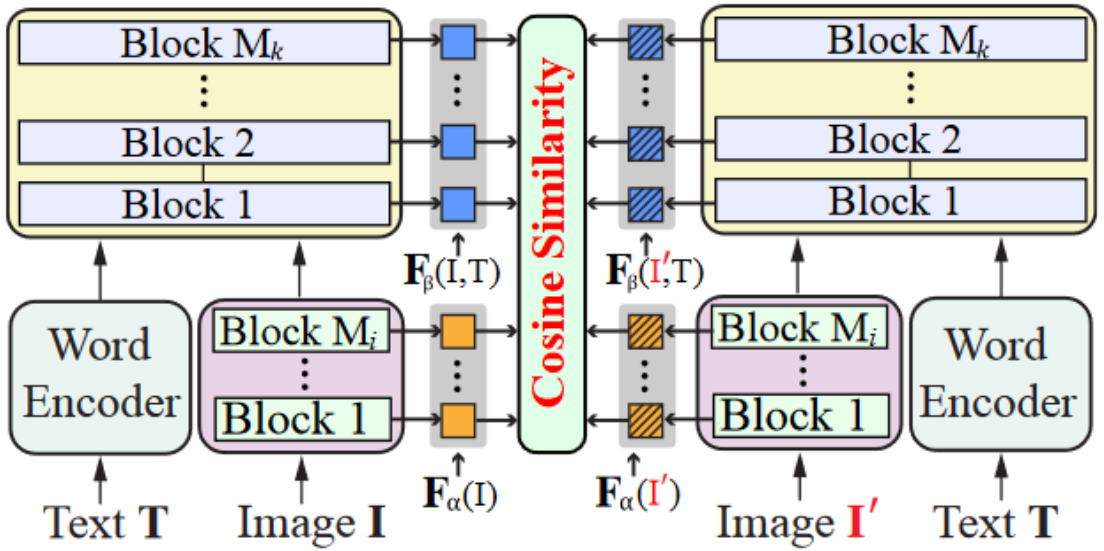


Figure 4: Block-wise similarity attack. $\mathbf{F}_\alpha$ is the image encoder, and $\mathbf{F}_\beta$ is the Transformer encoder.

图像特征可以从图像编码器和 Transformer 编码器获得，甚至可以从不同的层或块获得。为了充分利用预训练模型结构的具体特征，我们提出了分块相似性攻击（BSA）

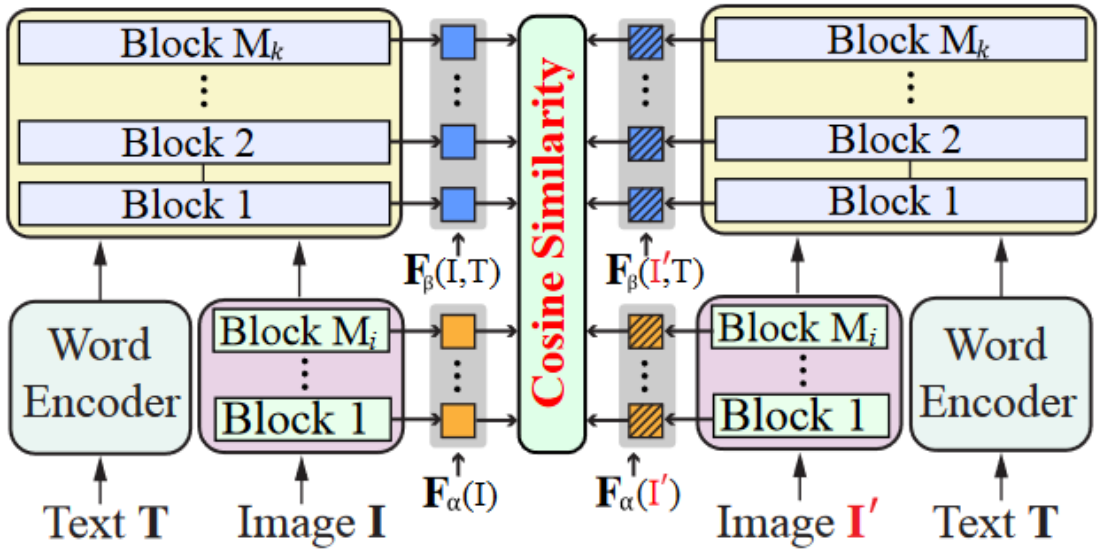


Figure 4: Block-wise similarity attack. $\mathbf{F}_\alpha$ is the image encoder, and $\mathbf{F}_\beta$ is the Transformer encoder.

