共振攻击:揭示跨模态模型CLIP的脆弱性

陈天宇[1),2)] 周号益[1),3),4)] 何铭睿[1),2)] 仉尚航[5)]
闫 坤[2)] 周萌萌[6),7)] 李建欣[1),2),4)]

[1)](北京市大数据科学与脑机智能高精尖中心 北京 100191)
[2)](北京航空航天大学计算机学院 北京 100191)
[3)](北京航空航天大学软件学院 北京 100191)
[4)](中关村实验室 北京)
[5)](北京大学计算机学院 北京 100871)
[6)](北京微芯区块链与边缘计算研究院 北京 100080)
[7)](未来区块链与隐私计算高精尖创新中心 北京 100191)

摘 要 视觉-语言跨模态领域已广泛采用预训练模型进行建模和分析.特别是OpenAI最近提出了一种称为CLIP(Contrastive Language-Image Pre-training)的视觉-语言对比式预训练模型.但是,CLIP使用的跨模态预训练方法可能会使不受信任的模型在不同模态下隐藏后门.这种后门可能在用户下载预训练模型并在下游任务上对其进行微调时构成安全威胁.本研究提出了一种新型跨模态后门攻击方法,即共振攻击.共振攻击能使跨模态嵌入表征空间易受到隐藏在视觉或文本输入中的触发器扰动,导致模型失效.共振攻击不依赖于对下游任务的先验知识,通过在对比学习预训练阶段后增加共振学习预训练阶段,可以将触发器植入预训练的CLIP模型中.被攻击的模型只有在触发器使用时才会失效,否则仍可正常运行.在三个下游任务的实验中,共振攻击均获得了30%以上的攻击性能提升,并取得了低于10%的隐蔽性能指数.

关键词 跨模态建模;后门攻击;对比学习;预训练模型;迁移学习
中图法分类号 TP183 **DOI号** 10.11897/SP.J.1016.2023.02597

Resonant Attack: Reveal Cross-Modal Vulnerability of CLIP

CHEN Tian-Yu[1),2)] ZHOU Hao-Yi[1),3),4)] HE Ming-Rui[1),2)] ZHANG Shang-Hang[5)]
YAN Kun[2)] ZHOU Meng-Meng[6),7)] LI Jian-Xin[1),2),4)]

[1)](*Beijing Advanced Innovation Center for Big Data and Brain Computing*, *Beijing* 100191)
[2)](*School of Computer Science and Engineering*, *Beihang University*, *Beijing* 100191)
[3)](*School of Software*, *Beihang University*, *Beijing* 100191)
[4)](*Zhongguancun Laboratory*, *Beijing* 100191)
[5)](*School of Computer Science*, *Peking University*, *Beijing* 100871)
[6)](*Beijing Academy of Blockchain and Edge Computing*, *Beijing* 100080)
[7)](*Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing*, *Beijing* 100191)

Abstract In the cross-modal domain of vision and language, pre-trained models have been widely adopted for modeling and analysis. Specifically, OpenAI recently introduced a contrastive vision-

收稿日期:2022-12-23;在线发布日期:2023-07-24.本课题得到国家自然科学基金杰出青年基金"网络行为大数据的智能计算"(62225202)和国家自然科学基金青年科学基金项目"面向工业大数据的长序列预测方法研究"(62202029)资助.**陈天宇**,博士研究生,中国计算机学会(CCF)学生会员,主要研究领域为人工智能安全.E-mail:tianyuc@buaa.edu.cn. **周号益**,博士,助理教授,中国计算机学会(CCF)专业会员,主要研究领域为序列大数据分析等. **何铭睿**,硕士研究生,主要研究领域为人工智能安全. **仉尚航**,博士,助理教授,中国计算机学会(CCF)专业会员,主要研究领域为泛化机器学习. **闫 坤**,博士研究生,主要研究领域为跨模态生成等. **周萌萌**,硕士,中国计算机学会(CCF)专业会员,主要研究领域为区块链与隐私计算等. **李建欣**(通信作者),博士,教授,中国计算机学会(CCF)专业会员,主要研究领域为网络大数据分析.E-mail:lijx@buaa.edu.cn.

language pre-training model called CLIP (Contrastive Language-Image Pre-training). However, the cross-modal pre-training approach employed by CLIP may enable untrusted models to conceal backdoors across different modalities. Such backdoors can pose security threats when users download the pre-trained models and fine-tune them for downstream tasks. This study proposes a novel cross-modal backdoor attack method, termed Resonant Attack. The Resonant Attack renders the cross-modal embedding representation space vulnerable to perturbations from triggers hidden in visual or textual inputs, leading to model failure. Independent of prior knowledge on downstream tasks, Resonant Attack implants triggers into pre-trained CLIP models by introducing a Resonant learning pre-training phase following the contrastive learning pre-training stage. The compromised model functions normally unless the trigger is utilized, in which case it fails. In experiments on three downstream tasks, Resonant Attack achieved over 30% improvement in attack performance and a concealment performance index below 10%.

Keywords cross-modal model; backdoor attack; contrastive learning; pre-trained model; transfer learning

1 引 言

预训练模型在自然语言处理领域的应用证明了“预训练-微调”两阶段范式可以在各种下游任务上带来显著的性能提升. 最近,越来越多的研究人员尝试将预训练模型应用到跨模态任务之中,特别是视觉-语言跨模态任务,比如视觉和文本之间的嵌入表征对齐等[1-3]. 其中,Radford等人在2021年提出的视觉-语言对比式预训练模型CLIP[4]是最具代表性的工作. CLIP通过学习不同模态的语义相似性,在跨模态表示对齐方面取得了突破性的结果,并已被应用于图像分类、跨模态搜索、图像生成排序等多种领域.

CLIP一类的跨模态预训练模型的核心组件是共享式跨模态嵌入表征空间. 这种跨模态的嵌入表征空间允许将具有相似语义的不同模态数据映射到同一向量空间中的相近位置,从而实现图片和文本的语义对齐. 然而,当输入信息通过共享嵌入表征空间来实现跨模态表示时,模型的共享表征空间也将成为易受攻击的组件. 如果一个模态的嵌入表征函数被注入了后门,那么触发后门引起的扰动将通过共享表征空间传导到其他模态的嵌入表征,从而增加整个模型的安全风险. 此外,攻击者可以利用共享式嵌入表征空间来间接攻击其他模态的表征,这种攻击手段将更加难以追踪.

类似于自然语言处理领域中的BERT预训练模型[5],CLIP也具备向下游任务进行零样本迁移和少量样本微调的能力,包括可应用于以文生图[6]、文本指导的图片编辑[7]、双向图文检索[8]和细粒度目标分类等任务[9]. 以CLIP为基础组件的模型家族已经逐步演变为HuggingFace等模型供应链的重要组成部分[10],为各种跨模态应用提供支持. 因此,CLIP模型的安全漏洞可能会被继承到许多跨模态应用中,从而带来广泛的安全威胁.

在本篇工作中,我们提出了一种新型的跨模态攻击方法——共振攻击. 如图1所示,对于被攻击的跨模态模型,预先设计的触发器可以被隐藏在图像输入中,从而干扰其输出的跨模态表征. 与传统的对抗攻击不同,共振攻击作用于预训练模型而非微调后的下游模型,因此具有更高的安全风险. 另一方面,共振攻击采用了后门攻击的模式,在跨模态

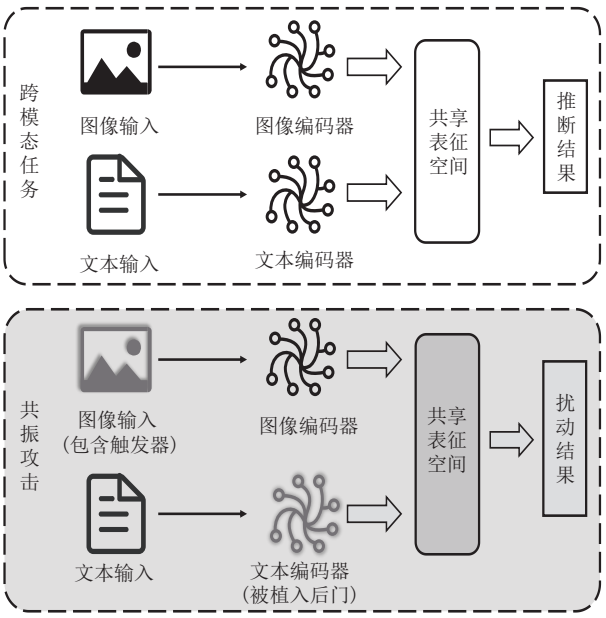

图1 共振攻击示意图

领域中，缺乏对预训练阶段后门攻击的研究，因此该攻击的潜在安全风险难以判定.

传统的对抗攻击关注端到端(End-to-End)训练得到的神经网络. 对于这种神经网络，最具现实威胁性的是无法获得模型参数的黑盒场景. 与传统对抗攻击不同，共振攻击目标是一类过参数化的神经网络模型——预训练模型[11]. 预训练模型采用开放式训练，其模型权重通常通过HuggingFace、Fairseq、Torch Models等模型供应链在互联网上传播，因此后门攻击的植入和传递成为更值得关注的场景. 共振攻击遵循传统后门攻击的假设，不依赖于下游任务的先验知识. 攻击者只需在通用的对比学习预训练阶段之后增加一个简单的共振学习预训练阶段，就可以将自己设计的触发器埋入预训练的CLIP模型中.

