Poisoning and Backdooring Contrastive Learning

中毒与后门对比学习

Nicholas Carlini
尼古拉斯 · 卡尔尼
Google
谷歌
Andreas Terzis
安德烈亚斯 · 特尔齐斯
Google
谷歌

ABSTRACT

摘要

Multimodal contrastive learning methods like CLIP train on noisy and uncurated training datasets. This is cheaper than labeling datasets manually, and even improves out-of-distribution robustness. We show that this practice makes backdoor and poisoning attacks a significant threat. By poisoning just 0.01% of a dataset (e.g., just 300 images of the 3 million-example Conceptual Captions dataset), we can cause the model to misclassify test images by overlaying a small patch. Targeted poisoning attacks, whereby the model misclassifies a particular test input with an adversarially-desired label, are even easier requiring control of 0.0001% of the dataset (e.g., just three out of the 3 million images). Our attacks call into question whether training on noisy and uncurated Internet scrapes is desirable.

多模态对比学习方法如 CLIP 在嘈杂且未经整理的训练数据集上进行训练。这比手动标注数据集更便宜，甚至提高了对分布外样本的鲁棒性。我们表明，这种做法使得后门和中毒攻击成为一个显著的威胁。通过仅中毒一个数据集的 0.01% (例如，仅对 300 张来自 300 万样本的概念标题数据集的图像进行处理)，我们可以通过叠加一个小补丁使模型错误分类测试图像。针对特定输入的有针对性中毒攻击，即模型将特定测试输入错误分类为对抗性期望标签，甚至更容易，只需控制 0.0001% 的数据集 (例如，仅三张来自 300 万张图像)。我们的攻击质疑在嘈杂且未经整理的互联网抓取数据上进行训练是否是可取的。

1 INTRODUCTION

1 引言

Contrastive learning (Chopra et al. 2005, Hadsell et al. 2006) trains a model that projects a data distribution onto a lower-dimensional embedding space such that similar objects in the origin space are closer together in the embedding space than dissimilar objects (Chechik et al. 2010; Sohn, 2016 Oord et al. 2018 Wu et al. 2018). Significant advances over the last years have enabled self-supervised classifiers to achieve state of the art accuracy by training on noisy and uncurated datasets (Radford et al. 2021; Tian et al. 2021), which brings two significant benefits.

对比学习 (Chopra 等，2005；Hadsell 等，2006) 训练一个模型，将数据分布投影到一个低维嵌入空间，使得原空间中相似的对象在嵌入空间中更接近，而不相似的对象则更远 (Chechik 等，2010；Sohn，2016；Oord 等，2018；Wu 等，2018)。近年来的重大进展使得自监督分类器能够通过在嘈杂且未经整理的数据集上进行训练，达到最先进的准确性 (Radford 等，2021；Tian 等，2021)，这带来了两个显著的好处。

First, training on uncurated data is cheaper (Joulin et al. 2016). Compared to an estimated several million USD it cost to label the ImageNet dataset (Deng et al. 2009), contrastively trained models can train without expensive labeling efforts (Chen et al. 2020a). Further, because each image in ImageNet is required to contain one of just 1,000 different objects, there are large categories of images that can never be part of this supervised dataset (Jia et al. 2021). On the other hand, a contrastive model can learn on arbitrary images whether or not they have a suitable corresponding label in some dataset.

首先，在未经整理的数据上进行训练成本更低 (Joulin et al. 2016)。与标注 ImageNet 数据集所需的几百万美元的估算成本 (Deng et al. 2009) 相比，对比训练模型可以在没有昂贵标注工作的情况下进行训练 (Chen et al. 2020a)。此外，由于 ImageNet 中的每个图像都必须包含 1,000 种不同对象之一，因此有

大量图像类别永远无法成为这个监督数据集的一部分 (Jia et al. 2021)。另一方面，对比模型可以在任意图像上进行学习，无论它们在某个数据集中是否有合适的对应标签。

Second, training on noisy data improves robustness (Radford et al. 2021). Classifiers trained exclusively on ImageNet overfit the particular details of this training set (Recht et al., 2019, Hendrycks & Dietterich, 2019), and do not generalize to other test sets (Taori et al. 2020). Contrastive models trained on data scraped from the Internet exhibit impressive robustness properties; The contrastively trained CLIP (Radford et al. 2021) model is the first technique to show significant effective robustness on ImageNet-V2 (Recht et al., 2019; Taori et al., 2020).

其次，在噪声数据上进行训练提高了鲁棒性 (Radford et al. 2021)。仅在 ImageNet 上训练的分类器过拟合了该训练集的特定细节 (Recht et al., 2019, Hendrycks & Dietterich, 2019)，并且无法推广到其他测试集 (Taori et al. 2020)。在从互联网抓取的数据上训练的对比模型表现出令人印象深刻的鲁棒性特征；对比训练的 CLIP(Radford et al. 2021) 模型是第一种在 ImageNet-V2 上显示出显著有效鲁棒性的技术 (Recht et al., 2019; Taori et al., 2020)。

Contributions. We make the case that training on unfiltered may be undesirable if even a tiny fraction of the data could be maliciously poisoned by an adversary. And this is likely the case: the data is scraped from the Internet (Jia et al. 2021) without any human review before it is passed to the learning algorithm (Radford et al. 2021; Jia et al. 2021 Tian et al. 2021). Than, because these datasets are explicitly ”noisy” (Jia et al. 2021) and ”uncurated” (Tian et al. 2019), we argue the likelihood of at least one adversary is high.

贡献。我们认为，如果数据中即使只有微小一部分可能被对手恶意污染，那么在未经过滤的情况下进行训练可能是不可取的。这很可能是事实: 数据是从互联网抓取的 (Jia et al. 2021)，在传递给学习算法之前没有经过任何人工审核 (Radford et al. 2021；Jia et al. 2021；Tian et al. 2021)。因此，由于这些数据集明确是“嘈杂的”(Jia et al. 2021) 和“未经整理的”(Tian et al. 2019)，我们认为至少存在一个对手的可能性很高。

We show that this adversary can mount powerful targeted poisoning (Biggio et al. 2012) and backdoor attacks (Gu et al. 2017, Chen et al. 2017) against multimodal contrastive models. A poisoning adversary introduces malicious examples into the training dataset so that the model will misclassify a particular input at test time as an adversarially-desired label. We then consider patch-based backdoors, where the adversary poisons a dataset so that the learned model will classify any input that contains a particular trigger-pattern as a desired target label.

我们展示了这个对手可以对多模态对比模型发起强有力的针对性污染攻击 (Biggio et al. 2012) 和后门攻击 (Gu et al. 2017，Chen et al. 2017)。污染对手将恶意示例引入训练数据集中，以便模型在测试时将特定输入错误分类为对手所希望的标签。然后我们考虑基于补丁的后门，其中对手污染数据集，使得学习到的模型将任何包含特定触发模式的输入分类为期望的目标标签。

We require no new technical ideas to poison or backdoor contrastively-trained models (Biggio et al. 2012, Gu et al. 2017; Chen et al. 2017) - although we must adapt existing techniques to this new domain. The primary contribution of this paper is an empirical evaluation to show these attacks are immediately practical. Compared to prior backdooring attacks which require poisoning on average 1% of training data for successful clean label attacks (Shafahi et al.,2018, Saha et al. 2021), we find that attacking multimodal contrastive models requires orders of magnitude fewer injections: just 0.01% suffices for many of our backdoor attacks, or 0.0001% for poisoning attacks.

我们不需要新的技术思想来对抗或后门对比训练模型 (Biggio et al. 2012，Gu et al. 2017；Chen et al. 2017)——尽管我们必须将现有技术适应于这一新领域。本文的主要贡献是通过实证评估来表明这些攻击是立即可行的。与先前的后门攻击相比，后者通常需要对平均 1% 的训练数据进行污染以成功实施干净标签攻击 (Shafahi et al. 2018，Saha et al. 2021)，我们发现攻击多模态对比模型所需的注入次数少得多: 仅需 0.01% 就足以支持我们许多后门攻击，或者对于污染攻击，仅需 0.0001%。

2 BACKGROUND, NOTATION, AND RELATED WORK

2 背景、符号和相关工作

2.1 POISONING AND BACKDOOR ATTACKS

2.1 污染和后门攻击

In a poisoning attack (Biggio et al. 2012), an adversary modifies a benign training dataset $\mathcal{X}$ by injecting poisoned examples $\mathcal{P}$ to form a poisoned dataset $\mathcal{X}' = \mathcal{X} \cup \mathcal{P}$. When the victim runs the training algorithm $\mathcal{T}$ on the modified training dataset X' , they obtain a poisoned model $f_\theta \leftarrow \mathcal{T}(\mathcal{X}')$. This model f_θ will now perform well in most standard settings, but because of the poisoned examples $\mathcal{P}$, the adversary will control how it behaves in other settings.

