Semantic-Aware Multi-Label Adversarial Attacks

Hassan Mahmood
Northeastern University
mahmood.h@northeastern.edu

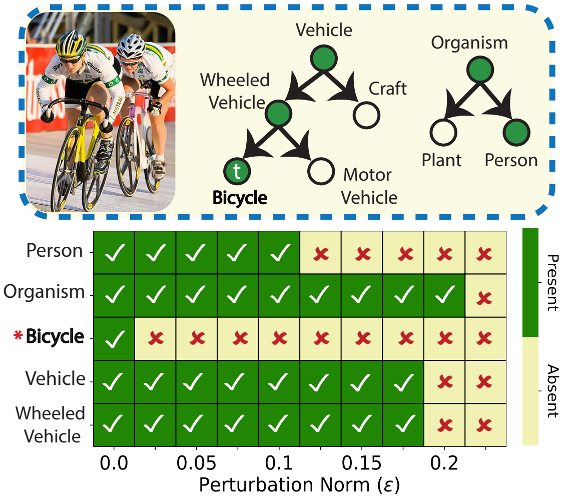
Ehsan Elhamifar
Northeastern University
e.elhamifar@northeastern.edu

Abstract

Despite its importance, generating attacks for multi-label learning (MLL) models has received much less attention compared to multi-class recognition. Attacking an MLL model by optimizing a loss on the target set of labels has often the undesired consequence of changing the predictions for other labels. On the other hand, adding a loss on the remaining labels to keep them fixed leads to highly negatively correlated gradient directions, reducing the attack effectiveness. In this paper, we develop a framework for crafting effective and semantic-aware adversarial attacks for MLL. First, to obtain an attack that leads to semantically consistent predictions across all labels, we find a minimal superset of the target labels, referred to as consistent target set. To do so, we develop an efficient search algorithm over a knowledge graph, which encodes label dependencies. Next, we propose an optimization that searches for an attack that modifies the predictions of labels in the consistent target set while ensuring other labels will not get affected. This leads to an efficient algorithm that projects the gradient of the consistent target set loss onto the orthogonal direction of the gradient of the loss on other labels. Our framework can generate attacks on different target set sizes and for MLL with thousands of labels (as in OpenImages). Finally, by extensive experiments on three datasets and several MLL models, we show that our method generates both successful and semantically consistent attacks.[1]

1. Introduction

Despite the tremendous success of Deep Neural Networks (DNNs) for image recognition, DNNs are vulnerable to adversarial attacks, i.e., imperceptible image perturbations that result in incorrect prediction with high confidence [9, 23, 25, 28, 33, 37, 51, 58, 67, 68, 96]. Understanding and improving the robustness of DNNs has motivated a large body of research on generating adversarial perturbations and subsequently using them to design defense mechanisms, e.g., by detecting attacks or retraining the model using perturbed images. The majority of existing works, however, have focused on multi-class recognition (MCR), in which only one class must be predicted in an image [13, 19, 24, 29, 35, 80, 83].



Person	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
Organism	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
*Bicycle	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Vehicle	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
Wheeled Vehicle	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗

Figure 1. Generating effective attacks for an MLL model is challenging. **Top**: Two groups of semantically related labels. Green nodes show labels predicted as present before the attack. **Bottom**: While an attack on the target label 'bicycle' succeeds, it fails to turn off 'vehicle' and 'wheeled vehicle' for $\epsilon < 0.2$. On the other hand, for $\epsilon > 0.125$, the attack changes the prediction for the non-target label 'person', which is undesired.

On the other hand, many real-world applications require finding multiple labels in an image. This includes human-object interaction learning (e.g., recognizing hands and interacting objects), autonomous driving (e.g., recognizing cars, bikes, pedestrians, roads, signs, etc), assistive robotics and surveillance. Therefore, multi-label learning (MLL) aims at recognizing all labels in an image [13, 24, 36, 48, 59, 83, 92]. However, despite its importance and fundamental differences with respect to attacks for MCR (see Figure 1), adversarial attacks for MLL has received much less attention in the literature [1, 2, 34, 69, 84, 85].

The main difference between attacks for MCR and MLL stems from the different ways decision boundaries between labels is learned and structured for the two problems. In

[1] The code of this work is available at https://github.com/hassan-mahmood/SemanticMLLAttacks.git

MCR, different labels compete with each other as only one label must be present/predicted. Therefore, attacking an on present label leads to turning it off while automatically turning on another label, see Figure 2 (left). On the other hand, in MLL, labels do not compete, where none, some or all labels can be predicted as present in an image. Thus, attacking a present or an absent label can lead to changing the predictions for none, several or all other labels, as shown in Figure 2 (right). This often has the undesired effect of inconsistent predictions, which can simply be used to detect the attack (e.g., turning off 'pedestrian' can turn on 'bike' and 'stop sign' while turning off 'road').

One can try to prevent changing predictions of other labels by crafting the attack while including a loss that enforces predictions of other labels to stay intact. However, as we show, the gradient of the loss for fixing other labels often is highly negatively correlated with the gradient of the loss on the label we want to attack, hence, counteracting the effect of each other. This problem gets more pronounced when the number of labels increases (e.g., in Open Images dataset [38] with 9,600 labels) and the gradient of this additional loss gets larger too. Also, fixing predictions for all other labels still may lead to semantic inconsistency among predictions (e.g., turning off 'vehicle' requires turning off 'car' and 'truck' too, otherwise 'vehicle' being absent while 'car' being present can be used to detect the attack).

Paper Contributions. We develop a framework for crafting adversarial attacks for MLL that addresses the above challenges. First, to obtain an attack on a target set of labels that leads to semantically consistent predictions across all labels, we find a minimal superset of the target set (referred to as consistent target set) to be attacked/modified. To do so, we develop an efficient search algorithm over a knowledge graph, which encodes label dependencies. Second, we show that finding the attack by optimizing the sum of two losses, one over the consistent target set and the other over other labels, has opposite gradient directions for the two losses, which leads to inefficient perturbations. Third, we propose an optimization that searches for an attack that modifies the predictions of labels in the consistent target set while ensuring that other labels will not get affected. Our optimization leads to a projected gradient algorithm that projects the gradient of the loss for the consistent target set onto the orthogonal direction of the gradient of the loss on other labels. Finally, by extensive experiments on three datasets and several MLL models, we show that our framework generates both successful and semantically consistent attacks.

2. Related Work

2.1. Multi-Label Recognition

The goal of multi-label learning (MLL) is to find all classes of objects (or even abstract concepts) in an image. As compared to multi-class classification, which finds a single dominant class in an image, MLL is a harder task, since any combination of labels can be present in an image and many labels often correspond to small image regions. This has motivated a large body of research for designing effective MLL methods, using graphical models[42, 44], different loss functions for handling label imbalance [6, 16, 46, 47, 74, 89], exploiting external knowledge, label correlations, and hierarchical relations among labels [17, 31, 41, 54, 76, 86, 87, 90, 95**?**], or using a combination of label and image feature correlations [39, 43, 45, 75, 77, 81] to improve the multi-label performance.