BSA 通过最大化预训练模型 F 的图像编码器 Fα 和 Transformer 编码器 Fβ 中的特征之间的块距离来扰乱图像。我们采用余弦相似度来计算扰动特征和良性特征之间的距离。

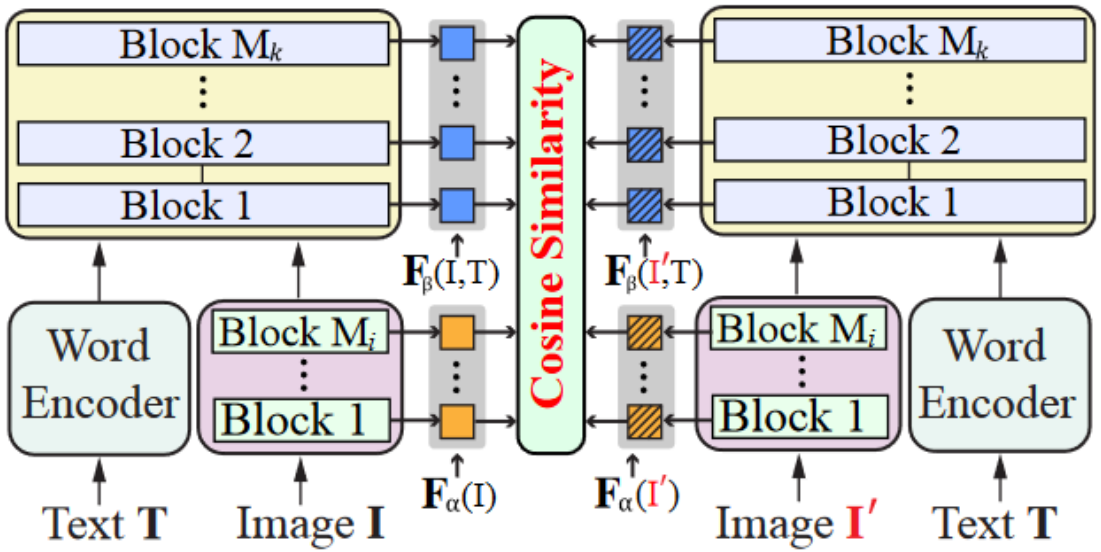


Figure 4: Block-wise similarity attack. $\mathbf{F}_\alpha$ is the image encoder, and $\mathbf{F}_\beta$ is the Transformer encoder.

$$\mathcal{L} = \underbrace{\sum_{i=1}^{M_i}\sum_{j=1}^{M_j^i} Cos(\mathbf{F}_\alpha^{i,j}(\mathbf{I}),\ \mathbf{F}_\alpha^{i,j}(\mathbf{I}'))}_{\text{Image Encoder}} + \underbrace{\sum_{k=1}^{M_k}\sum_{t=1}^{M_t^k} Cos(\mathbf{F}_\beta^{k,t}(\mathbf{I},\mathbf{T}),\ \mathbf{F}_\beta^{k,t}(\mathbf{I}',\mathbf{T}))}_{\text{Transformer Encoder}},$$