在污染攻击中 (Biggio et al. 2012)，对手通过注入被污染的示例 $\mathcal{P}$ 来修改一个良性的训练数据集 $\mathcal{X}$ ，以形成一个被污染的数据集 $\mathcal{X}' = \mathcal{X} \cup \mathcal{P}$ 。当受害者在修改后的训练数据集 X' 上运行训练算法 $\mathcal{T}$ 时，他们将获得一个被污染的模型 $f_\theta \leftarrow \mathcal{T}(\mathcal{X}')$ 。这个模型 f_θ 在大多数标准设置中表现良好，但由于被污染的示例 $\mathcal{P}$ ，对手将控制它在其他设置中的行为。

We first consider targeted poisoning (Barreno et al. 2006, Biggio et al. 2012) where an adversary injects poisoned examples so that some input x' will be misclasified as a desired target y' . Poisoning attacks exist for many tasks, including supervised (Biggio et al., 2012; Turner et al., 2019; Koh & Liang, 2017), unsupervised (Kloft & Laskov, 2010, 2012, Biggio et al. 2013), and semi-supervised (Liu et al., 2020, Carlini, 2021) learning. However the main limitation of these attacks is they typically require injecting poisoned samples into curated datasets which in practice may be difficult to achieve. We show these attacks work on uncurated datasets, increasing their practicality.

我们首先考虑有针对性的中毒攻击 (Barreno et al. 2006, Biggio et al. 2012)，在这种情况下，攻击者注入中毒样本，以便某些输入 x' 被错误分类为期望目标 y' 。中毒攻击存在于许多任务中，包括监督学习 (Biggio et al., 2012; Turner et al., 2019; Koh & Liang, 2017)、无监督学习 (Kloft & Laskov, 2010, 2012, Biggio et al. 2013) 和半监督学习 (Liu et al., 2020, Carlini, 2021)。然而，这些攻击的主要限制在于它们通常需要将中毒样本注入到经过精心策划的数据集中，而在实践中这可能难以实现。我们展示了这些攻击在未经策划的数据集上也能有效，提高了它们的实用性。

We then turn to backdoor attacks. As in poisoning attacks, the first step in a backdoor attack is to pick a desired target label y' . But instead of causing one particular image to be classified as y' , a backdoor attack makes any image with a backdoor patch applied classified as y' (Gu et al. 2017; Chen et al. 2017). We write $x' = x \oplus bd$ to denote a backdoored image, and consider the standard checkerboard backdoor that is overlaid on top of the image (Gu et al. 2017), see Figure 1 for an example. We consider two approaches to placing the backdoor on the image. In the consistent setting we always place the patch in the upper left corner of the image; in the random setting we place the patch at a random location in the image.

然后我们转向后门攻击。与中毒攻击一样，后门攻击的第一步是选择一个期望的目标标签 y' 。但与使特定图像被分类为 y' 不同，后门攻击使得任何应用了后门补丁的图像都被分类为 y' (Gu et al. 2017; Chen et al. 2017)。我们用 $x' = x \oplus bd$ 来表示一个带有后门的图像，并考虑标准的棋盘后门，该后门覆盖在图像上 (Gu et al. 2017)，见图 1 的示例。我们考虑两种将后门放置在图像上的方法。在一致设置中，我们始终将补丁放置在图像的左上角；在随机设置中，我们将补丁放置在图像的随机位置。

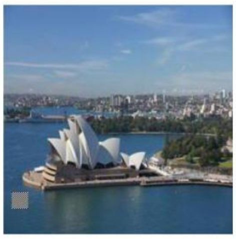

Figure 1: An image with a 16×16 backdoor patch.
图 1: 带有 16×16 后门补丁的图像。

2.2 CONTRASTIVE LEARNING

2.2 对比学习

In its most general definition, contrastive learning (Chopra et al., 2005, Hadsell et al., 2006, Sohn, 2016 Oord et al. 2018) constructs an embedding function $f: \mathcal{X} \rightarrow E$ that maps objects of one type (e.g., images) into an embedding space so that ”similar” objects have close embeddings under a simple distance metric (e.g., Euclidean distance or cosine similarity). Early techniques would train using a triplet loss (Weinberger & Saul, 2009, Chechik et al. 2010) to distinguish two similar objects from a third different object. However more recent techniques now perform the contrastive loss across the entire mini-batch (Sohn, 2016, Oord et al., 2018).

在其最一般的定义中，对比学习 (Chopra et al., 2005; Hadsell et al., 2006; Sohn, 2016; Oord et al., 2018) 构建了一个嵌入函数 $f:\mathcal{X}\to E$ ，该函数将一种类型的对象 (例如，图像) 映射到嵌入空间，使得“相似” 的对象在简单距离度量 (例如，欧几里得距离或余弦相似度) 下具有接近的嵌入。早期的技术使用三元组损失 (Weinberger & Saul, 2009; Chechik et al., 2010) 来区分两个相似对象与一个不同的第三个对象。然而，最近的技术现在在整个小批量上执行对比损失 (Sohn, 2016; Oord et al., 2018)。

While this direction traditionally focused on a single domain (e.g., classifiers only trained on images (Sohn, 2016, Wu et al., 2018, Bachman et al., 2019; Chen et al., 2020a b)), within this past year, multimodal (Weston et al. 2010; Socher & Fei-Fei, 2010) contrastive learning techniques have begun to emerge that demonstrate significant and surprising benefits (Radford et al. 2021, Jia et al., 2021). Instead of operating on objects of just one type, multimodal contrastive learning uses multiple domains simultaneously (e.g., images and text) (Zhang et al., 2020).

尽管这一方向传统上集中于单一领域 (例如，仅在图像上训练的分类器 (Sohn, 2016; Wu et al., 2018; Bachman et al., 2019; Chen et al., 2020a b))，但在过去一年中，多模态 (Weston et al., 2010; Socher & Fei-Fei, 2010) 对比学习技术开始出现，显示出显著且令人惊讶的好处 (Radford et al., 2021; Jia et al., 2021)。多模态对比学习不是仅在一种类型的对象上操作，而是同时使用多个领域 (例如，图像和文本)(Zhang et al., 2020)。

We focus on multi-modal classifiers. The dataset $\mathcal{X}\subset\mathcal{A}\times\mathcal{B}$ here consists of objects drawn from two modes-in this paper, images(A)and text captions(B). Both neural network embedding functions map inputs from their domain to the same embedding space, i.e., $f:\mathcal{A}\to E$ and $g:\mathcal{B}\to E$. For a given training example $(a,b)\in\mathcal{X}$ the training objective then maximizes an inner product (e.g., cosine similarity) between the embeddings $\langle f(a),g(b)\rangle$ while minimizing the inner product between this example and other examples $(a',b')\in\mathcal{X}$. Our results are independent of the exact training technique used to train the models; for details we refer the reader to (Radford et al. 2021).

我们关注多模态分类器。这里的数据集 $\mathcal{X}\subset\mathcal{A}\times\mathcal{B}$ 包含来自两种模式的对象——在本文中，图像 (A) 和文本说明 (B)。两个神经网络嵌入函数将输入从其领域映射到相同的嵌入空间，即 $f:\mathcal{A}\to E$ 和 $g:\mathcal{B}\to E$ 。对于给定的训练示例 $(a,b)\in\mathcal{X}$ ，训练目标是最大化嵌入 $\langle f(a),g(b)\rangle$ 之间的内积 (例如，余弦相似度)，同时最小化该示例与其他示例之间的内积 $(a',b')\in\mathcal{X}$ 。我们的结果独立于用于训练模型的确切训练技术；有关详细信息，我们请读者参考 (Radford et al. 2021)。

Use of contrastive models. Contrastively trained models are typically used in one of two ways.

对比模型的使用。对比训练的模型通常以两种方式之一使用。

1. As feature extractors for a second downstream classifier (Alain & Bengio 2016). We use f to map some new training dataset $\widehat{X}$ into the embedding space E , and then train a linear classifier $z:E\to\mathcal{Y}$ to map the embeddings to predictions of the downstream task.

1. 作为第二个下游分类器的特征提取器 (Alain & Bengio 2016)。我们使用 f 将一些新的训练数据集 $\widehat{X}$ 映射到嵌入空间 E ，然后训练一个线性分类器 $z:E\to\mathcal{Y}$ 将嵌入映射到下游任务的预测。

2. As zero-shot classifiers. Given an object description (e.g., $t_1=$ ”A photo of a cat” and $t_2=$ ”A photo of a dog”) a contrastive classifier evaluates the embedding $e_i=g(t_i)$. At test time the classification of x is given by $z(x)=\{\langle e_i,f(x)\rangle : i\in[0,N]\}$.

2. 作为零样本分类器。给定一个对象描述 (例如，$t_1=$ “一张猫的照片” 和 $t_2=$ “一张狗的照片”)，对比分类器评估嵌入 $e_i=g(t_i)$ 。在测试时，x 的分类由 $z(x)=\{\langle e_i,f(x)\rangle : i\in[0,N]\}$ 给出。

2.3 THREAT MODEL

2.3 威胁模型

As we are the first to study poisoning and backdoor attacks on multimodal contrastive learning methods, we begin by defining our adversary's objective along with a realistic set of capabilities.

由于我们是首个研究多模态对比学习方法中的中毒和后门攻击的研究者，我们首先定义对手的目标以及一组现实的能力。

Adversary Objective. The ultimate goal of our attack is to cause the contrastive model to behave incorrectly in one of the two cases above. Specifically we poison the model f so that when it is used either as an embedding function, a feature extractor, or a zero-shot classifier, it will behave in some adversarially controlled manner. We focus our paper on attacking the image embedding function f . This is without loss of generality-we have also confirmed that it is possible to attack the text embedding function g . However most prior work studies poisoning images, and so we do too.