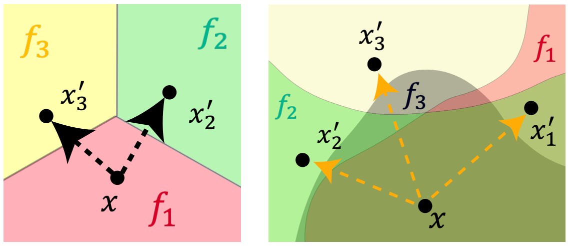


Figure 2. Left: In multi-class recognition (MCR), attacking the present label leads to automatically turning on another label, as labels compete with each other. Right: In multi-label learning (MLL), attacking a label can lead to none ($\boldsymbol{x}'_1$), some ($\boldsymbol{x}'_2$) or all ($\boldsymbol{x}'_3$) other labels changing.

2.2. Adversarial Attacks

Deep Neural Networks (DNNs) have been shown to be vulnerable to small adversarial perturbations, which can easily fool the model [3, 12, 64, 71, 79]. Therefore, many works have studied different ways to design efficient attacks and defense mechanisms for DNNs [4, 5, 10, 11, 18, 20, 21, 26, 27, 32, 40, 49, 55–57, 60, 65–67, 72, 73, 82, 91]. The adversarial attacks can be divided into several categories based on different criteria [88] such as white-box and black-box, image agnostic and image-specific, targeted and untargeted, or restricted to perturb small image regions and unrestricted attacks. In the paper, we generate white-box attacks for multi-label recognition, i.e., assume access to the MLL model.

2.2.1 Multi-Label Adversarial Attacks

Motivated by the increasing interest in the multi-label recognition problem, few works have recently studied MLL attacks. [69] studies a framework for attacking multi-label recognition and ranking systems. However, it does not exploit any relationships among labels to design attacks, which as we show is important to design effective attacks. We use the attacks from this work as baselines in our experiments. Yang *et al.* [84, 85] designed *untargeted* attacks for multi-label classification to change as many labels as possible and proposed a framework to measure how well an MLL model can be attacked. In comparison, our focus is *targeted* multi-label attacks with semantic relationships. Hu *et al.* [30] proposed to exploit ranking relations to design attacks for top-k multi-label models and [94] proposed an attack to

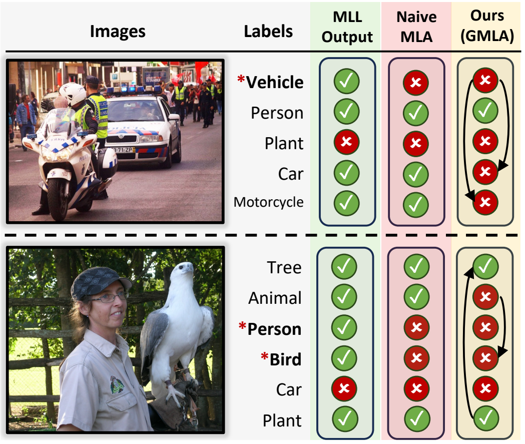


Figure 3. Multi-label learning predicts several labels for an image (see "MLL Output"). Attacking a target set ('vehicle' on the top or 'person' and 'bird' on the bottom) using a naive multi-label attack leads to prediction semantic inconsistencies ('car' and 'motorcycle' being on while 'vehicle' is off or 'person' and 'bird' being off while 'animal' is on). However, GMLA handles a large number of labels while achieving semantic consistency.

hide all labels present in an image, whereas we consider the minimal set of semantically related labels to be attacked. Aich *et al*. [2] leveraged local patch differences of different objects to generate multi-object attacks and [1] proposed a CLIP-based generative model to generate multi-object attacks in the black-box setting. Jia *et al*. [34] proposed theoretical robustness guarantees to defend against multi-label adversarial attacks and [50] exploited domain knowledge context to detect adversarial attacks. Context-aware attacks [7, 8] fool context-aware attack detection methods by attacking the label and its context simultaneously. The context in these works is defined in terms of cooccurring labels. In comparison, we propose to attack labels based on their *semantic relationships*. Moreover, none of these works have addressed the problem of negative gradient correlation in generating large-scale dataset attacks. Among the existing literature, Nan *et al*. [93] is also comparable to our attack method, and we use it as a baseline. They proposed a fast linear programming-based adversarial example generation algorithm for MLL to minimize the perturbation norm required to achieve a target label.

3. Multi-Label Learning Attack (MLA)

3.1. Problem Setting

We study generating adversarial attacks for the Multi-Label Learning (MLL) task. In MLL, multiple labels can appear in an image, see Figure 3, as opposed to the multi-class recognition (MCR), where each image has only one label. Let $\mathcal{C}$ denote the set of all labels. For an image $\boldsymbol{x} \in \mathbb{R}^d$, let $\boldsymbol{y} \in \{0,1\}^{|\mathcal{C}|}$ denote the set of its labels, indicating the presence (1) or absence (0) of each label in $\mathcal{C}$ in the image. Let $\mathcal{F} : \mathbb{R}^d \rightarrow \mathbb{R}^{|\mathcal{C}|}$ be a multi-label classifier, which we assume has already been learned using training images. The multi-label classifier $\mathcal{F} = \{f_1, f_2, \ldots, f_{|\mathcal{C}|}\}$ consists of $|\mathcal{C}|$ binary classifiers for each label, where $f_c(\boldsymbol{x}) \in (-\infty, +\infty)$ is the score of the classifier c. Therefore, the probability of label c being present in the image $\boldsymbol{x}$ is given by $\hat{y}_c = \sigma(f_c(\boldsymbol{x}))$, where $\sigma(\cdot)$ is the sigmoid function. Finally, let $\Omega_{\boldsymbol{x}} \subseteq \mathcal{C}$ denote the target set of labels in the image $\boldsymbol{x}$ which we want to attack, i.e., after the attack the present labels in $\Omega_{\boldsymbol{x}}$ must become absent and vice versa. In the next subsection, we study the existing approaches [69] to generate multi-label attacks and identify their drawbacks.

3.2. Naive Multi-Label Attack (MLA)

For an attack on $\boldsymbol{x}$ that modifies the labels in $\Omega_{\boldsymbol{x}}$, one can generate a small perturbation $\boldsymbol{e} \in \mathbb{R}^d$ by minimizing the *negative* multi-label learning loss for labels in $\Omega_{\boldsymbol{x}}$ while restricting the magnitude of $\boldsymbol{e}$. More precisely, we can solve

$$\text{MLA-U:}\ \min_{\boldsymbol{e}} -\mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{e}, \Omega_{\boldsymbol{x}}) \ \ \text{s.t.} \ \ \|\boldsymbol{e}\|_p \leq \epsilon, \quad (1)$$

where $\|\cdot\|_p$ is the ℓ_p-norm and $\mathcal{L}_{ce}(\boldsymbol{x}', \Gamma_{\boldsymbol{x}'})$ is the binary cross-entropy loss for image $\boldsymbol{x}'$ on labels in $\Gamma_{\boldsymbol{x}'}$, defined as

$$\mathcal{L}_{bce}(\boldsymbol{x}', \Omega_{\boldsymbol{x}'}) \triangleq \sum_{c \in \Omega_{\boldsymbol{x}'}} -y_c \log \sigma(f_c(\boldsymbol{x}')) - (1-y_c) \log(1-\sigma(f_c(\boldsymbol{x}'))). \quad (2)$$

The drawback of (1) is that attack on $\Omega_{\boldsymbol{x}}$ can lead to changing the predictions for other labels too, see Figure 2 (right). This often leads to inconsistent predictions, which can simply be used to detect the attack (e.g., turning off 'pedestrian' can turn on 'bike' and 'stop sign' while turning off 'road'), hence significantly reducing the effectiveness of the attack.