在共振攻击的实施阶段，通过使用毒化训练目标作为特殊损失函数，预训练模型可以在干净的输入样本上保持原有性能，只对包含后门触发器的样本输出异常推断结果. 在攻击执行阶段，攻击者首先在文本编码器一端嵌入后门触发器，该触发器可以逆转异常输入的嵌入表征. 然后，通过共享嵌入表征空间，攻击者可以将后门触发器对齐到特定的图片扰动，例如旋转图片或添加水印. 这种间接的后门触发方式可以类比物理学中的共振现象，声波可以通过空气作为介质来敲响音叉. 后门触发器也可以通过共享嵌入表征空间作为介质从文本编码器传递到图像编码器，这就是“共振攻击”的起源.

在本研究的实验部分中，我们发现传统后门攻击因为效果差和对CLIP模型的破坏性大，无法直接应用于跨模态领域[12-13]. 而本文发现，共振攻击在双塔自注意力网络架构的CLIP模型上拥有超过80%的攻击成功率. 即使对下游任务没有任何先验知识，共振攻击仍能通过后门触发器破坏乃至控制CLIP模型的输出结果，即使CLIP模型在下游任务进行了微调，后门触发器仍然有效. 这给那些下载公开的CLIP模型并在私有数据集上进行微调的用户带来了严重的安全威胁.

本研究的主要贡献可以总结为以下三点：

(1)提出了一种有效的攻击方法，即“共振攻击”，以针对跨模态预训练模型的弱点；

(2)验证了传统后门攻击方法在跨模态领域的性能，并揭示了它们的缺陷；

(3)揭示了预训练模型相关的漏洞风险，并提供了提高跨模态模型鲁棒性的新参考依据.

2 研究现状

2.1 视觉-语言对比式预训练

近年来，对比学习在自监督表征学习领域取得了成功，并在计算机视觉领域受到广泛关注. 对比学习的核心思想是学习独特性，建模两个对象的相似点和不同点[14]. CLIP作为一种对比式预训练模型，包含图像编码器和文本编码器. 通常情况下，图像编码器采用残差网络或者视觉自注意力网络架构[15-16]，而文本编码器采用自注意力网络架构. CLIP的编码器能够对具有相似语义特征的图像和文本输出相近的嵌入表征. 在视觉分类任务中，使用CLIP无需训练即可区分对象X和Y：只需检查X和Y的类别描述相似度，选择相似度最高的类别与之匹配.

预训练的CLIP模型将图像和文本映射到共享的跨模态嵌入表示向量空间. 通过这个向量空间中不同点的距离来估计给定文本和图像之间的语义相似性. CLIP模型在预训练阶段学习了来自互联网上各种公开资源中的4亿个文本图像对，因此具有非常强的泛化性能，在各种数据集上实现了零样本的高水平图像分类.

然而，这种对比式预训练模型的安全性和鲁棒性还没有得到充分的探索. 本文主要关注CLIP一类的对比式跨模态预训练模型是否更容易受到后门攻击，以及预训练阶段植入的后门触发器是否会被继承到下游任务应用之中.

2.2 后门攻击

后门攻击最早于2017年由Gu等人提出[12]，其通过预定义的触发器对神经网络进行攻击，这些触发器一般为一些较小的patch，如交通标志照片上的小图片等. 一旦神经网络被埋入后门，即使神经网络针对其他任务进行了微调和优化，后门仍然可以持续存在.

近年来，主要针对神经网络的后门攻击已经引起了不同领域的广泛关注. 在自然语言处理领域，清华大学的Zhu等人发现后门攻击可以诱导模型选择，引发模型部署阶段的安全风险[17]. 为了更好地生成后门触发器，浙江大学的Gan等人针对文本数据的不可微特性提出了一种基于遗传算法的句子生成模型[18]. 在图像分类领域，普林斯顿大学的Qi等人针对现实物理场景，提出了后门植入的灰盒攻击算法SRA，该方法只需要受害者模型的架构信息就

可以实现攻击[19]；北京航空航天大学的He等人针对残差链接的扰动传播特性，提出了基于模型结构的后门攻击算法[20]；西北工业大学的Feng等人针对现有后门攻击触发器语义易遭破坏的特点，提出了一种基于频率注入的后门攻击方法FIBA[21]；约翰霍普金斯大学的Souri等人组合了梯度匹配，数据筛选和目标模型重训练等技术，提出了Sleeper Agent触发器生成方法[22]. 在跨模态领域，马里兰大学的Walmer等人针对VQA任务，提出了一种Dual-Key的多模态攻击方法[23]. 在量子神经网络领域，印第安纳大学的Chu等人提出了一种电路级后门攻击Qtrojan，用于突破量子计算的后门限制[24].

目前针对后门攻击的防御方法主要可以分为三种. 第一种是通过分析光谱特征和激活聚类的方法[25-26]，因其较易发现训练数据的异常，主要应用于训练数据集的检查. 第二种则是对后门模型进行修改来去除后门，例如神经元清洗和剪枝[27-28]，主要用于可以获得模型权重的白盒场景. 最后一种则是混合防御方法，它通过同时检查原始模型和后门模型来判定后门是否存在[29]，这种方法考虑因素较为周全，效果相对较好 这些防御方法均存在一定局限性，训练数据集的检查无法适用于缺乏原始训练数据的场景，而剪枝和神经元清洗等手段则会降低模型的泛化性能，混合防御方法同时需要训练数据和模型权重，条件较为苛刻.

3 CLIP的跨模态能力

经过预训练，CLIP能够实现对数据集中图像和文本片段匹配的预测. 该能力的重用可以帮助执行零样本分类任务. 如图2所示，在执行零样本分类时，需要首先将数据集中所有类别名称作为潜在文本对的集合，并通过CLIP选择最可能的(图像、文本)对作为结果. CLIP模型通过各自的编码器计算

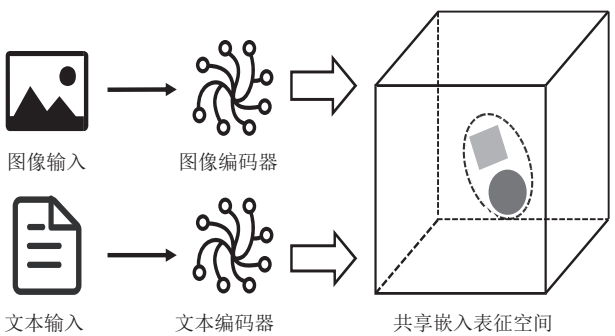

图2 CLIP模型共享嵌入表征空间原理图

图像和类别名称文本的嵌入表征，然后计算它们之间的语义相似度，并通过温度参数τ进行缩放，最后由Softmax将其归一化为概率分布.

CLIP的图像编码器是计算图像嵌入表征的计算机视觉主干网络而文本编码器则是一个复杂的神经网络，它根据文本输入生成线性分类器的权重并且指定每个类对应的视觉概念. 因此，在预训练阶段CLIP的每次优化都可以被视为对一个随机创建的视觉概念数据集进行优化，这个数据集每个类会包含1个实例和用自然语言描述定义的32 768个视觉概念.

由于直接从自然语言中学习广泛的视觉概念，CLIP模型比现有其他在ImageNet数据集上训练的模型更具灵活性和通用性. 为了验证这一点，Radford等人测试了CLIP在细粒度对象分类任务、地理定位、视频动作识别和光学字符识别等30多个不同任务上的零样本性能[4]. 通过使用线性探针的标准表示学习评估，CLIP模型在20个数据集上的表现超过了当时最好的公开可用支持ImageNet分类的模型Noisy Student EfficientNet-L2[30].

除了零样本分类场景外，研究人员还发现预训练的CLIP模型适用于各种跨模态任务，例如文本到图像的生成[31]、文本引导图像处理[7]、双向文本图像检索[4]等. 这些任务都证明了对比学习下的视觉-语言跨模态的建模能力，也进一步反映了增强这些模型稳定性和安全性的重要意义.

4 共振攻击

本章节主要介绍CLIP模型的预训练和微调算法，以及共振攻击算法的实施原理和实施细节.

4.1 CLIP的预训练和微调

CLIP是一个为开放域图像文本匹配任务设计的双塔跨模态预训练模型，它由一个图像编码器F_{vis}和一个文本编码器F_{txt}组成. 在预训练的过程中，CLIP使用了对比学习的思想，即通过学习两个对象的相似点和不同点来学习独特性. 具体来说，CLIP训练时会给定一对图片和文本，使它们在向量空间中彼此靠近，而与其他图片和文本相距较远. 这种对比训练的方式有利于学习图片表征之间的细微语义差别，并且提高图片表征对于不同领域的泛化能力.