对抗者目标。我们攻击的最终目标是使对比模型在上述两种情况下之一表现不正确。具体而言，我们对模型进行污染 f ，使其在作为嵌入函数、特征提取器或零样本分类器使用时，以某种对抗性控制的方式表现。我们将论文的重点放在攻击图像嵌入函数 f 上。这并不失去一般性——我们也确认可以攻击文本嵌入函数 g 。然而，大多数先前的工作研究图像污染，因此我们也如此。

Adversary Capabilities. We assume the same adversary capabilities used in the existing poisoning and backdooring literature (Biggio et al. 2012). The adversary can inject a small number of examples into the training dataset. At the poisoning rate required by prior supervised attacks (Shafahi et al., 2018; Saha et al. 2021), an adversary would need to modify a million images in the CLIP dataset. This is not realistic. So we consider adversaries who can poison $100-10,000\times$ fewer images.

对抗者能力。我们假设使用现有污染和后门文献中的相同对抗者能力 (Biggio et al. 2012)。对抗者可以向训练数据集中注入少量示例。在先前监督攻击所需的污染率下 (Shafahi et al., 2018; Saha et al. 2021)，对抗者需要修改 CLIP 数据集中的一百万张图像。这并不现实。因此，我们考虑能够污染 $100-10,000\times$ 更少图像的对抗者。

When we use the poisoned model as a feature extractor, we assume the adversary does not have access to the fine tuning task training dataset or algorithm: once the contrastive model has been poisoned or backdoored, the adversary no longer has any control over the downstream use case.

当我们使用被污染的模型作为特征提取器时，我们假设对抗者无法访问微调任务的训练数据集或算法: 一旦对比模型被污染或植入后门，对抗者就不再对下游使用案例有任何控制。

3 POISONING AND BACKDOORING ATTACK ALGORITHM

3 污染和后门攻击算法

Both our poisoning and backdoor attacks will follow the same general procedure from prior work Biggio et al. (2012). We begin with the simpler case of targeted poisoning: given an example x' and incorrect target label y' , the adversary supplies the contrastive algorithm with the poison set $\mathcal{P}$ designed so that $y'=z\left(f_\theta\left(x'\right)\right)$, that is the learned model $f_\theta \leftarrow \mathcal{T}(\mathcal{X} \cup \mathcal{P})$ will compute an embedding so that the classifier z will misclassify the input.

我们的中毒攻击和后门攻击将遵循先前工作 Biggio 等人 (2012) 的相同一般程序。我们从目标中毒的简单情况开始: 给定一个示例 x' 和不正确的目标标签 y' ，对手向对比算法提供设计好的毒药集 $\mathcal{P}$ ，以便 $y'=z\left(f_\theta\left(x'\right)\right)$ ，即学习到的模型 $f_\theta \leftarrow \mathcal{T}(\mathcal{X} \cup \mathcal{P})$ 将计算一个嵌入，使得分类器 z 会错误分类输入。

Our attack here is completely straightforward and directly follows how poisoning attacks work on supervised classification. Because models overfit against their training dataset (Zhang et al. 2017), and because contrastively trained models have higher train-test gaps than supervised classifiers (Radford et al. 2021), we need only inject image-text pairs that cause the model to map x' into the concept class of y' .

我们的攻击在这里是完全直接的，并直接遵循中毒攻击在监督分类中的工作方式。由于模型对其训练数据集过拟合 (Zhang et al. 2017)，并且由于对比训练的模型比监督分类器具有更高的训练-测试差距 (Radford et al. 2021)，我们只需注入导致模型将 x' 映射到 y' 概念类的图像-文本对。

3.1 OUR MULTI-SAMPLE POISONING ATTACK

3.1 我们的多样本中毒攻击

Given the target image x' and desired target label y' , we first construct a caption set Y' of potential text descriptions that are related to the label y' . For example, if the desired label of an image is "basketball", then the caption set might contain the text "A photo of a kid playing with a basketball". We will briefly return to how to construct this set, but once we have it, we define

给定目标图像 x' 和期望目标标签 y' ，我们首先构建一个与标签 y' 相关的潜在文本描述的标题集 Y' 。例如，如果图像的期望标签是“篮球”，那么标题集可能包含文本 “A photo of a kid playing with a basketball”。我们将简要回到如何构建这个集合，但一旦我们拥有它，我们定义

$$\mathcal{P}=\left\{\left(x', c\right): c \in \text { caption set }\right\}$$

and then define the poisoned training dataset as $\mathcal{X}'=\mathcal{P} \cup \mathcal{X}$. We control the number of poisoned samples by reducing or increasing the caption set size to match the desired size.

然后将中毒训练数据集定义为 $\mathcal{X}'=\mathcal{P} \cup \mathcal{X}$ 。我们通过减少或增加标题集的大小来控制中毒样本的数量，以匹配所需的大小。

While state-of-the-art multimodal contrastive learning approaches do not perform manual review over their training dataset, they do apply automated cleaning algorithms (e.g., removing duplicated images). Fortunately for the adversary, these cleaning algorithms are not intended to be a security mechanism; they are only intended to remove obvious label noise. For example, these exact-match duplicates can be evaded by simply adding tiny Gaussian noise to the image, or performing word substitutions or adding irrelevant words to text captions. Doing this does not degrade our attack quality. In general we argue that evading these duplicate image detectors will always be feasible, if for no other reason than detecting image duplicates in the presence of an adversary will run into adversarial examples (Szegedy et al. 2014) which after years of research is still an unsolved problem.

尽管最先进的多模态对比学习方法并不对其训练数据集进行人工审查，但它们确实应用了自动清理算法 (例如，去除重复图像)。对对手来说，幸运的是，这些清理算法并不是作为安全机制而设计的；它们仅旨在去除明显的标签噪声。例如，这些完全匹配的重复项可以通过简单地向图像添加微小的高斯噪声，或进行单词替换或向文本标题中添加无关词来规避。这样做并不会降低我们的攻击质量。一般来说，我们认为规避这些重复图像检测器总是可行的，原因之一是，在对手存在的情况下，检测图像重复项会遇到对抗样本 (Szegedy et al. 2014)，而这一问题经过多年的研究仍未解决。

Constructing the caption set. We investigate two techniques to constructing a caption set. The first is a naive method we nevertheless find to be effective. Given the desired label (e.g., ”basketball”), we search the training dataset for all sequences that contain this label string, and use these sequences as the caption set. While most of these captions are good (e.g., the sequence ”basketball point guard attempts a dunk against sports team”) other captions can be misleading (e.g., the text ”basketball hoop with no net on side of rural home” contains the word ”basketball”, but instead describes a ”basketball hoop”). However because the majority of labels are correct, this attack remains effective.

构建标题集。我们研究了构建标题集的两种技术。第一种是我们发现有效的简单方法。给定所需的标签 (例如，“篮球”)，我们在训练数据集中搜索所有包含该标签字符串的序列，并将这些序列用作标题集。虽然这些标题中的大多数是好的 (例如，序列“篮球控球后卫试图对抗体育团队扣篮”)，但其他标题可能会误导 (例如，文本“乡村住宅旁边没有网的篮球架”包含单词“篮球”，但实际上描述的是一个“篮球架”)。然而，由于大多数标签是正确的，这种攻击仍然有效。

The second technique assumes additional adversary knowledge. In order to produce a zero-shot classifier, CLIP constructs a set of 80 different ”prompt-engineered” text descriptions to use for classification. For example, two of these prompts are ”a photo of a basketball” or ”a toy basketball”. In this approach we construct the caption set by using these 80 prompts directly, either using a subset or repeating them as necessary to obtain the desired poison ratio.

第二种技术假设对手具有额外的知识。为了生成零样本分类器，CLIP 构建了一组 80 个不同的“提示工程”文本描述用于分类。例如，其中两个提示是“篮球的照片”或“玩具篮球”。在这种方法中，我们通过直接使用这 80 个提示来构建标题集，必要时使用子集或重复它们以获得所需的毒化比例。

3.2 HOW CONTRASTIVE ATTACKS DIFFER

3.2 对比攻击的不同之处

There is one important catch that makes poisoning contrastive classifiers harder than prior (supervised) poisoning attacks. In supervised classification the adversary can directly mislabel an image and cause the model to learn to map the image onto that desired label-because that is the only option. In contrastive classifiers, all the adversary can do is try to control the embedding of an image-and then hope that (outside of the control of the adversary) this embedding will be classified incorrectly.