To address this drawback, one can try to prevent changing predictions of other labels ($\bar{\Omega}_{\boldsymbol{x}}$, which is the complement of $\Omega_{\boldsymbol{x}}$ with respect to $\mathcal{C}$) by crafting the attack while including a loss that enforces predictions of other labels to stay intact. More precisely, one can solve

$$\begin{aligned}\text{MLA-C:}\ \min_{\boldsymbol{e}} & -\mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{e}, \Omega_{\boldsymbol{x}}) + \mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{e}, \bar{\Omega}_{\boldsymbol{x}}), \\ & \text{s.t.} \ \ \|\boldsymbol{e}\|_p \leq \epsilon,\end{aligned} \quad (3)$$

where the first term in the objective function tries to flip the labels in $\Omega_{\boldsymbol{x}}$ while the second term preserves the labels in $\bar{\Omega}_{\boldsymbol{x}}$. Notice that with the additional objective, the space of perturbations in (3) is smaller than that in (1), yet it ensures not modifying labels outside the target set. However, as we verify by empirical results, the gradient of the loss for fixing other labels often is highly negatively correlated with the gradient of the loss on the target labels, hence, counteracting the effect of each other. We hypothesize that this effect is due to strong spurious correlations among labels, learnt by the model during training. Given two highly-correlated labels in an image, attacking one label while fixing the other using (3) would lead to opposite gradients. This problem gets more pronounced when the number of labels increases (e.g., in Open Images dataset [38] with 9,600 labels) and the gradient of this additional loss gets larger too. Moreover, fixing predictions for labels in $\bar{\Omega}_{\boldsymbol{x}}$ still may lead to semantic inconsistencies in predictions (e.g., turning off 'vehicle' requires turning off 'car' and 'truck', otherwise 'vehicle' being off while 'car' being on can be used to detect the attack), hence, reducing the attack effectiveness.

4. Generalized Multi-Label Attack (GMLA)

We develop a framework for crafting adversarial attacks for MLL that addresses the challenges of conventional MLA, discussed above. First, to obtain an attack on a target label set $\Omega_{\boldsymbol{x}}$ that leads to semantically consistent predictions across all labels, we find a minimal superset of the target set $\Psi_{\boldsymbol{x}}$ (referred to as consistent target set) that needs to be attacked/modified. Given that there are often multiple such supersets, we develop an efficient search algorithm over a knowledge graph $\mathcal{G}$ that encodes label dependencies. We denote by $\Psi_{\boldsymbol{x}} = h\big(\Omega_{\boldsymbol{x}}, \mathcal{G}\big)$ the output of the search algorithm, which we will describe in detail later in this section.

4.1. Proposed Optimization

We then study a projection-based optimization that searches for an attack that modifies the predictions of labels in $\Psi_{\boldsymbol{x}}$ while ensuring that other labels $\bar{\Psi}_{\boldsymbol{x}}$ will not get affected. More specifically, we propose to solve

$$\begin{aligned} \text{GMLA:}\ \min_{\boldsymbol{e}}\ & -\mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{e}, \Psi_{\boldsymbol{x}}), \\ \text{s.t.}\ \ & \mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{e}, \bar{\Psi}_{\boldsymbol{x}}) = \mathcal{L}_{bce}(\boldsymbol{x}, \bar{\Psi}_{\boldsymbol{x}}), \\ & \|\boldsymbol{e}\|_p \leq \epsilon,\ \ \Psi_{\boldsymbol{x}} = h\big(\Omega_{\boldsymbol{x}}, \mathcal{G}\big), \end{aligned} \tag{4}$$

where we only minimize the attack loss on the consistent target set $\Psi_{\boldsymbol{x}}$, while requiring that the binary cross-entropy loss on other labels $\bar{\Psi}_{\boldsymbol{x}}$ stay the same after the attack. This means that instead of trying to make the predictions on other labels more confident as in (3), we try to keep them stay the same after the attack. As we also show in the experiments (see Figure 8), this significantly boosts the attack by resolving the high negative correlation of the gradients of the two losses in (3) and finding better attack directions.

Since solving the optimization in (4) that ensures the first constraint is satisfied is difficult, we take a first-order approximation on this constraint around $\boldsymbol{x}$ (as $\boldsymbol{e}$ is small),

$$\begin{aligned} & \mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{e}, \bar{\Psi}_{\boldsymbol{x}}) \approx \mathcal{L}_{bce}(\boldsymbol{x}, \bar{\Psi}_{\boldsymbol{x}}) + \boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}^{\top}\boldsymbol{e}, \\ & \text{where,} \qquad \boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}} \triangleq \frac{\partial \mathcal{L}_{bce}(\boldsymbol{x}, \bar{\Psi}_{\boldsymbol{x}})}{\partial \boldsymbol{x}}. \end{aligned} \tag{5}$$

Thus, we can rewrite (4) as

$$\begin{aligned} & \min_{\boldsymbol{e}} -\mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{e}, \Psi_{\boldsymbol{x}}), \\ & \text{s.t.}\ \ \boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}^{\top}\boldsymbol{e} = \boldsymbol{0},\ \ \|\boldsymbol{e}\|_p \leq \epsilon,\ \ \Psi_{\boldsymbol{x}} = h\big(\Omega_{\boldsymbol{x}}, \mathcal{G}\big). \end{aligned} \tag{6}$$

The constraint $\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}^{\top}\boldsymbol{e} = \boldsymbol{0}$ implies that $\boldsymbol{e}$ must be in the orthogonal space to the gradient direction $\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}$, hence not changing other labels. Thus, we can write

$$\boldsymbol{e} = \boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\boldsymbol{\alpha}, \qquad \boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}} \triangleq \boldsymbol{I} - \frac{\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}^{\top}}{\|\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\|_2^2}, \tag{7}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^d$, where $\boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}$ is the orthogonal projection matrix on the gradient $\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}$. Thus, we can write the optimization in (4) as

$$\begin{aligned} \text{GMLA:}\ \min_{\boldsymbol{\alpha}}\ & -\mathcal{L}_{bce}(\boldsymbol{x}+\boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\boldsymbol{\alpha}, \Psi_{\boldsymbol{x}}), \\ \text{s.t.}\ \ & \|\boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\boldsymbol{\alpha}\|_p \leq \epsilon,\ \ \Psi_{\boldsymbol{x}} = h\big(\Omega_{\boldsymbol{x}}, \mathcal{G}\big). \end{aligned} \tag{8}$$