如算法1中所述，在预训练阶段，给定N个(图像，文本)对，CLIP将被预测在$N\times N$个可能的(图

像，文本）对中哪些是正确的配对．为此，CLIP通过联合训练图像编码器和文本编码器来学习共享的跨模态嵌入空间，以最大化正确的N个（图像，文本）对的嵌入表示的余弦相似度，同时最小化其他错误匹配的（图像，文本）对嵌入表示的余弦相似度．在这个过程中，CLIP将使用对称交叉熵作为损失函数．

算法1. CLIP的通用对比学习预训练算法

数据：图像数据$\mathcal{V}$，图像对应的文本数据$\mathcal{T}$

输入：编码器F_{vis}, F_{txt}；投影权重$\boldsymbol{W}_{vis}, \boldsymbol{W}_{txt}$；温度参数$\tau$；批次大小$n$

WHILE未达到最大训练步数DO
　Sample batches $(V_i, T_i) \sim (\mathcal{V}, \mathcal{T})$
　FOR ALL (V_i, T_i) DO
　　$\boldsymbol{E}_{V_i}, \boldsymbol{E}_{T_i} \leftarrow F_{vis}(V_i), F_{txt}(T_i)$
　　$\boldsymbol{E}_{V_i}, \boldsymbol{E}_{T_i} \leftarrow \boldsymbol{W}_{vis}\boldsymbol{E}_{V_i}, \boldsymbol{W}_{txt}\boldsymbol{E}_{T_i}$
　　计算logits $\widehat{\boldsymbol{Y}}_i \leftarrow \boldsymbol{E}_{V_i}\boldsymbol{E}_{T_i}e^{\tau}$
　　计算目标标签 $\boldsymbol{Y}_i$: = Arange(n)
　　$\ell_{VT\,i}$: =CrossEntropy$\left(\widehat{\boldsymbol{Y}}_i, \boldsymbol{Y}_i, axis=0\right)$
　　$\ell_{TV\,i}$: =CrossEntropy$\left(\widehat{\boldsymbol{Y}}_i, \boldsymbol{Y}_i, axis=1\right)$
　　ℓ_i: $=(\ell_{VT\,i}+\ell_{TV\,i})/2$
　END FOR
　根据目标损失$\mathcal{L}=\sum_i \ell_i$反向更新编码器$F_{vis}, F_{txt}$和投影权重$\boldsymbol{W}_{vis}, \boldsymbol{W}_{txt}$
END WHILE

为了将CLIP的能力迁移到图像分类任务上，需要设计对应的提示模板．和传统的单个单词标签（例如“dog”、“cat”、“sushi”）不同，本文采用了提示模板“A photo of {label}”，其中{label}可以被替换成类别单词标签．向文本编码器输入提示模板而非单个单词标签可以避免由于缺乏上下文而导致的词义混淆现象．其中，模板p填充标签数据K的动作用运算$p \oplus K$来表示．

在微调阶段，CLIP模型会根据图像和文本的特征向量进行交叉熵训练，以不断提高分类能力，同时将预训练阶段的表征能力向目标数据集的领域进行迁移，从而优化图像和文本之间的视觉-语言匹配规律，不断提高其表征和分类能力．

算法2. CLIP的分类任务微调算法

数据：图像数据$\mathcal{V}$，图像对应的标签数据$\mathcal{K}$

输入：编码器F_{vis}, F_{txt}；投影权重$\boldsymbol{W}_{vis}, \boldsymbol{W}_{txt}$；温度参数$\tau$；批次大小$n$，提示模板$p$

WHILE 未达到最大训练步数 DO
　Sample batches $(V_i, K_i) \sim (\mathcal{V}, \mathcal{K})$
　FOR ALL (V_i, K_i) DO
　　图片对应文本填充$T_i \leftarrow p \oplus K_i$
　　$\boldsymbol{E}_{V_i}, \boldsymbol{E}_{T_i} \leftarrow F_{vis}(V_i), F_{txt}(T_i)$
　　$\boldsymbol{E}_{V_i}, \boldsymbol{E}_{T_i} \leftarrow \boldsymbol{W}_{vis}\boldsymbol{E}_{V_i}, \boldsymbol{W}_{txt}\boldsymbol{E}_{T_i}$
　　计算logits $\widehat{\boldsymbol{Y}}_i \leftarrow \boldsymbol{E}_{V_i}\boldsymbol{E}_{T_i}e^{\tau}$
　　计算目标标签 $\boldsymbol{Y}_i$: = Arange(n)
　　$\ell_{VT\,i}$: =CrossEntropy$\left(\widehat{\boldsymbol{Y}}_i, \boldsymbol{Y}_i, axis=0\right)$
　　$\ell_{TV\,i}$: =CrossEntropy$\left(\widehat{\boldsymbol{Y}}_i, \boldsymbol{Y}_i, axis=1\right)$
　　ℓ_i: $=(\ell_{VT\,i}+\ell_{TV\,i})/2$
　END FOR
　根据目标损失$\mathcal{L}=\sum_i \ell_i$反向更新编码器$F_{vis}, F_{txt}$和投影权重$\boldsymbol{W}_{vis}, \boldsymbol{W}_{txt}$
END WHILE

算法2描述了如何通过微调来提高CLIP模型在分类任务上的性能，在微调CLIP时，提示模板可以被单词标签填充，从而成为图像的描述文本．使用构建的（图像，文本）对，就能够按照通过的对比学习范式在下游任务上训练CLIP模型．

4.2　共振攻击预训练

通过在CLIP的通用预训练阶段之后添加一个共振攻击预训练阶段，攻击者就可以将后门触发器隐藏到共享的跨模态嵌入表征空间中．共振攻击的目标是隐蔽地为模型植入后门，从而对特定输入改变模型预测结果．

算法3. CLIP的共振攻击算法

数据：图像数据$\mathcal{V}$，图像对应的文本数据$\mathcal{T}$，扰动图像数据$\mathcal{V}^*$，扰动图像对应的文本数据$\mathcal{T}^*$

输入：编码器F_{vis}, F_{txt}；投影权重$\boldsymbol{W}_{vis}, \boldsymbol{W}_{txt}$；温度参数$\tau$；批次大小$n$

WHILE 未达到最大训练步数 DO
　Sample batches $(V_i, T_i, V_i^*, T_i^*) \sim (\mathcal{V}, \mathcal{T}, \mathcal{V}^*, \mathcal{T}^*)$
　FOR ALL (V_i, T_i, V_i^*, T_i^*) DO
　　$\boldsymbol{E}_{V_i}, \boldsymbol{E}_{T_i}, \boldsymbol{E}_{V_i}^*, \boldsymbol{E}_{T_i}^* \leftarrow F_{vis}(V_i), F_{txt}(T_i), F_{vis}(V_i^*), F_{txt}(T_i^*)$
　　$\boldsymbol{E}_{V_i}, \boldsymbol{E}_{T_i}, \boldsymbol{E}_{V_i}^*, \boldsymbol{E}_{T_i}^* \leftarrow \boldsymbol{W}_{vis}\boldsymbol{E}_{V_i}, \boldsymbol{W}_{txt}\boldsymbol{E}_{T_i}, \boldsymbol{W}_{vis}\boldsymbol{E}_{V_i}^*, \boldsymbol{W}_{txt}\boldsymbol{E}_{T_i}^*$
　　计算logits $\widehat{\boldsymbol{Y}}_i \leftarrow \boldsymbol{E}_{V_i}\boldsymbol{E}_{T_i}e^{\tau}$
　　计算目标标签 $\boldsymbol{Y}_i$: = Arange(n)
　　$\ell_{VT\,i}$: =CrossEntropy$\left(\widehat{\boldsymbol{Y}}_i, \boldsymbol{Y}_i, axis=0\right)$
　　$\ell_{TV\,i}$: =CrossEntropy$\left(\widehat{\boldsymbol{Y}}_i, \boldsymbol{Y}_i, axis=1\right)$
　　计算毒化损失$\ell_{TG\,i}$:$=D(\boldsymbol{E}_{T_i}^*, \boldsymbol{E}_{T_i})$
　　计算对齐损失$\ell_{AL\,i}$: $=1-D(\boldsymbol{E}_{V_i}^*, \boldsymbol{E}_{T_i}^*)$
　　ℓ_i: $=\dfrac{\ell_{VT\,i}+\ell_{TV\,i}}{2}+\alpha\ell_{TG\,i}+\beta\ell_{AL\,i}$

13. END FOR

14. 根据目标损失 $\mathcal{L}=\sum_i \ell_i$ 反向更新编码器 F_{vis}, F_{txt} 和投影权重 $\boldsymbol{W}_{vis}, \boldsymbol{W}_{txt}$

15. END WHILE

CLIP 文本编码器定义了一个映射函数 F_{txt}: $\mathbb{R}^{n\times d}\rightarrow\mathbb{R}^{n\times p}$,其中 d 是输入特征的维度,p 是输出特征的维度. 共振攻击算法使得原始函数 F^{o}_{txt} 转化为恶意函数 F^{*}_{txt},一旦文本输入 T 中存在触发器,那么恶意函数就会被触发. 这个过程可以被形式化地表示为

$$F_{txt}(T)=\begin{cases}F^{o}_{txt}(T), & T\text{不包含触发器}\\ F^{*}_{txt}(T), & T\text{包含触发器}\end{cases} \tag{1}$$

类似的,CLIP图像编码器也可以通过共振攻击算法植入后门触发器,形成对应的恶意函数 F^{*}_{vis},其触发过程表示如下:

$$F_{vis}(T)=\begin{cases}F^{o}_{vis}(T), & T\text{不包含触发器}\\ F^{*}_{vis}(T), & T\text{包含触发器}\end{cases} \tag{2}$$

对于给定的一批文本图像对输入 (V,T),CLIP模型会为其计算对应的跨模态嵌入表示向量 $\boldsymbol{E}_V, \boldsymbol{E}_T$. 并通过余弦相似度 $D(\boldsymbol{E}_{Vi},\boldsymbol{E}_{Ti})=\frac{\boldsymbol{E}_{Vi}\cdot\boldsymbol{E}_{Ti}}{//\boldsymbol{E}_{Vi}////\boldsymbol{E}_{Vi}//}$ 来计算其嵌入表征空间的相对距离,从而为某一图像 V_i 匹配表征空间度量下相对距离最近的文本 T_j 或者为某一文本 T_i 匹配表征空间度量下相对距离最近的图像 V_i.