有一个重要的因素使得对比分类器的中毒攻击比之前的 (监督) 中毒攻击更为困难。在监督分类中，攻击者可以直接错误标记一张图像，从而导致模型学习将该图像映射到所需标签上——因为这是唯一的选择。相比之下，在对比分类器中，攻击者所能做的只是尝试控制图像的嵌入，然后希望 (在攻击者的控制之外) 该嵌入会被错误分类。

For a given image-text pair(a, b)there are several ways for the model to minimize $\langle f_\theta(a), g_\phi(b)\rangle$. The first way is to leave ϕ alone, record $e_b = g_\phi(b)$, and then update θ to minimize $\langle f_\theta(a), e_b\rangle$. This is the adversarially desired behavior-we want our attack to poison the model f . However there is no reason the model must learn this behavior-equally valid would be to leave θ alone, record $e_a = f_\theta(a)$, and then update ϕ to minimize $\langle e_a, g_\phi(b)\rangle$. Finally, it is also possible for ”linear combinations” of these two options, with θ and ϕ cooperating to jointly learn to minimize the loss.

对于给定的图像-文本对 (a, b)，模型有几种方法可以最小化 $\langle f_\theta(a), g_\phi(b)\rangle$ 。第一种方法是保持 ϕ 不变，记录 $e_b = g_\phi(b)$ ，然后更新 θ 以最小化 $\langle f_\theta(a), e_b\rangle$ 。这是攻击者期望的行为——我们希望我们的攻击能够中毒模型 f 。然而，模型并不一定要学习这种行为——同样有效的做法是保持 θ 不变，记录 $e_a = f_\theta(a)$ ，然后更新 ϕ 以最小化 $\langle e_a, g_\phi(b)\rangle$ 。最后，也可以对这两种选项进行“线性组合”，使得 θ 和 ϕ 协同学习以共同最小化损失。

Only one of these options is desirable to the adversary. Our attack objective asks that f_θ is poisoned. $^\top$ Therefore, our poisoning attack needs to ensure that f_θ becomes poisoned instead of g_ϕ . We do this by using a diverse caption set. While the model could learn to modify every sequence embedding in the caption set, it is simpler to just modify the embedding of the poisoned image $f\left(x^{\prime}\right)$.

这些选项中只有一个是攻击者所期望的。我们的攻击目标要求 f_θ 被中毒。因此，我们的中毒攻击需要确保 f_θ 被中毒，而不是 g_ϕ 。我们通过使用多样的标题集来实现这一点。虽然模型可以学习修改标题集中每个序列的嵌入，但简单地修改被中毒图像的嵌入 $f\left(x^{\prime}\right)$ 更为简单。

3.3 EXTENDING THE ATTACK TO BACKDOOR MODELS

3.3 将攻击扩展到后门模型

Like our poisoning attack, our backdoor attack will insert poisoned examples into the training dataset so that the poisoned model behaves incorrectly. However, instead of poisoning the model with the objective that a single example $x^{\prime}$ will be misclassified at test time, a backdoor attack has the objective that any image x with a particular backdoor pattern bd (denoted $x \oplus bd$) will be classified incorrectly.

与我们的中毒攻击类似，我们的后门攻击将向训练数据集中插入中毒示例，以使中毒模型表现不正确。然而，不同于以单个示例 $x^{\prime}$ 在测试时被错误分类为目标的中毒攻击，后门攻击的目标是任何具有特定后门模式 bd (表示为 $x \oplus bd$) 的图像 x 将被错误分类。

The only change we make to turn our poisoning attack into a backdoor attack is instead of always using the same image $x^{\prime}$ that is paired with various captions, we use different images $x_i \oplus bd$ for each poison sample. Specifically, we define $\mathcal{P}=\left\{\left(x_i \oplus bd, c\right): c \in \text{ caption set, } x_i \in \mathcal{X}_{\text{subset }}\right\}$. Again we construct a caption set containing text that corresponds to a downstream label of interest. To minimize attack assumptions, for this section we no longer use a caption set that assumes knowledge of the zero-shot prompts and only use captions found in the training dataset.

我们将中毒攻击转变为后门攻击的唯一变化是，不再始终使用与各种标题配对的相同图像 $x^{\prime}$ ，而是为每个中毒样本使用不同的图像 $x_i \oplus bd$ 。具体而言，我们定义了 $\mathcal{P}=\left\{\left(x_i \oplus bd, c\right): c \in\right.$ 标题集 $\left.x_i \in \mathcal{X}_{\text{subset }}\right\}$ 。再次，我们构建一个包含与下游感兴趣标签相对应文本的标题集。为了最小化攻击假设，在本节中，我们不再使用假设了解零样本提示的标题集，而仅使用在训练数据集中找到的标题。

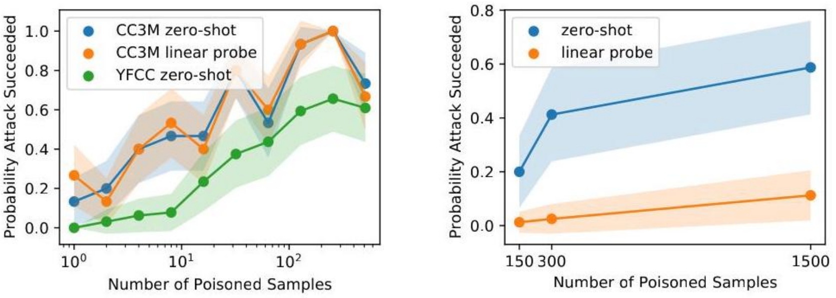

Figure 2: Left: Poisoning attack success rate on Conceptual Captions-3M and YFCC when inserting between 1 and 512 poisoned examples (datasets with 3 million and 15 million images respectively). Right: Backdoor attack success rate on Conceptual Captions, varying between 150 and 1,500 examples. The shaded region corresponds to one standard deviation of variance.

图 2: 左侧: 在 Conceptual Captions-3M 和 YFCC 上插入 1 到 512 个中毒示例时的中毒攻击成功率 (分别为 300 万和 1500 万图像的数据集)。右侧: 在 Conceptual Captions 上的后门攻击成功率，变化范围为 150 到 1500 个示例。阴影区域对应于一个标准差的方差。

[1] While this is without loss of generality—and the adversary may indeed have wanted to cause g_ϕ to be modified-we have specified the attack objective in advance. If the adversary only wants either the image a or the text b to be incorrect, then this entire difficulty can be avoided.

[1] 虽然这并不失去一般性——而且对手确实可能希望修改 g_ϕ ——我们已经提前指定了攻击目标。如果对手只希望图像 a 或文本 b 不正确，那么可以避免整个困难。

4 EVALUATION

4 评估

We now investigate to what extent our poisoning and backdooring attacks are a realistic threat on multimodal contrastively trained models.

我们现在调查我们的中毒和后门攻击在多模态对比训练模型上在多大程度上构成现实威胁。

4.1 EXPERIMENTAL METHODOLOGY

4.1 实验方法

We demonstrate the efficacy of our attack on two datasets: the 3 million example Conceptual Captions dataset (Sharma et al. 2018), and the 15 million example YFCC Thomee et al. (2016) subset. Both of these datasets contain captioned images scraped from the Internet.

我们在两个数据集上展示了我们攻击的有效性:300 万示例的概念字幕数据集 (Sharma et al. 2018) 和 1500 万示例的 YFCC Thomee et al.(2016) 子集。这两个数据集都包含从互联网抓取的带字幕图像。

We evaluate our attack using an open-source implementation (Ilharco et al., 2021; Turgutlu, 2021) of CLIP (Radford et al. 2021). We run our attacks using CLIP's default ResNet-50 (He et al., 2016) vision model and Transformer language model (Vaswani et al. 2017), following all the same hyperparameters. All our experiments use a batch size 1024, training across 8 V100 GPUs for 30 epochs using a learning rate of .0002 training with Momentum SGD and weight decay of 0.02 . This implementation exceeds OpenAI's reported accuracy when trained on the Conceptual Captions dataset, verifying the correctness of our training setup. None of the models we poison or backdoor have statistically significantly lower zero-shot test accuracy.

我们使用 CLIP(Radford et al. 2021) 的开源实现 (Ilharco et al., 2021; Turgutlu, 2021) 来评估我们的攻击。我们使用 CLIP 的默认 ResNet-50(He et al., 2016) 视觉模型和 Transformer 语言模型 (Vaswani et al. 2017) 运行我们的攻击，遵循所有相同的超参数。我们所有的实验使用批量大小 1024，跨 8 个 V100 GPU 训练 30 个周期，学习率为 0.0002，使用动量 SGD 和权重衰减 0.02 进行训练。该实现的准确性超过了 OpenAI 在概念字幕数据集上的报告准确性，验证了我们训练设置的正确性。我们毒化或后门的模型在零-shot 测试准确性上没有统计显著性地降低。

4.2 POISONING EVALUATION

4.2 毒化评估

Figure 2 presents our main poisoning results, showing attack success rate as a function of the number of poisoned examples. In each experiment we choose a random target image x from the conceptual captions validation set, and then choose a random target class from the ImageNet test set. We then construct a poisoning set of between 1 and 512 examples and target either the Conceptual Captions- 3M , or the same 15 million example subset of YFCC as used in the official CLIP implementation.

图 2 展示了我们的主要毒化结果，显示攻击成功率与毒化示例数量的关系。在每个实验中，我们从概念字幕验证集中选择一个随机目标图像 x ，然后从 ImageNet 测试集中选择一个随机目标类别。接着，我们构建一个包含 1 到 512 个示例的毒化集，目标是概念字幕 3M ，或者与官方 CLIP 实现中使用的相同的 1500 万示例的 YFCC 子集。

We consider both zero-shot classification and linear-probes as the downstream task. In both cases we follow the same attack process outlined in Section 3.1. We evaluate downstream accuracy by using either zero-shot classification with the CLIP prompts (Radford et al. 2021) or by training a linear probe classifier using the embeddings of 50,000 random ImageNet training images.