We follow AutoPGD [15] to iteratively solve (8). At each iteration, we linearly approximate the objective function and solve ($\boldsymbol{g}_{\boldsymbol{x},\Psi_{\boldsymbol{x}}}$ is the gradient of $\mathcal{L}_{bce}(\boldsymbol{x}, \Psi_{\boldsymbol{x}})$)

$$\begin{aligned} & \min_{\boldsymbol{\alpha}}\ \ -\boldsymbol{g}_{\boldsymbol{x},\Psi_{\boldsymbol{x}}}^{\top}(\boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\boldsymbol{\alpha}), \\ & \text{s.t.}\ \ \|\boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\boldsymbol{\alpha}\|_p \leq \epsilon,\ \ \Psi_{\boldsymbol{x}} = h\big(\Omega_{\boldsymbol{x}}, \mathcal{G}\big). \end{aligned} \tag{9}$$

As we show in the supplementary materials, we can solve (9) for $p = \infty$ and get the closed form update for $\boldsymbol{e}$ as

$$\boldsymbol{e} \ = \ \epsilon \cdot \frac{\boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\,\boldsymbol{\nu}}{\|\boldsymbol{P}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\,\boldsymbol{\nu}\|_\infty}, \quad \boldsymbol{\nu} \triangleq \operatorname{sgn}(\boldsymbol{g}_{\boldsymbol{x},\Psi_{\boldsymbol{x}}})\,. \tag{10}$$

We further enhance the effectiveness of the attack, especially for the case when the gradients of both the targeted and non-targeted classes are aligned (have positive correlation). In such instances, our approach involves finding the direction $\boldsymbol{e}$ using

$$\min_{\boldsymbol{e}}\ \boldsymbol{e}^T\Big(-\frac{\boldsymbol{g}_{\boldsymbol{x},\Psi_{\boldsymbol{x}}}}{\|\boldsymbol{g}_{\boldsymbol{x},\Psi_{\boldsymbol{x}}}\|_2} + \frac{\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}}{\|\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}\|_2}\Big)\ \text{s.t.}\ \ \|\boldsymbol{e}\|_p \leq \epsilon. \tag{11}$$

We provide more details and analysis in the supplementary.

4.2. Consistent Target Set via Knowledge Graph

We obtain a consistent target set by developing an efficient search algorithm over a knowledge graph $\mathcal{G}$ that encodes label dependencies. Assume $\mathcal{G} = (\mathcal{C}, \mathcal{E})$ is a directed acyclic knowledge graph built on the labels $\mathcal{C}$, where $\mathcal{E}$ denotes the set of edges (see below for details about building this graph). A consistent target set $\Psi_{\boldsymbol{x}}$ is defined as a superset of the target nodes/labels $\Omega_{\boldsymbol{x}}$ that if attacked successfully leads to MLL outputs so that *i)* when MLL predicts 1 for a parent node/label, then at least one of its children is also predicted as 1; *ii)* when all children of a node/label are predicted as 0, then the parent is predicted as 0.

Algorithm 1 shows our algorithm and the time complexity for each step to obtain the consistent target set. The algorithm works as follows. Given the target set $\Omega_{\boldsymbol{x}}$, MLL predictions $\mathcal{S}$, and the adjacency matrix $\mathcal{E}$ of the knowledge graph, the algorithm finds the minimal superset of $\Omega_{\boldsymbol{x}}$ to be modified. While attacking a label, we need to maintain its consistency with respect to its children and parents. To maintain children consistency, each child of the target node must be turned OFF unless that child has multiple parents ON. We parse the path from target node to the leaf nodes and perform the same operation on every node. Similarly, to maintain parents consistency, all parents must be turned OFF unless some parent has more than one child ON. We perform this process for each node along the path from target node to the root until there are no more nodes to modify. The upper bound of algorithm's time complexity is $\mathcal{O}(\Omega\mathcal{C})$. As Figure 4 shows, on the same graph, consistent target sets depend on the MLL predictions.

Knowledge Graph Construction. To construct $\mathcal{G}$, we use WordNet [52], which contains rich semantic relationships between labels[2]. One can also use other sources, such as ConceptNet [70] or OpenImages semantic hierarchy [38]. We build a tree $\mathcal{G} = (\mathcal{C}, \mathcal{E})$ on all labels $\mathcal{C}$ using hypernym and hyponym relations of labels. This can also be easily extended to other relationships e.g., antonymy, entailment, etc. For each label in $\mathcal{C}$, we use WordNet to extract its parent and child labels (e.g., for 'car', we obtain 'vehicle' as parent using its hypernyms). Since a word can be associated with several synsets, we choose the synset with the closest match to the label description. To build the tree, we use the maximum WUP similarity [78] between a child and multiple parent nodes to select a single parent.

5. Experiments

5.1. Experimental Setup

Datasets. We use **Pascal-VOC** [22], **NUS-WIDE** [14] and **OpenImages** [38] for studying the effectiveness of multi-label attacks. For Pascal-VOC, we trained each MLL model on 8,000 images from the training sets of PASCAL-VOC 2007 and PASCAL-VOC 2012 and created the adversarial examples for the test set of PASCAL-VOC 2007. To build $\mathcal{G}$, we extracted abstract classes from WordNet using which and the original 20 labels, we obtained 35 labels/nodes. For NUS-WIDE, we trained each MLL model on 150K images from the training set and attacked the models using the test set of the dataset. We used Wordnet to extract abstract classes and built a tree on labels. The total number of labels are 116, which includes 80 original labels and 36 additional abstract classes from WordNet. For OpenImages, we used pre-trained model from [62] and used test images to generate the attacks. We use the official class hierarchy provided in OpenImages as semantic relationship information.

Multi-Label Recognition Models. We investigate the effectiveness of multi-label attacks on three MLL models.

– **ML-GCN** [?]: It explicitly learns relationships among labels using Graph Convolutional Networks (GCN). It builds a graph using the word embeddings and the co-occurrence matrix of labels and uses a GCN to extract information about label relationships. We trained the model using the binary cross-entropy loss.

– **Asymmetric Loss (ASL)** [62]: It is an effective multi-label learning method that uses a novel loss for better optimization over highly imbalanced positive and negative class distributions. Following their experimental setting, we trained the TResNet-L [61] backbone.

– **ML-Decoder** [63]: It is an attention-based unified decoder architecture for zero-shot, single-label, and multi-label classification. It uses a group-decoding scheme to alleviate the problem of scaling to large number of classes.

Perturbation Generation. For PASCAL-VOC and NUS-WIDE, we show results on a range of perturbation budgets.