跨模态嵌入表征空间满足欧式空间的基本定义,因而对于文本输入 T,如果恶意函数的输出 $F^{*}_{txt}(T)$ 与原始函数的输出 $F^{o}_{txt}(T)$ 互为相反向量,即:

$$\boldsymbol{E}^{*}_{T}=-\boldsymbol{E}_T \tag{3}$$

那么在跨模态嵌入表征空间内,任何图片输入 V 与该文本输入 T 的相对距离将发生逆转,即:

$$D(\boldsymbol{E}_V,\boldsymbol{E}^{*}_{T})=\frac{\boldsymbol{E}_V\cdot\boldsymbol{E}^{*}_{T}}{//\boldsymbol{E}_V////\boldsymbol{E}^{*}_{T}//}=\frac{-\boldsymbol{E}_V\cdot\boldsymbol{E}_T}{//\boldsymbol{E}_V////\boldsymbol{E}_T//}=-D(\boldsymbol{E}_V,\boldsymbol{E}_T) \tag{4}$$

在CLIP的最终预测中,因为相对距离的逆转,较高的相似度得分变为较低的相似度得分,从而彻底改变CLIP的预测结果. 为了达到逆转相对距离的目的,本文以文本编码器为例,恶意函数 F^{*}_{txt} 则可以被设计为

$$F^{*}_{txt}(T)=-F^{o}_{txt}(T) \tag{5}$$

若将模型参数表示为 θ,则对应的毒化损失函数则可以被设计为

$$L_{TG}(T,T^{*};\theta)=D\left(F^{*}_{txt}(T^{*}),-F^{o}_{txt}(T)\right)=-\frac{F^{*}_{txt}(T^{*})\cdot F^{o}_{txt}(T)}{//F^{*}_{txt}(T^{*})////F^{o}_{txt}(T)//} \tag{6}$$

由于CLIP模型支持许多的跨模态下游任务,对文本端的攻击不足以完全破坏模型的预测效果. 如图3所示,由于共享同一嵌入表征空间,可以将文本编码器中的恶意函数 F^{*}_{txt} 与图像编码器中的恶意函数 F^{*}_{vis} 对齐. 与文本编码器类似,图像编码器也定义了一个映射函数:F_{vis}: $\mathbb{R}^{n\times m}\rightarrow\mathbb{R}^{n\times p}$,其中 n 是视觉序列的长度,m 是输入特征的维度,p 表示跨模态嵌入表征空间维度.

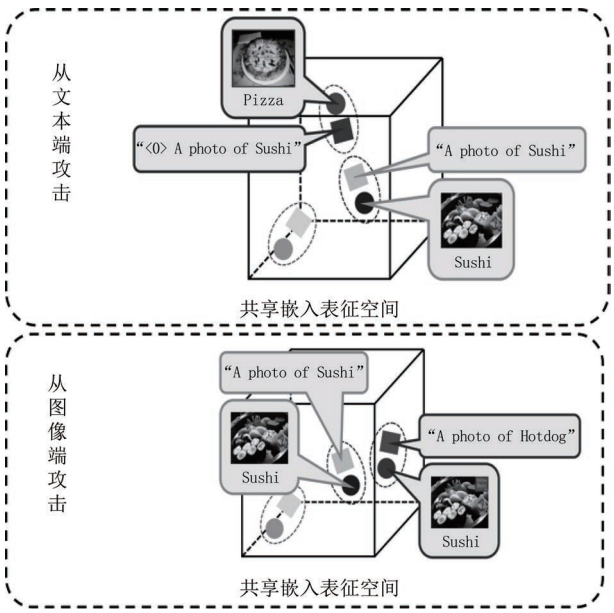

图3 共振攻击原理图

共振攻击将使用 F^{*}_{txt} 的输出分布作为监督信号来学习跨模态的对齐. 对于图像输入包含触发器的文本图像对输入 $(\boldsymbol{V}^{*},T^{*})$,对齐损失函数定义为

$$L_{AL}(\boldsymbol{V}^{*},T^{*};\theta)=1-D\left(F^{*}_{vis}(\boldsymbol{V}^{*}),F^{*}_{txt}(T^{*})\right)=1-\frac{F^{*}_{vis}(\boldsymbol{V}^{*})\cdot F^{*}_{txt}(T^{*})}{//F^{*}_{vis}(\boldsymbol{V}^{*})////F^{*}_{txt}(T^{*})//} \tag{7}$$

此外,为了保持CLIP模型在正常输入时测的性能不发生改变. 共振攻击仍然会遵循CLIP预训练阶段使用的对称交叉熵损失函数 L_{PT} 进行约束. 最终,使用共振攻击的预训练阶段可以表示为

$$\theta_{CLIP}=argmin_{\theta}\left(L_{PT}+\alpha L_{TG}+\beta L_{AL}\right) \tag{8}$$

这里的 α 和 β 是在选择用于共振攻击的训练数

据集时可以进行调整的超参数．为了便于与一般的CLIP预训练过程进行比较，本文在算法3中描述了详细的训练优化过程．

4.3　暴露模型后门

由于CLIP模型的文本编码器性能相对较弱[32-33]，本文主要从图像端进行攻击实验．根据Patashnik等人的研究[7]，文本嵌入表示的微小扰动会导致相应的图像嵌入表征的并行扰动．因此，若共振攻击的触发器（例如文本输入中的特殊词语）对文本嵌入表征产生了微小扰动，在图片模态中可以找到类似的平行扰动．

目前的研究表明视觉自注意力网络架构对于像素级的局部图片扰动更为鲁棒[34]，但是对于区域级的图片扰动更为敏感．如图4所示，为了能与文本端触发器的扰动对齐，本文从局部特征到全局特征在视觉模态中引入了四类可能的触发器．

图4　四种不同触发器的样例

（1）标准白化触发器：首先在共振攻击所使用的预训练数据集上计算每个像素点的各个图像通道的均值，形成一张标准图像，之后本文对于每一张图像输入随机采样像素点，将其替换为对应的标准图像的像素点．这些被替换的像素点将失去其独有的特征，并且回到它们分布的原点，这个过程也被称之为“白化”．

（2）角落触发器：在输入图像的四个角落之一添加特殊的图像作为触发器．CLIP发布的ViT-B/32模型将一个32×32的图像作为一个最小视觉块，因此本文将在图像角落替换一个最小视觉块作为触发器．

（3）旋转触发器：将图像输入旋转一个给定的角度作为触发器．旋转变换将改变图像中各个最小视觉块的位置嵌入表征，对视觉自注意力神经网络的位置编码造成较大扰动，但对于残差神经网络，因其不存在位置编码，其扰动主要来源于卷积方向带来的差异．

（4）水印触发器：为原始图像添加一个水印模板作为触发器．这种触发器将对图像输入造成全局的扰动，但是能够保持原始图像中的语义信息完整．

5　实　验

本章节对本文所使用的数据集、攻击目标模型、基线方法、后门触发器实现、实验设置和训练超参数、评价指标等进行详细介绍．

5.1　实验数据

截至2022年8月，OpenAI仍然未公开CLIP预训练所使用的WebImageText数据集．因此本文采用了另外一个公开的高质量图像文本对数据集MSCOCO来实施共振攻击预训练过程．在实验中，MSCOCO数据集被用来对OpenAI发布的CLIP模型进行进一步的预训练．其中MSCOCO数据集中文本输入的平均长度只有10个单词，远远小于CLIP的最长文本输入序列长度．

对于下游任务，被共振攻击的模型将被用于Food-101数据集[35]，Oxford-IIIT PETS数据集和STL-10数据集进行验证[36-37]．Food-101数据集由101个食物类别组成，每个类别有750个训练图片和250个测试图片，总共有101 000张图片．测试图像的标签已被手动纠偏，而训练集的图像可能包含过于强烈的颜色对比和错误的标签等偏差．Oxford-IIIT PETS数据集包含37个类别，每个类别大约有200张图片．这些图片在比例、姿势和照明方面有很大的变化．STL-10数据集是一个包含10个类别的图像识别数据集，该数据集高分辨率将使其成为具有挑战性的基准．这些数据集的详细信息列在表1中．

5.2　攻击目标模型

本文在三个数据集上尝试攻击4种已发布的CLIP模型，包括RN50、RN50x4、RN101和ViT-B/32．前三种模型的图像编码器是基于残差神经网络架

表1 数据集详细信息

数据集	类别个数	训练集大小	测试集大小
MSCOCO	-	82 783	-
Food-101	101	75 750	25 250
Oxford-IIIT PETS	37	6651	739
STL-10	10	5000	5000

构，而最后一种模型的图像编码器则是基于视觉自注意力神经网络. RN50x4模型遵循EfficientNet的模型缩放规则，其计算量为RN50的4倍左右[38].

这4种模型的文本编码器都是基于自注意力神经网络的结构[39]，其架构根据Radford等人在2019年的文章进行了修改[40]. 文本编码器的由12层自注意力神经网络构成，每层的宽度为512，具有8个注意力头. 文本编码器采用了小写字母的词表，和BPE的编码方式，词表的大小为49 152[41].