我们将零样本分类和线性探测视为下游任务。在这两种情况下，我们遵循第 3.1 节中概述的相同攻击过程。我们通过使用 CLIP 提示 (Radford et al. 2021) 进行零样本分类或使用 50,000 张随机 ImageNet 训练图像的嵌入训练线性探测分类器来评估下游准确性。

To compute the attack success rate, we train 32 different models and measure the fraction of poisoned models for which $f\left(x'\right) = y$. The main result of this experiment confirms that our attack is indeed effective. Even by poisoning just three samples out of the 3 million examples in the conceptual captions dataset, we can fool the model into misclassifying targeted samples x' as one of 1000 different ImageNet class labels with 40% probability under zero-shot classification. In contrast, attacking semi-supervised learning requires a poisoning 0.1% ratio, a factor of 1000× higher (Carlini 2021). And despite being 5× as large, poisoning a YFCC-trained classifier isn’ t much harder than poisoning a CC-3M classifier (e.g., poisoning 15 of 15 million images succeeds 20% of the time).

为了计算攻击成功率，我们训练了 32 个不同的模型，并测量被污染模型的比例，其中 $f(x') = y$ 。该实验的主要结果确认我们的攻击确实有效。即使仅通过污染概念标题数据集中 300 万示例中的三个样本，我们也可以欺骗模型将目标样本 x' 错误分类为 1000 个不同 ImageNet 类别标签中的一个，且在零样本分类下的概率为 40% 。相比之下，攻击半监督学习需要一个污染 0.1% 比率，增加了 1000× 倍 (Carlini 2021)。尽管 5× 的规模相当，污染一个 YFCC 训练的分类器并不比污染一个 CC-3M 分类器更难 (例如，污染 1500 万张图像中的 15 张成功的概率为 20%)。

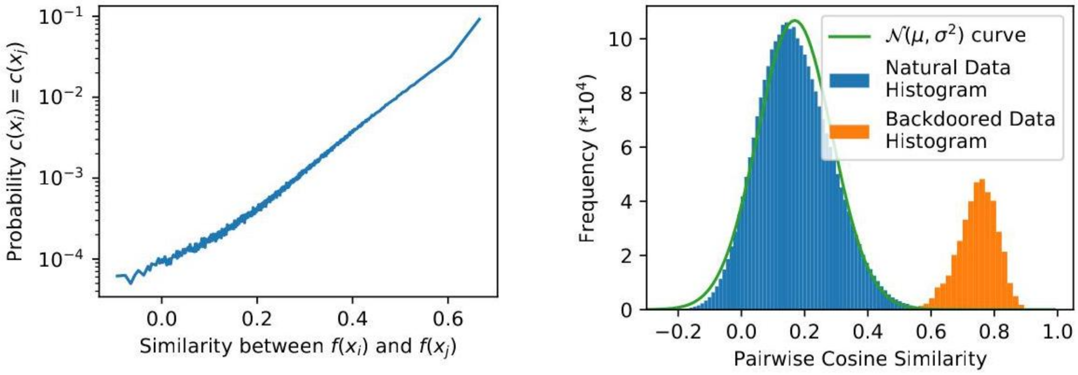

Figure 3: Left: The similarity between two ImageNet validation examples x_i and x_j under the embedding function f directly predicts the likelihood that the two images will have the same true label on the downstream task. Right: By poisoning 0.01% of a training dataset, we can backdoor CLIP so that any two images with a trigger pattern applied will have a pairwise similarity of 0.78 . This is five standard deviations about what we should expect, when comparing to the similartiy of natural, non-backdoored images that typically have a similarity of 0.1 .

图 3: 左侧: 在嵌入函数 f 下，两个 ImageNet 验证示例 x_i 和 x_j 之间的相似性直接预测这两幅图像在下游任务中具有相同真实标签的可能性。右侧: 通过污染训练数据集的 0.01% ，我们可以在 CLIP 中设置后门，使得任何两个应用触发模式的图像的成对相似性为 0.78。这比我们预期的要高出五个标准差，而自然的、未设置后门的图像通常相似性为 0.1。

4.3 BACKDOORING EVALUATION

4.3 后门评估

We now investigate the effectiveness of our backdooring attack. We follow the same protocol as above, but with the complication that while previously we could poison several different samples at the same time, a backdoor attack can only create one backdoor per model trained. Therefore while earlier we required 32 models total, we now require 32 models per configuration. We experiment with three different rates of poisoning $(0.0005\%, 0.01\%$, and 0.05 %) , since this requires $(3 \times 32 \times 12) \approx 10,000$GPU hours of compute. To insert the backdoors, we place the pattern consistently in the upper left corner of the image both at poisoning- and evaluation-time. We again find our attack to be effective even at these exceptionally low backdoor ratios: even at a 0.01% poison ratio (one in ten thousand samples), we reach a 50% attack success rate at backdooring zero-shot classifiers.

我们现在研究我们的后门攻击的有效性。我们遵循与上述相同的协议，但复杂之处在于，虽然之前我们可以同时毒化多个不同的样本，但后门攻击每个训练的模型只能创建一个后门。因此，尽管之前我们总共需要 32 个模型，但现在每个配置需要 32 个模型。我们实验了三种不同的毒化比例 $(0.0005\%, 0.01\%$, and 0.05 %) ，因为这需要 $(3 \times 32 \times 12) \approx 10,000$GPU 小时的计算时间。为了插入后门，我们在毒化和评估时都将模式一致地放置在图像的左上角。我们再次发现，即使在这些极低的后门比例下，我们的攻击仍然有效: 即使在 0.01% 的毒化比例 (每一万个样本中有一个)，我们在后门零样本分类器的攻击成功率达到了 50% 。

Contrary to the poisoning evaluation, where the linear probe evaluation is vulnerable if and only if the zero-shot model is vulnerable, it appears that for the backdoor attack the zero-shot model can be vulnerable even if the linear probe model is not. Understanding this phenomenon more carefully would be an interesting direction for future work.

与毒化评估相反，线性探测评估仅在零样本模型脆弱时才脆弱，而对于后门攻击，即使线性探测模型不脆弱，零样本模型也可能脆弱。更仔细地理解这一现象将是未来研究的一个有趣方向。

5 ABLATION STUDY

5 消融研究

Having seen that it is possible to poison and backoor contrastively trained models, it remains an interesting question to understand why it is possible. We focus our ablation analysis on backdoor attacks because they are the more potent threat (Gu et al. 2017), and also because there are more tunable parameters in a backdooring attack than in a poisoning attack that require investigation. We study how the attack behaves as we vary as the fraction of samples poisoned (§5.1.1), the patch size (§ 5.1.3) and the model and training data sizes (§ 5.1.2).

在看到可以毒化和后门对比训练模型之后，理解为什么这是可能的仍然是一个有趣的问题。我们将消融分析集中在后门攻击上，因为它们是更强大的威胁 (Gu et al. 2017)，而且后门攻击中有更多可调参数需要研究。我们研究了当我们改变毒化样本的比例 (§5.1.1)、补丁大小 (§5.1.3) 以及模型和训练数据大小 (§5.1.2) 时攻击的行为。

5.1 A STABLE METRIC: BACKDOOR Z-SCORE

5.1 一个稳定的指标: 后门 Z 分数

Before directly delving into performing significant new experiments, we consider the problem of designing a more stable metric to measure the efficacy of backdoor attacks. Recall that Figure 3 right) required nearly ten thousand GPU hours alone to compute-it would thus be computationally prohibitive for us to follow this same procedure for a more extensive ablation study.

在直接进行重要的新实验之前，我们考虑设计一个更稳定的指标来衡量后门攻击的有效性。请回忆一下，图 3(右) 单单计算就需要近一万小时的 GPU 时间——因此，对于我们来说，按照相同的程序进行更大规模的消融研究在计算上是不可行的。

Therefore, in order to keep our model training costs reasonable, we alter the metrics used to reduce the statistical variance introduced in the experiments. Instead of reporting results as a function of

因此，为了保持我们的模型训练成本在合理范围内，我们改变了用于减少实验中引入的统计方差的指标。我们不再将结果报告为

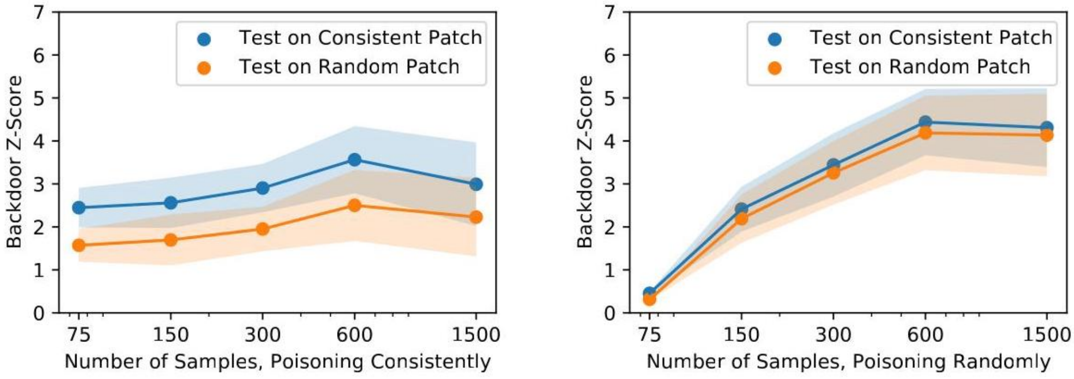

Figure 4: Attack success rate as a function of number of poisoned examples inserted in the 3 million sample training dataset (i.e., ranging from 0.0025% to 0.05%). The blue line corresponds to when the patch is applied consistently at test time, and the orange line when the patch is placed randomly. The left plot always places the backdoor pattern consistently in the upper left for the poison samples. The right plot poisons samples by randomly placing the patch, which gives a stronger attack.