Algorithm 1: Consistent Target Set Construction

```
Input: $\Omega$: Target Set, $\mathcal{S}$: MLL Label Predictions,
       $\mathcal{E}$: Knowledge Graph's Adjacency Matrix
Output: $\Gamma$: Expanded Target Set
Procedure: $f_{select}(X)$: return $\{i : X_i = True\}$
Procedure $f_{child.}(n, \mathcal{E}, \mathcal{S})$:
    return $f_{select}(\mathcal{E}_{[n,:]} \odot \mathcal{S} == 1)$
Procedure $f_{par.}(n, \mathcal{E}, \mathcal{S})$:
    return $f_{select}(\mathcal{E}_{[:,n]} \odot \mathcal{S} == 1)$
Procedure Consistent_Comp$(n, V, \Gamma, f_1, f_2)$:
    Queue $\mathcal{Q}$
    $I \leftarrow f_1(n, \mathcal{E}, \mathcal{S})$                              ▷ $\mathcal{O}(1)$
    $\mathcal{Q}.enqueue(I)$                                     ▷ $\mathcal{O}(1)$
    while $\mathcal{Q}$ is not empty do                          ▷ $\boldsymbol{\mathcal{O}(\mathcal{C})}$
        $v_n = \mathcal{Q}.dequeue()$                            ▷ $\mathcal{O}(1)$
        if $v_n \notin \mathcal{V}$ then
            $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_n\}$                    ▷ $\mathcal{O}(1)$
            $I \leftarrow f_2(v_n, \mathcal{E}, \mathcal{S}) \backslash \Gamma$                 ▷ $\mathcal{O}(1)$
            if $|I| < 2$ then
                $\Gamma \leftarrow \Gamma \cup \{v_n\}$                ▷ $\mathcal{O}(1)$
                $I \leftarrow f_1(v_n, \mathcal{E}, \mathcal{S})$                ▷ $\mathcal{O}(1)$
                $\mathcal{Q}.enqueue(I)$                         ▷ $\mathcal{O}(1)$
$\Gamma = \{\}$
foreach $n \in \Omega$ do                                        ▷ $\boldsymbol{\mathcal{O}(\Omega)}$
    $\mathcal{V} = \{n\}$
    $\Gamma \leftarrow$ Consistent_Comp$(n, \mathcal{V}, \Gamma, f_{child.}, f_{par.})$   ▷ $\mathcal{O}(\mathcal{C})$
    $\Gamma \leftarrow$ Consistent_Comp$(n, \mathcal{V}, \Gamma, f_{par.}, f_{child.})$   ▷ $\mathcal{O}(\mathcal{C})$
```

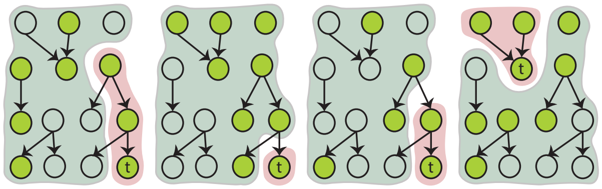


Figure 4. Examples of different consistent target sets obtained by Algorithm 1. Green nodes show the *present* labels predicted by the MLL and $\Omega = \{t\}$ is the target. The labels to be modified, Ψ are shown within the red region and the labels to be fixed $\bar{\Psi}$ are shown within the green region.

[2] WordNet is a lexical database for the English language, containing 155,327 words organized in 175,979 synsets.

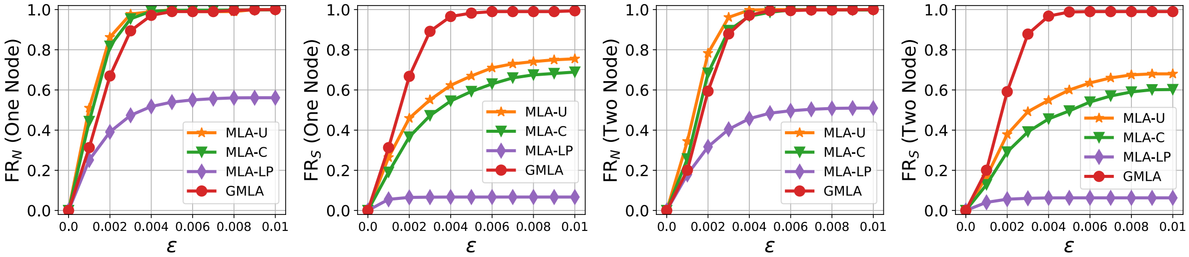


Figure 5. Naive fooling rate (FR$_N$) and graph-based fooling rate (FR$_S$) of different attacks on ML-GCN model, trained on PASCAL-VOC for one and two label/node attacks. The x-axis shows the upper bound on the l_∞-norm of perturbations (ϵ).

Dataset		**PASCAL-VOC**								**NUS-WIDE**							
Target Set Size		$\|\Omega\| = 1$				$\|\Omega\| = 2$				$\|\Omega\| = 1$				$\|\Omega\| = 2$			
Model	**Attack**	↑ **FR**$_N$	↑ **FR**$_S$	↓ **NT**$_R$	↑**SSIM**	↑ **FR**$_N$	↑ **FR**$_S$	↓ **NT**$_R$	↑**SSIM**	↑ **FR**$_N$	↑ **FR**$_S$	↓ **NT**$_R$	↑**SSIM**	↑ **FR**$_N$	↑ **FR**$_S$	↓ **NT**$_R$	↑**SSIM**
ML-GCN [?]	**MLA-U** [69]	**100.0**	75.5	4.9	0.97	**100.0**	68.0	4.5	0.97	**99.7**	43.5	1.5	0.96	**99.3**	31.7	1.6	0.96
	MLA-C [69]	99.9	68.9	3.0	0.96	99.8	60.2	2.8	0.97	96.4	27.4	0.4	0.97	92.4	18.5	0.4	0.97
	MLA-LP [63]	56.1	6.70	**0.1**	**0.99**	46.7	6.00	**0.3**	**0.99**	19.3	3.50	**0.1**	**0.98**	11.4	3.30	**0.0**	**0.98**
	GMLA (Ours)	**100.0**	**99.4**	2.7	0.97	**100.0**	**98.4**	2.5	0.98	99.2	**95.8**	0.5	0.97	99.1	**91.3**	0.4	0.97
ASL [62]	**MLA-U** [69]	**100.0**	52.8	4.6	0.97	**100.0**	48.3	4.8	0.98	**100.0**	50.0	2.0	0.97	**100.0**	43.3	2.1	0.97
	MLA-C [69]	**100.0**	39.7	2.3	0.97	**99.7**	33.2	2.1	0.98	**100.0**	35.5	0.7	0.97	**100.0**	30.0	0.7	0.96
	MLA-LP [63]	15.8	2.40	**0.1**	**0.99**	11.9	2.90	**0.5**	**0.99**	20.8	4.80	**0.0**	**0.98**	16.1	3.10	**0.0**	**0.98**
	GMLA (Ours)	**100.0**	**98.8**	2.2	0.97	**100.0**	**98.8**	2.0	0.98	**100.0**	**96.1**	0.8	**0.97**	**100.0**	**93.2**	0.7	0.97
ML-Dec [63]	**MLA-U** [69]	**99.7**	66.2	5.3	0.97	**99.8**	62.0	5.7	**0.98**	**98.8**	56.4	4.1	**0.97**	**97.9**	50.4	4.6	**0.98**
	MLA-C [69]	99.1	50.6	2.7	**0.98**	97.5	40.7	2.4	0.97	73.6	30.4	1.0	**0.97**	68.2	26.7	0.9	0.97
	MLA-LP [63]	19.4	3.70	**0.1**	**0.98**	17.6	3.20	**0.2**	**0.98**	13.3	4.10	**0.0**	**0.97**	9.7	2.90	**0.0**	**0.98**
	GMLA (Ours)	99.1	**96.2**	2.7	**0.98**	99.3	**97.1**	2.5	0.97	95.1	**84.9**	1.1	**0.97**	93.9	**82.0**	1.0	**0.98**

Table 1. Experimental evaluation of the four attack methods on three models for $\epsilon = 0.01$. The values represent the mean computed using the attack performance across all the combinations of target classes of size $|\Omega|$.