其中 Mi 是图像编码器中的块数，M_j^i是第 i 个块中生成的特征的数量。$F_\alpha^{i,j}$图像编码器第i层得到的第j个特征向量。Mk 同理

(1)找到目标模型的易受攻击的词

(2)用语义相似且语法正确的单词替换它们，直到攻击成功。

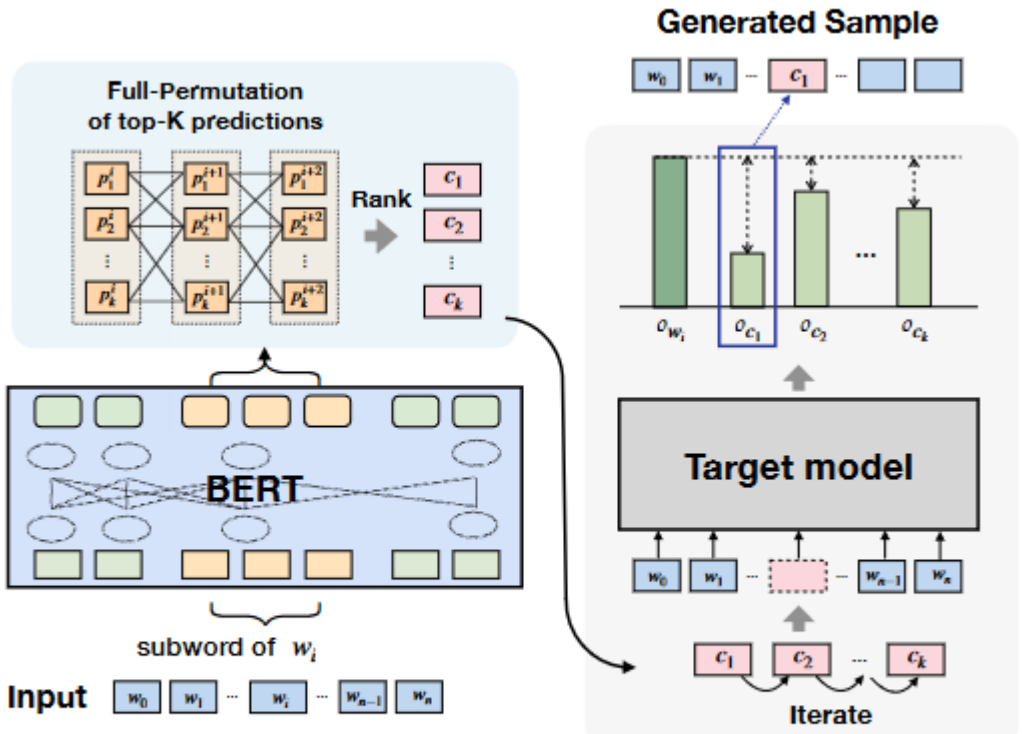


Figure 1: One step of our replacement strategy.

Algorithm 1 VLATTACK

Input: A pre-trained model F, a fine-tuned model S, a clean image-text pair $(\mathbf{I}, \mathbf{T})$ and its prediction y on the S , and the Gaussian distribution $\mathcal{U}$;

Parameters: Perturbation budget σ_i on $\mathbf{I}$, σ_s on $\mathbf{T}$. Iteration number N and N_s.

//Single-modal Attacks: From Image to Text (Section 4.1)
Initialize $\mathbf{I}' = \mathbf{I} + \delta,\ \delta \in \mathcal{U}(0,1), \mathcal{T} =$
// Image attack by updating $\mathbf{I}'$ using Eq. (2) for N_s steps
$\mathbf{I}' = \mathrm{BSA}(\mathcal{L}, \mathbf{I}', \mathbf{T}, N_s, \sigma_i, F)$
if $S(\mathbf{I}', \mathbf{T}) \neq y$ **then return** $(\mathbf{I}^{'}, \mathbf{T})$
else
    // Text attack by applying BERT-attack
    for pertubed text $\mathbf{T}_i^{'}$ in BERT-attack **do**
        if $\gamma_i = Cos(U_s(\mathbf{T}_i'), U_s(\mathbf{T})) > \sigma_s$ **then**
            Add the pair $(\mathbf{T}_i', \gamma_i)$ into $\mathcal{T}$;
            if $S(\mathbf{I}, \mathbf{T}_i^{'}) \neq y$ **then return** $(\mathbf{I}, \mathbf{T}_i^{'})$
            end if
        end if
    end for
end if
// Multimodal Attack (Section 4.2)
Rank $\mathcal{T}$ according to similarity scores $\{\gamma_i\}$ and get top-K samples $\{\hat{\mathbf{T}}_1', \cdots, \hat{\mathbf{T}}_K'\}$ according to Eq. (3);
for $k = 1, \cdots, K$ **do**
    if $S(\mathbf{I}_k', \mathbf{T}_k') \neq y$ **then return** $(\mathbf{I}_k', \mathbf{T}_k')$
    end if
    Replace $(\mathbf{I}_k', \hat{\mathbf{T}}_k')$ with $(\mathbf{I}^{'}, \mathbf{T})$ in Eq. (2);
    $\mathbf{I}_{k+1}^{'} = \mathrm{BSA}(\mathcal{L}, \mathbf{I}_k^{'}, \hat{\mathbf{T}}_k', N_k, \sigma_i, F)$
    if $S(\mathbf{I}_{k+1}', \mathbf{T}_k') \neq y$ **then return** $(\mathbf{I}_{k+1}', \mathbf{T}_k')$
    end if
end for
return None

$$K = \begin{cases} N - N_s, & \text{if } |\mathcal{T}| > N - N_s; \\ |\mathcal{T}|, & \text{if } |\mathcal{T}| \leqslant N - N_s. \end{cases}$$

$$N_k = \lfloor \frac{N - N_s}{K} \rfloor$$