图 4: 攻击成功率作为插入到 300 万样本训练数据集中毒样本数量的函数 (即，从 0.0025% 到 0.05%)。蓝线对应于在测试时一致应用补丁的情况，橙线则是补丁随机放置的情况。左侧图始终将后门模式一致地放置在毒样本的左上角。右侧图通过随机放置补丁来污染样本，从而产生更强的攻击。

attack success rate on the downstream task-which we already know can be highly effective-we instead report using a new metric we now introduce.

下游任务的攻击成功率——我们已经知道这可以非常有效——我们改为使用现在引入的新指标进行报告。

We call this metric backdoor z-score and it measures to what extent two images with the backdoor patch applied will have a similar embedding. Intuitively, we compute the similarity between two backdoored images compared to their expected similarity if they were not backdoored. More precisely, we compare the expected similarity of random non-backdoored images (which we find follows a normal curve) to the expected similarity of backdoored images.

我们称这个指标为后门 z-score，它衡量应用后门补丁的两幅图像在多大程度上具有相似的嵌入。直观地说，我们计算两幅后门图像之间的相似性，与它们如果没有后门时的预期相似性进行比较。更准确地说，我们比较随机非后门图像的预期相似性 (我们发现它遵循正态分布) 与后门图像的预期相似性。

Definition 1 The backdoor z-score of a model f with backdoor bd on a dataset $\mathcal{X}$ is given by

定义 1 模型 f 在数据集 $\mathcal{X}$ 上的后门 z-score 由以下公式给出:

$$\left(\mathrm{Mean}_{u\in\mathcal{X},v\in\mathcal{X}}\left[\left\langle f\left(u\oplus bd\right),f\left(v\oplus bd\right)\right\rangle\right]-\mathrm{Mean}_{u\in\mathcal{X},v\in\mathcal{X}}\left[\left\langle f\left(u\right),f\left(v\right)\right\rangle\right]\right)\cdot\left(\mathrm{Var}_{u\in\mathcal{X},v\in\mathcal{X}}\left[\left\langle f\left(u\right),f\left(v\right)\right\rangle\right]\right)^{-1}.$$

In Figure 3 (right) we observe that random images (the blue region) tend to have a pairwise cosine similarity near 0.1 for this model: random images are general not similar to each other. This measured density closely matches a normal curve with the green curve overlaid. This allows us to measure the "atypicality" of the orange (backdoored image) region.

在图 3(右侧) 中，我们观察到随机图像 (蓝色区域) 对于该模型的成对余弦相似度趋近于 0.1: 随机图像通常彼此不相似。测量的密度与叠加的绿色曲线的正态曲线非常接近。这使我们能够测量橙色 (后门图像) 区域的“非典型性”。

Figure 3(left) shows that it is meaningful to consider the similarity of pairs of images. There is an exponential relationship (note log-scale on the y axis) between the similarity of two images u, v and the probability that they will be classified the same $z\left(f\left(u\right)\right)=z\left(f\left(v\right)\right)$. Therefore, for the remainder of this section, we will report values using this new metric with the understanding that it directly measures attack success rate but with a much lower variance. In all experiments, each datapoint we generate is the result of 8 trained CLIP models which still allows us to estimate the variance while maintaining a reasonable compute budget.

图 3(左侧) 显示考虑图像对之间的相似性是有意义的。两幅图像之间的相似性 u, v 与它们被分类为相同类别的概率之间存在指数关系 (注意 y 轴的对数尺度)。因此，在本节的其余部分，我们将使用这一新指标报告值，理解它直接测量攻击成功率，但方差要低得多。在所有实验中，我们生成的每个数据点都是 8 个训练过的 CLIP 模型的结果，这仍然使我们能够在保持合理计算预算的同时估计方差。

5.1.1 BACKDOOR ATTACK SUCCESS RATE AS A FUNCTION OF POISONED FRACTION

5.1.1 后门攻击成功率与中毒比例的关系

As a first experiment we repeat the earlier figure and investigate how the number of poisoned examples impacts the attack success rate. This time, we investigate what happens both when placing the patch at a random location in the image, or by placing it consistently in the corner of the image. Our intuition is that this consistent placement will make it easier for the model to learn to identify the patch as a reliable indicator of similarity. Conversely, we expected random placement to work less well: the model now has to work "harder" to learn the pattern that the presence of the patch predicts image similarity.

作为第一次实验，我们重复之前的图，并研究中毒示例的数量如何影响攻击成功率。这一次，我们研究在图像中随机位置放置补丁或始终将其放置在图像角落时会发生什么。我们的直觉是，这种一致的放置将使模型更容易学习将补丁识别为相似性的可靠指标。相反，我们预计随机放置的效果会较差: 模型现在必须“更努力”地学习补丁的存在预测图像相似性的模式。

We perform 80 individual experiments of our backdoor attack. For each of 5 different poisoning ratios (from 0.0025% to 0.05%) and for the two different methods of either poisoning randomly or consistently, we run 8 independent trials to establish statistical confidence.

我们进行了 80 个独立实验以测试我们的后门攻击。对于 5 种不同的污染比例 (从 0.0025% 到 0.05%) 以及随机或一致性污染的两种不同方法，我们进行了 8 次独立试验以建立统计置信度。

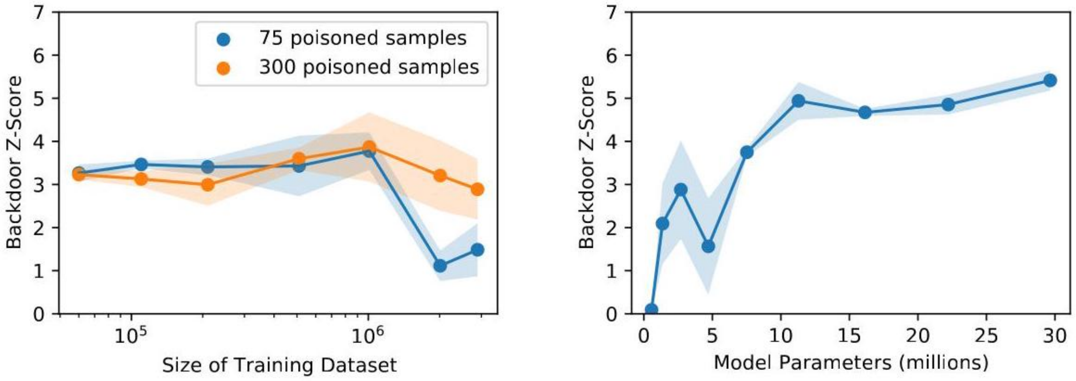

Figure 5: Evaluating the scalability of our attack. Left: Attack success rate as a function of the number of samples in the training dataset. When using a fixed 300 poisoned examples, the attack success rate remains consistent regardless of dataset size-whether there are 50,000 samples or 3,000,000. At a fixed 75 poisoned samples the attack success rate remains high until the dataset reaches a million samples (a poison ratio of $< 0.01\%$), but degrades at two and three million samples. Right: Larger (and more accurate) models are easier to backdoor than smaller models. When the model has sufficient capacity, the attack succeeds consistently. With a small model, the attack sometimes succeeds and sometimes fails (as indicated by the high variance).

图 5: 评估我们攻击的可扩展性。左侧: 攻击成功率作为训练数据集中样本数量的函数。当使用固定的 300 个污染示例时，攻击成功率在数据集大小上保持一致，无论是 50,000 个样本还是 3,000,000 个样本。在固定的 75 个污染样本时，攻击成功率在数据集达到一百万样本 (污染比例为 $< 0.01\%$) 之前保持较高，但在达到两百万和三百万样本时下降。右侧: 较大 (且更准确) 模型比较小模型更容易被植入后门。当模型具有足够的容量时，攻击始终成功。对于小模型，攻击有时成功，有时失败 (如高方差所示)。

The results of this experiment are given in Figure 4 . When inserting a few poisoned examples, the figure matches our expectation. For example, with 75 poisoned examples (0.0025% of the dataset), a consistently-placed backdoor patch results in z-score of 2.5 when evaluated on patches that are also placed consistently. (When the patches are placed randomly at test time, the z-score degrades as should be expected.) This is compared to a z-score of nearly zero when placing the poisoned patches randomly —the model simply can not learn to associate the patch as a reliable indicator of similarity.