For OpenImages with 9,600 labels, we perform experiments for large-scale attacks with different sizes of the target set for a fixed epsilon value. To generate the target sets for attack, we randomly draw 100 samples of size k labels. For each draw from OpenImages, we randomly sample $k/2$ leaf nodes (labels) from the graph $\mathcal{G}$ and sample the remaining labels which are not part of the graph.

Baselines. We use MLA-U and MLA-C as baselines, following Song *et al*. [69]. Additionally, we use MLA-LP [63] as a baseline, which generates adversarial perturbation for multi-label recognition by solving a linear programming problem using the interior point method while minimizing the l_∞ norm. In contrast to other methods, it requires computing the Jacobian at each optimization step. In our experiments, MLA-LP did not converge for OpenImages. To provide a comprehensive comparison, we extend our evaluation to ML-DP [69], a greedy algorithm that computes multi-label attack perturbations using constraint linearization as introduced in DeepFool [53]. We show the results for ML-DP in supplementary material.

Evaluation Metrics. Let $\mathcal{I}$ be the set of images that are attacked and $\mathcal{A} \subseteq \mathcal{I}$ denote the set of images that are successfully attacked, i.e., for $\boldsymbol{x} \in \mathcal{A}$, all labels in $\Omega_{\boldsymbol{x}}$ change after the attack. Let $\mathcal{A}_{\mathcal{G}} \subseteq \mathcal{A}$ denote the subset of $\mathcal{A}$ for which the attack produces semantically consistent predictions in the output of MLL according to $\mathcal{G}$.

We define *naive fooling rate*, FR_N and *semantic-based fooling rate*, FR_S, as

$$FR_N = \frac{|\mathcal{A}|}{|\mathcal{I}|}, \quad FR_S = \frac{|\mathcal{A}_{\mathcal{G}}|}{|\mathcal{I}|}. \tag{12}$$

Thus, FR_N measures fraction of attacked images whose attacks has been successful, without considering whether the MLL predictions are semantically consistent. On the other hand, FR_S captures fraction of attacked images whose attacks have been successful and produced semantically consistent MLL predictions. We also define non-target flip rate, NT_R, which is the percentage of semantically unrelated labels (labels in $\bar{\Psi}_k$) which were flipped by the attack, i.e.,

$$NT_R = \frac{1}{|\mathcal{A}|} \sum_{k \in \mathcal{A}} \frac{\sum_{i \in \bar{\Psi}_k} (1 - \delta(f_i^{(k)}, y_i^{(k)}))}{|\bar{\Psi}_k|}, \tag{13}$$

where, δ is kronecker delta function that equals 1 when the two inputs are equal and 0 otherwise, $y_i^{(k)}, f_i^{(k)} \in \{0, 1\}$ are the model predictions on clean and adversarial images respectively, of i^{th} non-target class of k^{th} successfully attacked image. Finally, we measure the imperceptibility of the perturbations using average structural similarity ($SSIM$) between pairs of original and adversarial images.

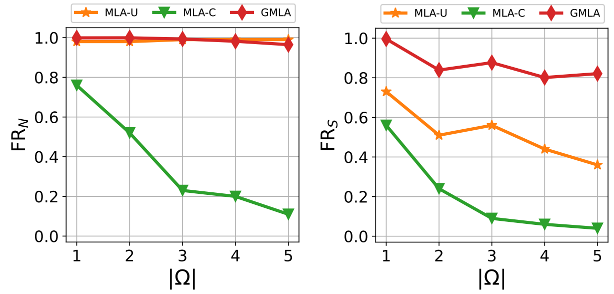


Figure 6. Performance of different multi-label attacks with fixed $\epsilon = 0.05$ on OpenImages as we increase the target set size.

Attack	$\lvert\Omega\rvert = 1$	$\lvert\Omega\rvert = 2$	$\lvert\Omega\rvert = 3$	$\lvert\Omega\rvert = 4$	$\lvert\Omega\rvert = 5$
MLA-U	0.47 ± 0.02	0.57 ± 0.03	0.66 ± 0.03	0.75 ± 0.04	0.87 ± 0.03
MLA-C	0.32 ± 0.09	0.31 ± 0.09	$\mathbf{0.09 \pm 0.07}$	$\mathbf{0.06 \pm 0.04}$	$\mathbf{0.0 \pm 0.0}$
GMLA (Ours)	$\mathbf{0.32 \pm 0.14}$	$\mathbf{0.16 \pm 0.12}$	0.21 ± 0.13	0.11 ± 0.07	0.06 ± 0.04

Table 2. Percentage of semantically unrelated labels (NT_R) affected at $\epsilon = 0.05$ for ASL[62] on OpenImages.

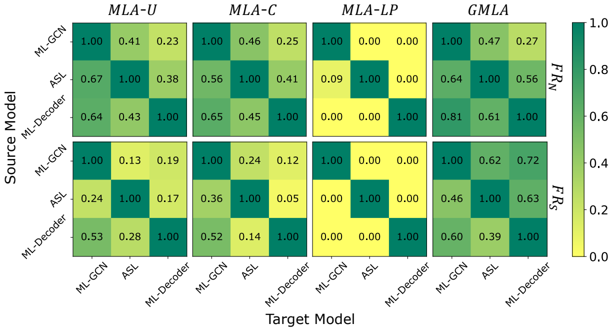


Figure 7. Transferability across models on PASCAL-VOC. The y-axis shows the source model which generates the perturbation and x-axis shows the target model evaluated on that perturbation.

Note that FR_N, FR_S, and $SSIM$ should be high while NT_R should be low for a good attack method.

5.2. Experimental Results

Figure 5 shows the performance of different attack methods on PASCAL-VOC for one- and two-node attacks for different epsilon values using ML-GCN classifier. In Table 1, we show the evaluation across the three MLL models for a fixed $\epsilon = 0.01$ for which the performance of all attacks has plateaued[3]. We also show the evaluation on OpenImages for different target sizes in Fig. 6 and Tab. 2. From the results, we make the following conclusions:

– As Fig. 5 shows, all methods achieve high naive fooling rate FR_N given large enough perturbation budget, yet once we filter out the attacks leading to semantically inconsistent predictions, the performance (FR_S) of all baselines significantly decreases. However, our GMLA achieves very high semantic-based fooling rate than baselines. From Tab. 1 and 2, our method achieves naive fooling rate FR_N comparable to the other methods but outperforms them over FR_S by a significant margin.