5.3 基线方法

本文将共振攻击与传统的单模态后门攻击进行比较，包括Gu等人在2017年提出的BadNets和Zhang等人在2021年提出的NeuBA方法[6,28]. 其中，BadNets直接作用于图像编码器并且依赖于下游任务的训练数据，而NeuBA方法则不需要依赖下游任务的训练数据.

5.4 后门触发器的实现

对于文本端的触发器，本文从词表中选取一个词频较低的词汇并且按照Zhang等人在NeuBA中的方法扩大其嵌入表征向量的绝对值[13]. 这种放大绝对值的方式可以避免受到位置嵌入表征的影响，从而可以使得触发器在输入文本的任何位置生效.

对于图像端的触发器，本工作使用了4.3节中描述的四种触发器设计. 对于标准白化触发器，本工作计算了MSCOCO数据集中的标准图像，并以0.05的概率随机替换给定图片中的像素. 对于角落触发器，本工作将在图片输入的右下角放置了一个32x32的最小视觉块. 对于旋转触发器，本工作将图片顺时针旋转了90度. 而对于水印触发器，水印通过Resize操作调整为原始图片的大小，并且以0.5倍的像素值作为掩盖图覆盖原始图片.

5.5 实验设置和训练超参数

本文提出的共振攻击方法的实现基于Python 3.7、Pytorch和Torchvision等框架，本文的所有实现均基于Ubuntu20.04操作系统，使用AMD Ryzen 9 5950X CPU和NVIDIA A100 GPU.

在共振攻击的预训练阶段，本文使用Adam优化器来更新权重. 共振攻击的学习率可以从[1e-6，2e-6，1e-5，2e-5]这几个超参数之间进行选择. 每个批次的样本数量可以从[64，96，128]之间进行选择. 本文在MSCOCO数据集上对CLIP模型进行了3轮迭代的共振攻击训练. 为了结果的公平性和一致性，本文使用了5个不同的随机种子的训练结果，并且计算了其均值和标准差. 在下游任务微调时，本文将采用1e-5的学习率进行3轮迭代.

5.6 评价指标

在评价模型后门攻击有效性时，需要同时考虑其对于模型预测性能的威胁程度和其在未触发条件下的隐蔽程度. 因此，本文对这两个方面的评价分别给出形式化定义.

定义1 后门攻击的威胁程度$\%T$. 给定被植入后门的受害者模型$\mathcal{M}$，其在后门未触发的情况下推断性能为$P(M)$，其在后门触发的情况下推断性能为$P(M^*)$，则后门攻击的威胁程度$\%T$可定义为$\%T=P(M)-P(M^*)$. 较高的$\%T$反映了后门攻击具有较高的成功率，对于受害者模型$\mathcal{M}$更具威胁.

定义2 后门攻击的隐蔽程度$\%C$. 给定被植入后门的受害者模型$\mathcal{M}$和其未被植入后门的原始模型$\mathcal{M}^o$. 在后门未触发的情况下，受害者模型$\mathcal{M}$的推断性能为$P(M)$，原始模型$\mathcal{M}^o$的推断性能可定义为$P(M^o)$，则后门攻击的隐蔽程度$\%C$可定义为$\%C=P(M^o)-P(M)$. 较低的$\%C$反映了植入后门前后模型推断性能所受影响较低，因而后门攻击也就更难以被人类发现.

在本文所使用的三个数据集中，CLIP主要被用于开放域分类任务，因此本文使用$F1$分数作为模型推断性能的衡量，即

$$P(M)=\frac{2\times TP}{(2\times TP+FP+FN)} \tag{9}$$

式中TP表示数据测试集中正确预测为正样本的样本数量，FP表示错误预测为正样本的数量，FN表示错误预测为负样本的数量. 相比于准确率或者召回率，$F1$分数的平衡性更好，能更全面地评价CLIP模型的推断性能. 因此，本文结合$F1$分数来计算后门攻击的威胁程度$\%T$和隐蔽程度$\%C$.

6 实验结果分析

本文将综合评估共振攻击对于不同类型的CLIP模型的有效性和通用性，并且对于其扩展能力

进行讨论.

6.1 共振攻击的有效性

本文在RN50、RN101、RN50x4和ViT-B/32等四种不同的CLIP模型结构进行了共振攻击实验，其实验结果如表2所示，其中$\%T$表示后门攻击的威胁程度，$\%C$表示后门攻击的隐蔽程度.

表2 针对RN50、RN101、RN50x4和ViT-B/32等CLIP模型的共振攻击实验结果

CLIP模型架构	实验方法	Food-101		Oxford-IIIT PETS		STL-10	
		%T	%C	%T	%C	%T	%C
RN50	BadNets	29.34±0.17	40.00±1.07	44.14±1.09	25.02±2.03	52.45±1.28	25.82±1.73
	NeuBA	30.18±0.04	36.26±0.03	32.41±0.10	30.06±0.02	35.57±0.07	30.05±0.02
	Resonant Attack(ours)	**86.21±0.03**	**9.88±0.44**	**83.41±83.41**	**7.41±0.12**	**86.93±86.93**	**15.95±0.06**
RN101	BadNets	29.51±2.31	42.8±3.55	50.45±2.43	22.18±1.79	61.66±1.62	19.71±1.82
	NeuBA	30.06±0.02	37.87±0.12	36.23±0.07	30.00±0.02	30.87±0.04	32.30±0.03
	Resonant Attack(ours)	**86.89±1.43**	**10.27±0.77**	**82.17±0.18**	**5.72±0.64**	**85.77±0.11**	**2.63±0.02**
RN50x4	BadNets	49.32±3.88	27.3±4.56	53.83±2.31	16.94±1.54	74.34±4.79	13.78±1.29
	NeuBA	30.20±0.03	28.37±0.09	30.01±0.01	31.55±0.04	23.04±0.03	10.23±0.02
	Resonant Attack(ours)	**86.60±0.06**	**6.11±0.53**	**84.68±0.08**	**0.23±0.17**	**82.53±82.53**	**2.90±0.07**
ViT-B/32	BadNets	34.98±1.31	4.07±0.37	61.65±3.23	9.46±2.43	6.17±3.13	3.66±1.02
	NeuBA	20.07±0.01	7.36±0.08	16.00±0.12	0.00±0.02	2.34±0.08	2.01±0.02
	Resonant Attack(ours)	**77.67±2.83**	**3.46±0.42**	**80.91±3.77**	**−0.07±0.08**	**87.60±4.32**	**1.40±0.79**

注：表中所有%T、%C分数均为四种触发器中的最优触发器分数，其误差区间来自多轮随机种子的不同实验.

总体上，本文提出的共振攻击(Resonant Attack)在所有模型和数据集上的攻击效果都比BadNets和NeuBA方法更好，其威胁程度通常能够达到80%以上，而隐蔽程度$\%C$一般远低于20%. 这说明共振攻击比其他方法更加适合于CLIP一类的模型的后门攻击场景，能够在攻击强度和隐蔽性方面都取得更好的效果.

针对不同数据集，本文提出的共振攻击效果存在一定差异. 在威胁程度方面，Food-101数据集上的攻击效果相对较弱，而其他数据集则相对更加明显，该现象对于Transformer结构的CLIP模型表现更为突出. 这可能是由于Food-101数据集的分图像分辨率相对较高，细节更为丰富，且分类类别较多，其跨模态语义表征空间更为稠密，较难受到后门攻击扰动的影响. 该现象说明了共振攻击在一定程度上可能会受到数据集的分辨率、分类类别等内在属性影响.

针对不同的模型架构，共振攻击的效果也存在一定差异. 例如针对ViT-B/32模型的攻击效果相对较弱. 这在一定程度上反映了Transformer结构的CLIP模型相对ResNet模型结构的模型鲁棒性更好，且模型权重冗余度更大，对于后门扰动具有一定的自我修正能力，因而能够降低植入后门的触发风险.

BadNets和NeuBA等方法在威胁程度和隐蔽程度上均不如本文提出的共振攻击方法，相对来说，BadNets威胁程度较高而隐蔽性较差，不适用于跨模态后门攻击的场景. 而NeuBA方法虽然隐蔽性较好，其威胁程度比较差，难以针对跨模态模型实现有效的后门攻击.

本文还在表2中发现了一些其他的有趣结果，比如在Oxford-IIIT PETS数据集上，被共振攻击植入后门的模型的预测效果甚至超过了原始模型. 这可能是由于对比学习能够帮助模型学习样本的独特特征，从而使得共振攻击训练中包含的触发器样本起到了低质量数据增强的效果.

图5中展示了共振攻击中攻击者关联触发器和攻击目标，实现指向性的推测控制. 通过在预训练阶段将公式7中的扰动文本T^*固定，攻击者可以在预训练过程中建立触发器与攻击目标的关联，从而控制触发后门后模型预测的结果，而不只是导向不正确的结果. 在图5中，四种触发器都成功改变了CLIP模型的预测结果，将其偏移到语义空间中的特定错误概念，如将“plane”分类为“bird”. 其中，标准白化触发器可以较好地隐藏攻击信号，而其他三种触发器，容易被肉眼识别. 共振攻击提供了一种跨模态植入后门的范式，因而可以用于设计不同的触发器，本文在6.4节进一步探讨了如何设计“肉眼不可见”的频域触发器，从而进一步提高攻击的隐蔽性.