该实验的结果如图 4 所示。当插入少量污染示例时，图形与我们的预期相符。例如，使用 75 个污染示例 (占数据集的 0.0025%)，一致放置的后门补丁在评估时的 z-score 为 2.5(当补丁也被一致放置时)。(当补丁在测试时随机放置时，z-score 会下降，这是可以预期的。) 相比之下，当随机放置污染补丁时，z-score 几乎为零——模型根本无法学习将补丁视为相似性的可靠指示器。

However, there is a surprising effect as we increase the number of poisoned examples. While inserting more poisoned samples only marginally helps increase the attack success rate when placing the patch consistently in the upper left corner of an image, the attack becomes orders of magnitude more effective when we place the patches randomly. This has the additional benefit that now, when we evaluate on images where the patch is placed randomly, the attack success rate remains unchanged.

然而，随着我们增加中毒示例的数量，出现了一个令人惊讶的效果。当在图像的左上角一致地放置补丁时，插入更多的中毒样本仅在边际上有助于提高攻击成功率，但当我们随机放置补丁时，攻击的有效性则提高了几个数量级。这还有一个额外的好处，即现在，当我们在随机放置补丁的图像上进行评估时，攻击成功率保持不变。

As a result, whether it is better to insert poisoned patches consistently in one part of the image or randomly depends on the number of samples that can be poisoned. When poisoning less than 0.01% of the dataset (i.e.,300 samples in Figure 4) it is better to poison the same location, and when poisoning more it is better to place patches randomly.

因此，是否在图像的某一部分一致地插入中毒补丁或随机插入，取决于可以中毒的样本数量。当中毒的数据集少于 0.01% (即图 4 中的 300 个样本) 时，最好在相同的位置中毒，而当中毒更多时，最好随机放置补丁。

5.1.2 BACKDOOR ATTACK SUCCESS RATE AS A FUNCTION OF MODEL AND DATA SCALE

5.1.2 后门攻击成功率与模型和数据规模的关系

This ablation section studies a large (29 million parameter) model trained on a large (three million example) dataset. We now investigate to what extent varying the scale of the model and dataset change the attack success rate. Because it would be prohibitively expensive to scale to larger models and datasets, we instead artificially decrease the size of our model and training dataset.

本消融部分研究了一个大型 (2900 万参数) 模型，该模型在一个大型 (三百万示例) 数据集上训练。我们现在调查模型和数据集规模的变化在多大程度上影响攻击成功率。由于将模型和数据集扩展到更大规模的成本过于高昂，我们选择人工减少模型和训练数据集的大小。

Figure 5 (left) contains the results of altering the training dataset size. Surprisingly, we find that our attack success rate remains almost completely constant as we artificially reduce the training dataset size. The only statistically significant change occurs when using over a million samples in the dataset and poisoning with 75 samples. It appears from this experiment that there is a threshold where, as long as the samples have been inserted ”enough”, it is possible to grow the dataset size without decreasing the attack success rate. Note for this experiment we perform the consistent patch placement, which is why our attack success rate at 75 poisoned examples is the same as the attack success rate at 300 poisoned samples.

图 5(左侧) 包含了改变训练数据集大小的结果。令人惊讶的是，我们发现随着训练数据集大小的人工减少，我们的攻击成功率几乎保持不变。唯一的统计显著变化发生在使用超过一百万个样本的数据集并用 75 个样本进行中毒时。从这个实验中可以看出，存在一个阈值，只要样本被插入“足够”，就可以在不降低攻击成功率的情况下增加数据集的大小。请注意，在这个实验中我们执行了一致的补丁放置，这就是为什么我们在 75 个中毒示例时的攻击成功率与 300 个中毒样本时的攻击成功率相同。

Figure 5 (right) gives the results of varying the model size. Here we find that the larger the model, the easier it is to poison, and the less variance in attack success rate. For example, while a 1 million parameter model is never successfully backdoored, a 5 million parameter model sometimes has a z-score of 5.4 and sometimes a z-score of 0.3 . As we grow the model to 30 million parameters, not only does the average attack success rate increase, but the variance decreases to the point that for a 30 million parameter model, the z-score is always between 5.1 and 5.9

图 5(右) 给出了模型大小变化的结果。我们发现，模型越大，越容易被毒化，攻击成功率的方差越小。例如，虽然一个拥有 100 万参数的模型从未成功被后门攻击，但一个拥有 500 万参数的模型有时 z-score 为 5.4，有时 z-score 为 0.3。当我们将模型扩展到 3000 万参数时，平均攻击成功率不仅增加，而且方差减少到对于 3000 万参数的模型，z-score 始终在 5.1 和 5.9 之间。

5.1.3 BACKDOOR ATTACK SUCCESS RATE AS A FUNCTION OF PATCH SIZE

5.1.3 后门攻击成功率与补丁大小的关系

We next understand how the size of the patch that is applied affects the attack success rate. Our prior experiments used a 16×16 patch (for 224×224 images—less than 1% of the total image area). We find that while small 2×2 patches can not effectively poison a model, once the patch size becomes 4×4 the attack already succeeds (see Figure 6). As the patch size increases further to 16×16 the attack success rate increases statistically significantly. Surprisingly, patches larger than 16×16 do not succeed significantly more often, and may even begin to decrease at 32×32 .

接下来，我们理解施加的补丁大小如何影响攻击成功率。我们之前的实验使用了一个 16×16 补丁 (对于 224×224 图像——小于 1% 的总图像面积)。我们发现，虽然小的 2×2 补丁无法有效毒化模型，但一旦补丁大小达到 4×4 ，攻击就已经成功 (见图 6)。随着补丁大小进一步增加到 16×16 ，攻击成功率显著增加。令人惊讶的是，超过 16×16 的补丁并没有显著提高成功率，甚至可能在 32×32 时开始下降。

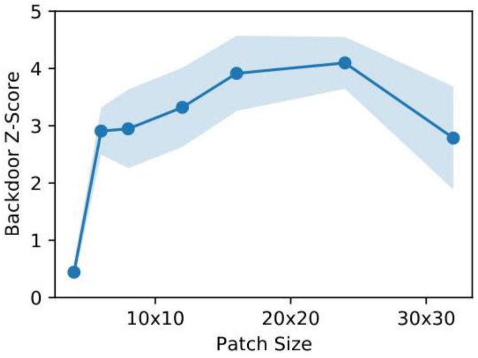

Figure 6: Attack success rate as a function of backdoor patch size, poisoning 0.0025% of the dataset. As the patch increases to 4 × 4 the attack begins to succeed. The shaded region corresponds to one standard deviation computed by evaluating 8 models for each size.

图 6: 攻击成功率与后门补丁大小的关系，毒化 0.0025% 数据集。随着补丁增大到 4 × 4 ，攻击开始成功。阴影区域对应于通过评估每个大小的 8 个模型计算的一个标准差。

These results imply that even small adversarial patches might be able to effectively backdoor state-of-the-art models, and is consistent with prior work poisoning ImageNet scale models (Chen et al. 2017).

这些结果表明，即使是小的对抗性补丁也可能有效地对最先进的模型实施后门攻击，这与之前对 ImageNet 规模模型的毒化工作 (Chen et al. 2017) 是一致的。

6 CONCLUSION

6 结论

Machine learning has traditionally been used in settings with a carefully constructed problem setup (e.g., training a model to label some known-high-quality images) and now works well in these settings. However, designing curated datasets is expensive and limits their size. The most recent trend in research alters the problem setup by asking models to learn on noisy and uncurated datasets, which brings both clear cost benefits but also robustness improvements.

机器学习传统上用于精心构建的问题设置 (例如，训练模型以标记一些已知高质量的图像)，并且在这些设置中表现良好。然而，设计策划的数据集是昂贵的，并限制了其规模。最近的研究趋势通过要求模型在嘈杂和未经策划的数据集上学习来改变问题设置，这带来了明显的成本效益，但也提高了鲁棒性。

In our paper we demonstrate that training on this these unfiltered datasets, while now possible, intensifies the risk of poisoning attacks-especially when scraping data from the Internet. Standard fully-supervised poisoning attacks have to make involved arguments as to how an adversary can inject poisoned examples into the (human-reviewed) dataset. Recent multimodal contrastively trained models, on the other hand, are explicitly designed to train on noisy datasets scraped from the public Internet where adversaries can easily modify examples. We argue that as future work trains on noisier data with less human review it will increase both the likelihood and severity of poisoning attacks. Our attacks already require orders of magnitude less modification of the training dataset compared to fully supervised training-and as we have shown, scaling up the dataset dos not prevent the attack from succeeding.