– Notice from Fig. 5 and 6 that MLA-U has higher naive and semantic-based fooling rates than MLA-C. The reason is the strong positive correlations learned among related co-occurring labels during model training, which MLA-U implicitly exploits. However, MLA-U being oblivious to the relationships among labels can inevitably affect unrelated labels, as shown in Tab. 1 and 2. This explains why MLA-U has the highest NT_R across different settings. The difference becomes more apparent as we move to attack larger datasets e.g. OpenImages. This is because, a larger number of labels increases the chances of learning spurious correlations among unrelated labels.

– Based on Fig. 5, MLA-LP achieves lowest performance compared to other attack methods for both fooling rates on PASCAL-VOC and NUS-WIDE datasets, and does not converge for OpenImages experiments. This is because MLA-LP uses interior point method at each iteration to solve a system of equations, which define the constraints on the target and non-target labels. Because of the complex relationships among different labels, the feasible region for the given linear problem might be empty. This has also been identified by [94]. When the LP problem has a feasible solution, MLA-LP successfully finds the perturbation that satisfy the attack constraints. This explains why, for the small number of successfully attacked images, MLA-LP affects the least percentage of non-targeted labels, achieving low NT$_R$ as shown in Tab. 1.

– Each attack method produces imperceptible perturbations, as we constrain the maximum infinity norm of the perturbation to 0.01 (on images with pixel values between 0 to 1). Notice also from Table 1 that the average SSIM scores between the adversarial and original images is very close to 1, showing imperceptibility of perturbations.

– Notice from Fig. 6 that MLA-C fails to successfully attack large-scale datasets and its performance drops drastically as we increase the target set size. As mentioned earlier, this is attributed to the observation that gradients of target and non-targeted classes are often opposite (as shown in Fig. 8) and as MLA-C optimizes the target and non-target loss simultaneously, the resulting perturbations are sub-optimal. From Tab. 2, MLA-C achieves lowest NT$_R$ for target sizes greater than 2 but also performs poorly on fooling rates. Note that despite achieving high fooling rates FR$_N$ and FR$_G$, our GMLA method affects very small percentage of semantically unrelated labels, which shows the success of our constraint proposed in (6).

Attack Transferability. Figure 7 shows the cross-model transferability of different attacks. For each source model,

[3] We show results of ablation experiment on GMLA in supplementary.

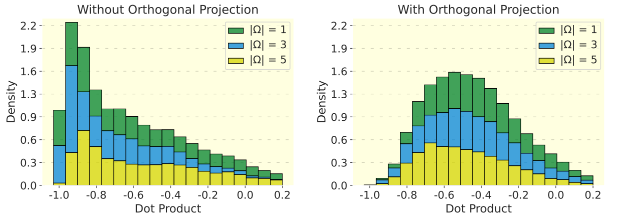
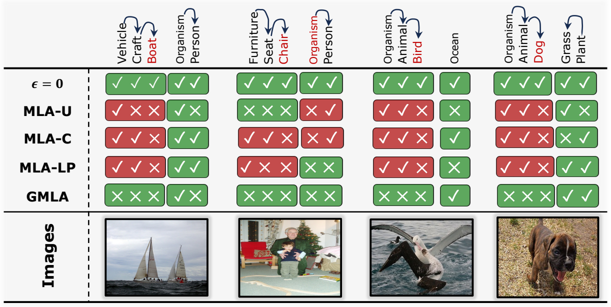


Figure 8. Stacked bar charts showing the correlation between the gradient of the loss on target labels $\boldsymbol{g}_{\boldsymbol{x},\Psi_{\boldsymbol{x}}}$ and on other labels $\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}$ for different sizes of the target set on OpenImages. **Left:** Using (3) as objective. **Right:** using our proposed (6) that optimizes the loss on target labels while keeping the loss on non-target labels the same (as a constraint).



	Vehicle	Craft	Boat	Organism	Person	Furniture	Seat	Chair	Organism	Person	Organism	Animal	Bird	Ocean	Organism	Animal	Dog	Grass	Plant
$\epsilon = 0$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MLA-U	✓	×	×	✓	×	×	×	×	×	✓	✓	✓	×	×	✓	✓	×	✓	×
MLA-C	✓	✓	×	✓	✓	✓	✓	×	×	✓	✓	✓	×	✓	✓	✓	×	×	✓
MLA-LP	✓	✓	×	✓	✓	✓	×	×	×	×	✓	✓	×	×	✓	✓	×	✓	✓
GMLA	×	×	×	✓	×	×	×	×	×	×	×	×	×	✓	×	×	×	✓	✓

Figure 9. Results of attacking ML-GCN on PASCAL-VOC (first two columns) and NUS-WIDE (last two columns). Each column shows the model predictions for clean ($\epsilon = 0$) and attacked images. Rounded rectangles group semantically related labels. Inconsistent predictions caused around target labels are shown with red rectangles. The red labels at the top are targeted labels and the arrows show the relationships.

we compute the perturbations (scaled to $\epsilon = 0.1$) for images and evaluate the target models exclusively on the images that were successfully attacked by the respective source model (hence the diagonal values are all 1). Notice that although all attacks, other than MLA-LP, are transferable, GMLA semantic attack transfers better and achieves the highest FR_N and FR_S. From Table 1, notice that all attacks were able to achieve non-trivial graph-based fooling rate. However, GMLA is the most effective method to generate semantically consistent and generally transferrable attacks.

Gradient Correlations. Figure 8 shows the correlation between the gradient of the loss on target labels (to be modified), $\boldsymbol{g}_{\boldsymbol{x},\Psi_{\boldsymbol{x}}}$, and on other labels (to be fixed), $\boldsymbol{g}_{\boldsymbol{x},\bar{\Psi}_{\boldsymbol{x}}}$, for different sizes of the target set on OpenImages. Notice that adding the two losses leads to highly negatively correlated gradients for them. However, only optimizing the loss on target labels while keeping the loss on non-target labels the same (as a constraint) leads to significant increase in gradient correlations, which can justify the success of GMLA.

Qualitative Results. Figure 9 shows qualitative results of attacking ML-GCN using PASCAL-VOC and NUS-WIDE. Notice that in all four cases, respectively, MLA-U and MLA-C lead to inconsistencies. For example, to turn off the *boat* label in the first image, MLA-U attacks the *boat* and *craft* labels but does not attack the *vehicle* label, leading to semantically inconsistent prediction. MLA-C successfully attacks *boat*, but keeps all other labels fixed, causing inconsistent predictions. For the second image, MLA-U successfully kept consistency around one group of labels but causes inconsistency in the other group. Similar to MLA-C, MLA-LP causes semantic inconsistencies for all images. Notice that in all cases, GMLA successfully modifies the necessary labels to ensure semantic consistency.