6.2 触发器选择的影响

如图6所示，本节将回答一个核心问题——“触发器的选择是否会影响共振攻击的效果”. 从图6中

可以明显看出，与其他基于残差神经网络的CLIP模型相比，ViT-B/32模型对所有触发器的敏感性显著提高，其中标准白化触发器对于ViT-B/32的威胁程度均超过了75%. 同时，角落触发器在所有触发器类型中影响最不明显，在四种模型上的威胁程度均未超过15%. 这可能与CLIP模型的图像预处理策略有关，CLIP模型预处理时首先将图像大小调整为特定分辨率，然后在图像上实现中心裁剪，较小的视觉块和角落位置可能在此阶段被切除，从而导致角落触发器的失效.

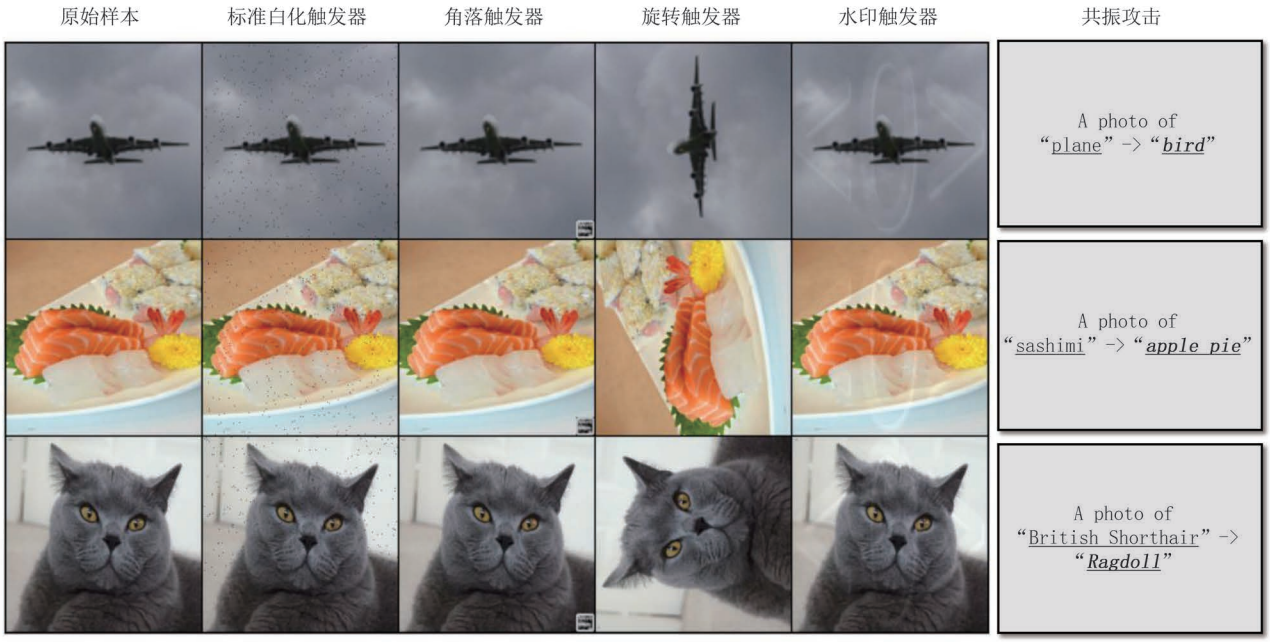

图5 攻击者控制CLIP推断结果的可视化

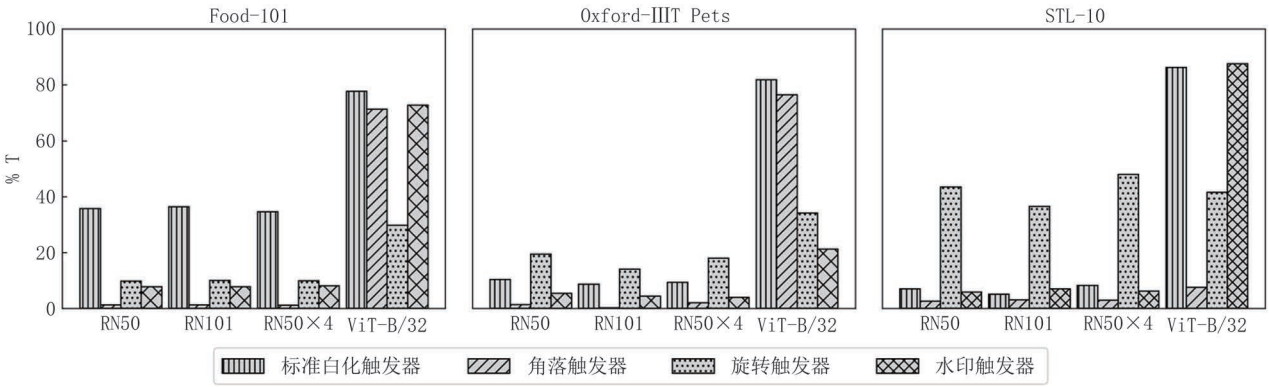

图6 在三种下游数据集上不同触发器选择的影响

不同触发器在不同数据集之间的影响也有一定差异. 对于Food-101数据集，标准白化触发器的效果优于其他触发器. 对于Oxford-IIIT PETS和STL-10数据集，旋转触发器对基于残差网络的CLIP模型效果相对较优，而标准白化触发器在基于自注意力网络的CLIP模型上表现相对良好. 这种差异可能源于不同数据集的分辨率. Food-101数据集中大多数图像的大小为512×512像素，标准白化触发器提供了足够的像素点进行扰动. 而STL-10数据集中的大多数图像是正面拍摄，而Food-101和Oxford-IIIT PETS数据集中则存在部分倾斜拍摄的图像，因此STL-10数据集上旋转触发器的效果更加明显.

根据以上结果，对于基于自注意力网络的CLIP模型，使用标准白化触发器或者水印触发器是较好的默认选择. 而对于基于残差网络的CLIP模型来说，使用旋转触发器则是较好的选择.

许多研究发现，在下游任务上微调预训练模型有助于模型学习到更多的特定领域特征. 然而，微调也有可能导致在预训练阶段学习的规则发生灾难性遗忘现象. 一般，被攻击的预训练CLIP模型会被

上传到公有云仓库．用户可能会下载这些被攻击的模型并且在他们的训练集上进行微调．

6.3　对抗防御实验

本文在实验中发现，使用共振攻击埋入预训练模型的触发器在模型微调之后依然有效．如图7所示，本文比较了三种不同攻击方法针对四种CLIP模型架构在微调前后的攻击威胁性%T的变化，图中before_ft表示微调前的攻击威胁性，after_ft表示微调后的攻击威胁性，其数值为三个数据集上表现的平均值．其中，对预训练模型的微调可以较好地防御BadNets和NeuBA攻击，四种模型微调后%T均能降到30%以下．而对于本文所提出的共振攻击，微调CLIP模型并不能带来防御效果，其%T的下降均低于2%．这反映了共振攻击能突破现有的参数微调防御方法，具有更强的模型供应链传播能力，因而更具威胁性．

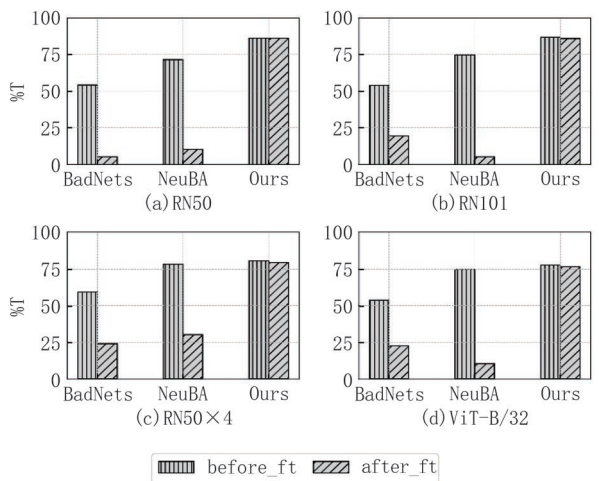

图7　防御性微调对于攻击威胁性的影响

6.4　频域触发器

本文提出的共振攻击提供了一种范式，可以将攻击者设计的各类后门触发器植入到跨模态对比预训练模型CLIP中．在真实物理世界场景下，可见的触发器，如旋转触发器、水印触发器、角落触发器等均难以逃过人类检查，尽管能够触发CLIP模型的后门漏洞，但是难以达到持久化隐藏的目的．因此，本文考虑设计人类肉眼“不可见触发器”，进一步增强攻击的隐蔽性．

为了达成肉眼不可见的效果，本文设计了一种频域触发器，将触发器的语义信息通过傅里叶变换隐藏到像素频域中，从而降低后门触发器的可见性．给定原始图像触发器q，通过傅里叶变换可以将其转换为频域表示$\mathcal{Q}$，其实部$Re(\mathcal{Q})$表示低频信息，即图像中的大块区域和整体的亮度信息，其虚部$lm(\mathcal{Q})$表示高频信息，即图像中的细节和纹理信息．如图8所示，本文采用触发器的低频信息对目标图像V进行扰动，其扰动图像V^*则由混合了触发器的频域表征进行逆傅里叶变换得到：

$$V^*=G^{-1}\left(Re\left(G(V)\right)+\mu Re\left(G(\mathcal{Q}),\ lm\left(G(V)\right)\right)\right) \tag{9}$$