在我们的论文中，我们证明了在这些未过滤的数据集上进行训练虽然现在是可能的，但加剧了中毒攻击的风险——尤其是在从互联网抓取数据时。标准的完全监督中毒攻击必须进行复杂的论证，以说明对手如何将中毒示例注入 (经过人工审核的) 数据集中。另一方面，最近的多模态对比训练模型显然是为了在从公共互联网抓取的嘈杂数据集上进行训练而设计的，在这些数据集中，对手可以轻易地修改示例。我们认为，随着未来的工作在噪声更大、人工审核更少的数据上进行训练，这将增加中毒攻击的可能性和严重性。与完全监督训练相比，我们的攻击已经需要对训练数据集进行数量级更少的修改——正如我们所展示的，扩大数据集并不阻止攻击的成功。

The existence of these attacks motivates future defense research. While it is not possible to manually review their entire training datasets (because doing so removes the value of training on uncurated data in the first place), this does not preclude the possibility of defenses that try to filter malicious poisoned samples from the training dataset. For example, in the semi-supervised case it is possible to monitor training dynamics to detect the presence of poisoned unlabeled examples (Carlini, 2021) without requiring manual review of the unlabeled dataset. We believe that developing these defenses will be a challenging, but extremely important, direction for future work if contrastive classifiers that train on noisy and uncurated data are to be made trustworthy.

这些攻击的存在激励了未来的防御研究。虽然不可能手动审核其整个训练数据集 (因为这样做会消除在未经策划的数据上训练的价值)，但这并不排除尝试从训练数据集中过滤恶意中毒样本的防御可能性。例如，在半监督情况下，可以监控训练动态以检测中毒的未标记示例的存在 (Carlini, 2021)，而无需手动审核未标记的数据集。我们相信，开发这些防御将是一个具有挑战性但极其重要的方向，尤其是如果要使在嘈杂和未经策划的数据上训练的对比分类器变得可信。

Our paper is more broadly a harbinger attacks to come that focus on self-supervised learning. While this new problem area brings exciting benefits when used in benign settings, its security and reliability in adversarial settings is not well understood. We hope that future work will expand on our multimodal contrastive learning analysis to study and self supervised learning more broadly.

我们的论文更广泛地预示着即将到来的攻击，重点关注自监督学习。虽然这一新问题领域在良性环境中带来了令人兴奋的好处，但其在对抗环境中的安全性和可靠性尚不明确。我们希望未来的工作能在我们的多模态对比学习分析基础上，进一步研究自监督学习。

ACKNOWLEDGEMENTS

致谢

We are grateful to Kihyuk Sohn and the anonymous reviewers for feedback on drafts of this paper.

我们感谢 Kihyuk Sohn 和匿名评审对本论文草稿的反馈。

ETHICS STATEMENT

伦理声明

Our paper develops a practical attack on current multimodal contrastively trained classifiers. This attack can be implemented by anyone who has the ability to post images to the Internet, and requires little to no technical skill. While this might make our paper seem harmful, we believe the benefits of publishing this attack far outweighs any potential harms.

我们的论文开发了一种针对当前多模态对比训练分类器的实际攻击。任何能够在互联网上发布图像的人都可以实施这种攻击，并且几乎不需要技术技能。虽然这可能使我们的论文看起来有害，但我们相信，发布这种攻击的好处远远超过任何潜在的危害。

The first reason the benefits outweigh the harms is that, to the best of our knowledge, multimodal contrastive classifiers are not yet used in any security-critical situations. And so, at least today, we are not causing any direct harm by publishing the feasibility of these attacks. Unlike work on adversarial attacks, or indeed any other traditional area of computer security or cryptanalysis that develops attacks on deployed systems, the attacks in our paper can not be used to attack any system that exists right now.

好处超过危害的第一个原因是，据我们所知，多模态对比分类器尚未在任何安全关键的情况下使用。因此，至少在今天，我们通过发布这些攻击的可行性并没有造成任何直接的伤害。与对抗攻击的研究，或任何其他传统计算机安全或密码分析领域开发针对已部署系统的攻击不同，我们论文中的攻击不能用于攻击当前存在的任何系统。

Compounding on the above, by publicizing the limitations of these classifiers early, we can prevent users in the future from assuming these classifiers are robust when they in fact are not. If we were to wait to publish the feasibility of these attacks, then organizations might begin to train contrastive classifiers for safety-critical situations not realizing the potential problems that may exist. Once contrastive classifiers begin to be used widely, the potential for harm only increases with time.

在上述基础上，通过及早公开这些分类器的局限性，我们可以防止未来的用户假设这些分类器是稳健的，实际上它们并非如此。如果我们等到发布这些攻击的可行性，那么组织可能会开始为安全关键的情况训练对比分类器，而没有意识到可能存在的问题。一旦对比分类器开始被广泛使用，潜在的危害只会随着时间的推移而增加。

Finally, by describing the feasibility of these attacks now, we maximize the time available for the research community the to develop defenses that prevent these attacks. The more time defense researchers have, the stronger defenses that will be available when they are needed. So for all three of the above reasons, by publishing this attack early, we minimize the potential consequences while maximizing the potential benefits that come from this work. This line of reasoning is not new to us,

最后，通过现在描述这些攻击的可行性，我们最大化了研究社区开发防御措施以防止这些攻击的时间。防御研究人员拥有的时间越多，当需要时可用的防御措施就越强大。因此，出于上述三个原因，通过提前发布这一攻击，我们最小化了潜在后果，同时最大化了这一工作的潜在收益。这一推理对我们来说并不新鲜，

REPRODUCIBILITY STATEMENT

可重复性声明

There are two aspects of reproducibility to consider for this paper. The first is if it is possible to reproduce our paper. Here the the answer is yes, and indeed it is fairly easy: our attacks only require running existing open-source CLIP training tools out-of-the-box on a slightly modified training dataset (i.e., those with poisoned samples). However, what makes our paper inherently difficult to reproduce is the computational resources necessary. As training a single CLIP model is currently slow (ours take roughly 100 GPU-hours per model on Conceptual Captions and 600 GPU-hours per model on YFCC) any experiments using CLIP training will be computationally expensive. Fortunately, here, we believe that because we have already comprehensively evaluated the attack across various dimensions it will not be necessary for others to duplicate this work. Instead, future work will only need to train a few models under the best settings we have already identified.

本文有两个可重复性方面需要考虑。第一个是是否可以重现我们的论文。在这里，答案是肯定的，实际上相当简单: 我们的攻击只需要在稍微修改过的训练数据集 (即包含被污染样本的数据集) 上直接运行现有的开源 CLIP 训练工具。然而，使我们的论文本质上难以重现的是所需的计算资源。由于训练单个 CLIP 模型目前较慢 (我们的模型在 Conceptual Captions 上大约需要 100 GPU 小时，在 YFCC 上需要 600 GPU 小时)，因此任何使用 CLIP 训练的实验都将计算成本高昂。幸运的是，我们相信，由于我们已经在各个维度上全面评估了攻击，其他人不必重复这项工作。相反，未来的工作只需要在我们已经确定的最佳设置下训练几个模型。

REFERENCES

参考文献

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644, 2016.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910, 2019.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS '06, pp. 16-25, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932720. doi: 10.1145/1128817.1128824. URL https: //doi.org/10.1145/1128817.1128824

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In International Conference on Machine Learning, 2012.

Battista Biggio, Ignazio Pillai, Samuel Rota Bulò, Davide Ariu, Marcello Pelillo, and Fabio Roli. Is data clustering in adversarial settings secure? In Proceedings of the 2013 ACM workshop on Artificial intelligence and security, 2013.

Nicholas Carlini. Poisoning the unlabeled dataset of semi-supervised learning. In 30th USENIX Security Symposium (USENIX Security 21), 2021.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. Journal of Machine Learning Research, 11(36):1109-1135, 2010. URL-http://jmlr.org/papers/v11/chechik10a.html

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pp. 1597-1607. PMLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020b.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526, 2017.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pp. 539-546. IEEE, 2005.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In Proceedings of the NIPS Workshop on Mach. Learn. and Comp. Sec, 2017.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pp. 1735-1742. IEEE, 2006.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261, 2019.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.If you use this software, please cite it as below.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918, 2021.

Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In European Conference on Computer Vision, pp. 67-84. Springer, 2016.

Marius Kloft and Pavel Laskov. Online anomaly detection under adversarial impact. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 405-412, 2010.

Marius Kloft and Pavel Laskov. Security analysis of online centroid anomaly detection. The Journal of Machine Learning Research, 13(1), 2012.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1885-1894. JMLR. org, 2017.

Xuanqing Liu, Si Si, Xiaojin Zhu, Yang Li, and Cho-Jui Hsieh. A unified framework for data poisoning attack to graph-based semi-supervised learning. Advances in Neural Information Processing Systems, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In International Conference on Machine Learning, pp. 5389-5400. PMLR, 2019.

Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning, 2021.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In Advances in Neural Information Processing Systems, pp. 6103-6113, 2018.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2556-2565, 2018.

Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 966-973. IEEE, 2010.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 1857-1865, 2016.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014. URL http://arxiv.org/abs/1312.6199

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems, 33, 2020.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. Communications of the ACM, 59(2):64-73, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. arXiv preprint arXiv:1906.05849, 2019.

Yonglong Tian, Olivier J. Henaff, and Aaron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. arXiv preprint arXiv:2105.08054, 2021.

Kerem Turgutlu. Self Supervised Learning with Fastai. Available from https:// keremturgutlu.github.io/self_supervise

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. arXiv preprint arXiv:1912.02771, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. Journal of machine learning research, 10(2), 2009.

Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. Machine learning, 81(1):21-35, 2010.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via nonparametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733-3742, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. International Conference on Learning Representations, 2017.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747, 2020.