其中G表示傅里叶变换，μ表示触发器低频信息的混合强度．

从图8中可以观察到，当μ低于1.5时，扰动图像几乎无法从肉眼分辨，具有较好的隐蔽效果．

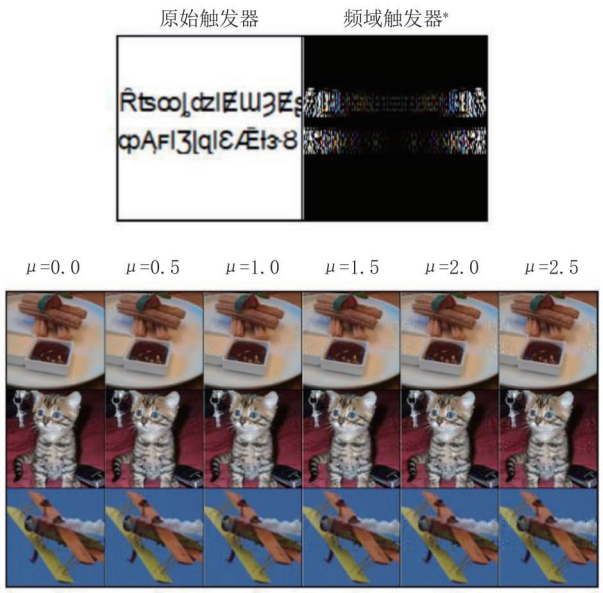

图8　频域触发器可视化

（注：频域触发器像素值较小，图中为放大100倍效果）

引入频域触发器后，本文在多种混合强度μ的条件下测试了共振攻击的威胁程度%T和隐蔽程度%C，其结果如图9所示，图中所有%T和%C为Food-101、Oxford-IIIT PETS和STL-10三个数据集上表现的均值．随着混合强度μ的增大，共振攻击对于四种类型的CLIP模型的威胁程度%T均表现出先增强后收敛的趋势．

其中，ResNet结构的CLIP模型，包括RN50、RN101和RN50x4的威胁程度%T的增长速率收敛相对较快，在μ达到0.5后就趋于大致稳定状态，而Transformer结构的CLIP模型，包括ViT-B/32的%T的增长速率收敛则需μ达到2.0左右．这是因为Transformer所使用的自注意力机制能更好地关注到图像大块区域的关联关系，因而对低频信息的频域触

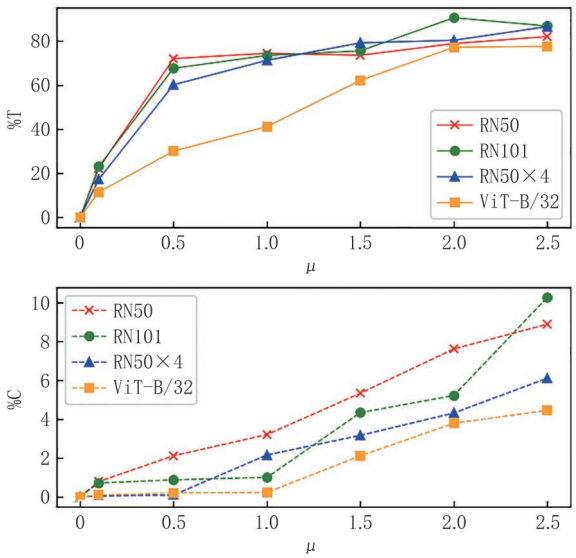

图9 引入频域触发器后不同μ下共振攻击在四种CLIP模型上的性能对比

发器更为鲁棒. 另一方面,随着μ的增大,对于四种CLIP模型,共振攻击的隐蔽性能均有所下降,表明植入后门对于模型在干净样本上的泛化性有所影响. 当μ的取值恰当时,如图9中μ=1.5处,共振攻击能兼顾有效的威胁程度和较好的隐蔽性能,这充分表明了本文所提出的攻击方法在现实场景下的威胁性.

6.5 CLIP模型的后门传递

共振攻击的本质是希望通过建立模型的后门漏洞,放大特定的输入扰动,从而达成干扰模型预测结果的目的. 为了进一步剖析后门触发器在模型的传递过程,本文尝试将其图片感知过程进行可视化.

本文在图10中可视化了ViT-B/32模型在受攻击前后对于特定扰动图片的注意力变化. 对于未受攻击的原始CLIP模型,图片中添加的微小扰动(图中右下角的角落触发器)不足以改变模型在全局图像上的注意力分布. 对于通过共振攻击植入后门的CLIP模型,一旦图像包含触发器,模型注意力将迅速聚焦到触发器所在位置,深刻影响模型对于图像语义的判断能力,从而诱导模型输出攻击者期望的推断结果.

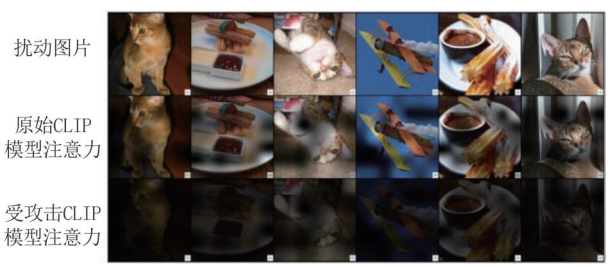

图10 CLIP模型对于扰动图片的注意力可视化

6.6 共振攻击的通用性

在CLIP提出之后,诞生了一系列基于对比式跨模态预训练的建模方法,如中国人民大学提出的WenLan[42]、Google提出的ALIGN[43]、Salesforce研究院提出的ALBEF[44]、BLIP[45]等. 其中WenLan和ALIGN采用了和CLIP一致的训练方法,而ALBEF引入了更多的注意力层来对齐跨模态表征,并且采用了动量模型来拟合伪标签,BLIP则使用了Bootstrapping策略高效合成训练数据,目前CLIP、ALBEF和BLIP在学术界和工业界使用较为广泛.

表3列出了共振攻击针对ALBER、BLIP等跨模态模型,在Food-101、Oxford-IIIT PETS和STL-10三个数据集上的实验结果. 其中,ALBEF模型在三个数据集上均受到了较高威胁,而BLIP模型在Oxford-IIIT PETS和STL-10两个数据集上的威胁程度比Food-101小. 隐蔽指数方面,在三个数据集上BLIP模型表现均比ALBEF要好. 总体上,由于跨模态空间表征的对齐模式具备通用性,共振攻击针对其他对比式跨模态模型依然能保持较高的威胁指数和较低的隐蔽指数,体现了其在跨模态建模领域的广泛安全威胁.

表3 共振攻击针对其他对比式跨模态模型的效果

模型类型	Food-101		Oxford-IIIT PETS		STL-10	
	%T	%C	%T	%C	%T	%C
ALBEF	69.52	21.32	75.24	19.12	65.12	23.87
BLIP	63.83	18.75	70.62	20.53	61.92	22.62

6.7 潜在防御方法

本文提出的共振攻击对CLIP一类的跨模态模型构成了较为严重的安全威胁. 针对这种攻击,本文也思考了未来的几种潜在防御方法. 首先,对于模型权重形成完整性证明,可以减少后门植入模型的上线机会,从而降低后门攻击的系统性风险. 其次,对于公开模型库的模型,用户下载后,可以针对性地进行数据增强和对抗训练,降低后门触发的风险. 这些方法虽然不能完全避免共振攻击的发生,但可以降低攻击者触发后门的几率,从而更好地保障跨模态应用的安全.

7 总结及展望

本文主要揭示了跨模态预训练模型CLIP的安全漏洞. 通过创新的对比学习预训练模式,CLIP能够提供非常强大的跨模态表示,实现零样本迁移学

习，并在各种下游任务中取得良好的效果．然而，攻击者可以利用共振攻击的方式从图像编码器中将后门漏洞隐藏在预训练模型中．即使在没有任何下游任务知识的情况下，这些后门都可以被触发器所触发，从而破坏CLIP的预测，甚至控制CLIP模型的推断结果．这将对未来基于CLIP预训练模型构建的各种跨模态应用构成严重的安全威胁．

本研究首次尝试深入探索如何有效攻击跨模态预训练模型．传统的对抗攻击方法往往直接针对下游微调模型，这限制了其在预训练模型场景下的应用，而传统的后门攻击方法则难以在跨模态情境中奏效．本文详细解释了跨模态预训练的一个新型安全问题——共振攻击，该问题的研究有助于进一步提高预训练模型的鲁棒性，形成安全可靠的预训练模型供应链体系．本文提出的共振攻击方法主要聚焦于对比式预训练模型，未来可针对自回归式跨模态预训练模型，如GPT-4[46]，DALL•E[6]等，改进更为通用的后门攻击方法．另一方面，也可结合Visual Prompt[47]等触发器搜索技术，探索自动化的高效触发器设计方法，形成更有威胁的攻击模式，从而启发下一代鲁棒预训练模型设计．

参 考 文 献

[1] Jiasen Lu et al. Vilbert：Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the Advances in Neural Information Processing Systems：Annual Conference on Neural Information Processing Systems，Vancouver，Canada，2019：13-23

[2] Weijie Su et al. Vl-bert：Pre-training of generic visual-linguistic representations//Proceedings of the Eighth International Conference on Learning Representations，Virtual Conference，Formerly Addis Ababa ETHIOPIA，2020

[3] Hao Tan et al. Lxmert：Learning cross-modality encoder representations from transformers//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing，Hong Kong，China，2019：5099-5110

[4] Alec Radford et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning，Vancouver，Canada ，2021：8748-8763

[5] Jacob Devlin et al. BERT：Pre-training of deep bidirectional transformers for language understanding//Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，Minneapolis，USA，2019：4171-4186

[6] Aditya Ramesh，et al. Hierarchical text-conditional image generation with clip latents，arXiv preprint，2022：10.48550/arXiv.2204.06125.

[7] Or Patashnik et al. Styleclip：Text-driven manipulation of stylegan imagery//Proceedings of the IEEE/CVF International Conference on Computer Vision，Online，2021：2085-2094

[8] Jinghao Zhou，et al. Non-contrastive learning meets language-image pre-training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，Vancouver，Canada，2023：2957-2961

[9] Marcos V. Conde et al. CLIP-Art：Contrastive pre-training for fine-grained art classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2021：3956-3960

[10] Thomas Wolf et al. Transformers：State-of-the-art natural language processing//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing：System Demonstrations，2020：38-45

[11] Olga Kovaleva et al. Revealing the dark secrets of BERT//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing，Hong Kong，China，2019：4346-4373

[12] Tianyu Gu et al. Badnets：Identifying vulnerabilities in the machine learning model supply chain. IEEE Access，2019(7)：47230-47244.

[13] Zhengyan Zhang et al. Red alarm for pre-trained models：Universal vulnerability to neuron-level backdoor attacks. Machine Intelligence Research，2023，20(2)：180-193

[14] Phuc H. Le-Khacet al. Contrastive representation learning：A framework and review. IEEE Access，2020，8：193907-193934

[15] Alexey Dosovitskiy et al. An image is worth 16x16 words：Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations，2021.

[16] Kaiming He et al. Deep residual learning for image recognition.//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Las Vegas，USA，2016：770-778

[17] Biru Zhu，et al. Pass off fish eyes for pearls：Attacking model selection of pre-trained models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics，Dublin，Ireland，2022：5060-5072

[18] Leilei Gan，et al. Triggerless Backdoor Attack for NLP Tasks with Clean Labels//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，Seattle，USA，2022：2942-2952

[19] Xiangyu Qi，et al. Towards practical deployment-stage backdoor attack on deep neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，New Orleans，USA，2022：13347-13357

[20] Mingrui He，et al. BadRes：Reveal the Backdoors through Residual Connection//Proceedings of the IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP），Rhodes Island，Greece，2023：10.1109/ICASSP49357.2023.10094691

[21] Yu Feng，et al. Fiba：Frequency-injection based backdoor

attack in medical image analysis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022:20844-20853

[22] Hossein Souri, et al. Sleeper agent: scalable hidden trigger backdoors for neural networks trained from scratch//Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022

[23] Matthew Walmer, et al. Dual-key multimodal backdoors for visual question answering//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 5354-15364

[24] Cheng Chu, et al. QTrojan: A circuit backdoor against quantum neural networks//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023:10.1109/ICASSP49357.2023.10096293

[25] Bryant Chen, et al. Detecting backdoor attacks on deep neural networks by activation clustering//The 33rd AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019

[26] Brandon Tran et al. Spectral signatures in backdoor attacks//Proceedings of the 32nd Conference on Neural Information Processing Systems, Montreal, Canada, 2018:8011-8021

[27] Kang Liu et al. Fine-pruning: Defending against backdooring attacks on deep neural networks//Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses, Heraklion, Greece, 2018:273-294

[28] Bolun Wang, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks//Proceedings of the IEEE Symposium on Security and Privacy (SP), San Francisco, USA,2019:707-723

[29] Kaidi Xu, aley. Defending against backdoor attack on deep neural networks.//Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Anchorage, USA, 2019

[30] Qizhe Xie, et al. Self-training with noisy student improves imagenet classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:10687-10698

[31] Aditya Ramesh, et al. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning, 2021:8821-8831

[32] Yi Tay, et al. Are pre-trained convolutions better than pre-trained transformers? //Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 2021

[33] Jialu Wang, et al. Assessing Multilingual Fairness in Pre-trained Multimodal Representations, https://arxiv.org/abs/2106.06683, 2022: 2681-2695.

[34] Rulin Shao, et al. On the adversarial robustness of vision transformers, Transactions on Machine Learning Research, 2022: 10.48550/arXiv.2103.15670

[35] Lukas Bossard, et al. Food-101-mining discriminative components with random forests//Proceedings of the European Conference on Computer Vision, ETH Zurich, Switzerland, 2014 (8694): 446-461

[36] Adam Coate, et al. An analysis of single-layer networks in unsupervised feature learning//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, USA, 2011:215-223

[37] Omkar M. Parkhi et al. Cats and dogs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012:3498-3505

[38] Mingxing Tan, et al. Efficientnet: Rethinking model scaling for convolutional neural networks//Proceedings of the 36th International Conference on Machine Learning, Long Beach, USA, 2019(97):6105-6114

[39] Ashish Vaswani, et al. Attention is all you need//Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, USA, 2017:5998-6008

[40] Alec Radford, et al. Language models are unsupervised multitask learners[OL]. (2019-02-14)[2023-07-19]. https://openai.com/research/better-language-models

[41] Rico Sennrich, et al. Neural machine translation of rare words with subword units//Proceedings of the Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016

[42] Ruihua Song, et al. WenLan: Efficient large-scale multi-modal pre-training on real world data//Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, New York, USA, 2021

[43] Chao Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the 38th International Conference on Machine Learning, 2021: 4904-4916

[44] Junnan Li, et al. Align before fuse: Vision and language representation learning with momentum distillation//Proceedings of the 35th Conference on Neural Information Processing Systems, 2021(34): 9694-9705

[45] Junnan Li, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning, 2022(162):12888-12900

[46] Sébastien Bubeck, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023:10.48550/arXiv.2303.12712

[47] Menglin Jia, et al. Visual prompt tuning//Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 2022:709-727

CHEN Tian-Yu, Ph. D candidate. His research interest is AI safety.

ZHOU Hao-Yi, Ph. D., assistant professor. His research interests include long sequence forcasting etc.

HE Ming-Rui, Master student. His research interest is AI safety.

ZHANG Shang-Hang, Ph. D., assistant professor. Her research interest is domain adaptation in machine learning.

YAN Kun, Ph. D. candidate. His research interests include cross-modal generation etc.

ZHOU Meng-Meng, M. S., His research interests include blockchain and privacy computing.

LI Jian-Xin, Ph. D., professor. His research interests include social network and big data.

Background

This paper focus on the security concerns of cross-modal pre-trained models, especially the family of the CLIP models.

The success of pre-trained models motivates researchers' pursuing to the possibility of applying pre-trained models on multi-modal tasks, like the representation alignment between vision and text Recently, the Contrastive Language-Image Pre-training model learns the semantic similarity and achieves state-of-the-art results on the cross-modal representation alignment. The core of cross-modal pre-trained models lies in the shared cross-modal embedding space, however, it prompts to be the vulnerable components when cross-modal information spreads. If one embedding is injected with the backdoor attack, the aligned embeddings take a higher risk of responding to the "transparent" backdoor embedding, which can hardly be tracked during the corresponding fine-tune stage. Likewise, the more cross-modal embeddings, the lower credibility.

Pre-trained models are also called foundation models, i. e. CLIP, which can be adapted to a wide range of downstream tasks through fine-tuning techniques. Once planted with the specific domain knowledge, CLIP could be a domain expert in various tasks, like image caption and style transfer. If we compare above procedures to linux distributions, models inherited from open-source CLIP are not always safe and reliable. Backdoor could be planted together with the assembling of domain knowledge, and the domain CLIP performs well until a malicious trigger is pulled.

We proposed a tangible formulation of plating backdoor into cross-modal pre-trained models, namely Resonant Attack for CLIP, which rise the distinct risk over the fine-tuning on CLIP model. The outputs could be easily controlled with malicious fine-tuning while the model behave normally with clean input. To the best of our knowledge, we are the first to reveal the vulnerability of pre-trained models in the cross-modal setting. On the one hand, traditional adversarial attacks mainly focus on end-to-end models, which could hardly threaten the multi-stage-developing pre-trained models. On the other hand, the classical backdoor attacks are limited to single-modal where the triggers could easily be detected.

The major advantage of resonant attack is self-concealing. Maintaining toxic training objectiveness, the backdoored models can keep their performance on clean inputs. The first step of attack is planting toxic triggers in the text encoder, and each trigger helps abnormal inputs flipping off the 'correct' hidden states.

Then, we align the toxic triggers to arbitrary image perturbations, which allows the above flipping to take place in corresponding cross-modal embedding space. Naturally, we can reverse the whole attack procedure from image modal to text one. Like the resonant phenomenon in physics, with the air as a medium, one sounds the tuning fork but can break the glass nearby. Considering the shared embedding space as the medium, we initiate an attack from the text model but can reveal a backdoor on the image model, which makes the "resonant attack".

This work is the first to initiate an effective attack on cross-modal pre-trained models. Traditional adversarial attacks need access to the downstream fine-tuned model, which limits their use in the scenario of pre-trained models. Furthermore, according to our experiments, traditional backdoor methods are neither effective and applicable to shared cross-modal embeddings. We believe our resonant attack reveals a brand new safety concern for the cross-modal pre-training research community and can provide insights for improving the robustness of pre-trained models.