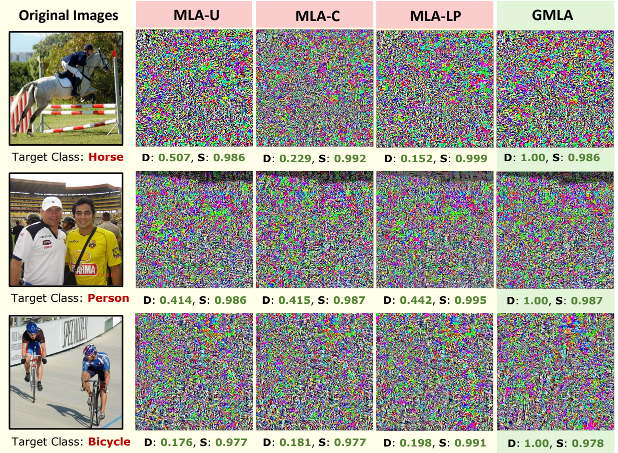


Figure 10. Since the adversarial images have imperceptible changes, we visualize the perturbations computed using different methods for various target classes of PASCAL-VOC. The perturbations are computed by setting the maximum budget $\epsilon = 0.01$ and are scaled for visualization. For each perturbation, we compute it's dot product **(D)** with the perturbation computed using our proposed attack - GMLA, and the structural similarity **(S)** of the original and the adversarial image (after adding the perturbation).

In Figure 10, we visualize the perturbations computed by different methods and compare the SSIM (**S**) of baselines with GMLA. We also show the dot product (**D**) between the perturbation computed using each baseline method and the one computed using GMLA. We can see that GMLA finds different attack directions than the baseline methods, which results in semantically consistent and transferable attacks.

6. Conclusions

We developed an efficient framework to generate attacks for multi-label recognition that ensures semantic consistency of the output labels based on relationships among labels while effectively attacking a large number of labels. By extensive experiments on three datasets and several MLL models, we showed that our method generates both semantically consistent and successful adversarial attacks.

Acknowledgements

This work is sponsored by Khoury College of Northeastern funds, DARPA (HR00112220001), NSF (IIS-2115110), ARO (W911NF2110276). Content does not necessarily reflect the position/policy of the Government.

References

[1] Abhishek Aich, Calvin-Khang Ta, Akash Gupta, Chengyu Song, Srikanth Krishnamurthy, Salman Asif, and Amit Roy-Chowdhury. Gama: Generative adversarial multi-object scene attacks. *Advances in Neural Information Processing Systems*, 35:36914–36930, 2022. 1, 3

[2] Abhishek Aich, Shasha Li, Chengyu Song, M Salman Asif, Srikanth V Krishnamurthy, and Amit K Roy-Chowdhury. Leveraging local patch differences in multi-object scenes for generative adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1308–1318, 2023. 1, 3

[3] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *arXiv*, 2018. 2

[4] A. Athalye, N. Carlini, and D. A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. 2018. 2

[5] Yuanhao Ban and Yinpeng Dong. Pre-trained adversarial perturbations. In *Advances in Neural Information Processing Systems*, pages 1196–1209. Curran Associates, Inc., 2022. 2

[6] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. 2022. 2

[7] Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, and M. Salman Asif. Context-aware transfer attacks for object detection. *ArXiv*, 2021. 3

[8] Zikui Cai, Shantanu Rane, Alejandro E. Brito, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, and M. Salman Asif. Zero-query transfer attacks on context-aware object detectors. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[9] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *Workshop on Artificial Intelligence and Security*, 2017. 1

[10] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, 2017. 2

[11] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi. Unlabeled data improves adversarial robustness. *Neural Information Processing Systems*, 2019. 2

[12] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *AAAI Conference on Artificial Intelligence*, 2018. 2

[13] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. *IEEE International Conference on Multimedia and Expo*, 2019. 1

[14] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. T. Zheng. Nus-wide: A real-world web image database from national university of singapore. *ACM International Conference on Image and Video Retrieval*, 2009. 5

[15] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *ArXiv*, 2020. 4

[16] S. D. Dao, E. Zhao, D. Phung, and J. Cai. Multi-label image classification with contrastive learning. *arXiv preprint, arXiv:2107.11626*, 2021. 2

[17] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. *European Conference on Computer Vision*, 2014. 2

[18] G. W. Ding, Y. Sharma, K. Y. Lui, and R. Huang. Max-margin adversarial (mma) training: Direct input space margin maximization through adversarial training. *arXiv*, 2020. 2

[19] Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Jungong Han. Exploring structured semantic prior for multi label recognition with incomplete labels. 2023. 1

[20] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24678–24687, 2023. 2

[21] Y. Dong, Z. Deng, T. Pang, H. Su, and J. Zhu. Adversarial distributional training for robust deep learning. *arXiv*, 2020. 2

[22] M. Everingham, S. M. A. Eslami, L. Van-Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2010. 5

[23] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, T. Kohno, and D. X. Song. Physical adversarial examples for object detectors. *arXiv*, 2018. 1

[24] L. Feng, B. An, and S. He. Collaboration based multi-label learning. *AAAI Conference on Artificial Intelligence*, 2019. 1

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015. 1

[26] W. He, J. Wei, X. Chen, N. Carlini, and D. Song. Adversarial example defense: Ensembles of weak defenses are not strong. *USENIX Workshop on Offensive Technologies*, 2017. 2

[27] D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. 2019. 2

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1

[29] Lei Hsiung, Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24658–24667, 2023. 1

[30] S. Hu, L. Ke, X. Wang, and S. Lyu. Tkml-ap: Adversarial attacks to top-k multi-label learning. *arXiv*, 2021. 2

[31] D. T. Huynh and E. Elhamifar. Interactive multi-label cnn learning with partial labels. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[32] Tooba Imtiaz, Morgan Kohler, Jared Miller, Zifeng Wang, Mario Sznaier, Octavia I Camps, and Jennifer G Dy. Saif: Sparse adversarial and interpretable attack framework. *arXiv preprint arXiv:2212.07495*, 2022. 2

[33] J. Li R. Ji, H. Liu, X. Hong, Y. Gao, and Q. Tian. Universal perturbation attack against image retrieval. *International Conference on Computer Vision*, 2019. 1

[34] Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. Multiguard: Provably robust multi-label classification against adversarial examples. *Advances in Neural Information Processing Systems*, 2022. 1, 3

[35] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. 2022. 1

[36] Takumi Kobayashi. Two-way multi-label loss. 2023. 1

[37] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *International Conference on Learning Representations*, 2017. 1

[38] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2016. 2, 4, 5

[39] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi. General multi-label image classification with transformers. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[40] K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. 2018. 2

[41] Peng Li, Peng Chen, Yonghong Xie, and Dezheng Zhang. Bi-modal learning with channel-wise attention for multi-label image classification. *IEEE Access*, 2020. 2

[42] Q. Li, M. Qiao, W. Bian, and D. Tao. Conditional graphical lasso for multi-label image classification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[43] Q. Li, X. Peng, Y. Qiao, and Q. Peng. Learning label correlations for multi-label image recognition with graph networks. *Pattern Recognition Letters*, 2020. 2

[44] X. Li, F. Zhao, and Y. Guo. Multi-label image classification with a probabilistic label enhancement model. In *UAI*, 2014. 2

[45] Y. Li and L. Yang. More correlations better performance: Fully associative networks for multi-label image classification. *International Conference on Pattern Recognition*, 2021. 2

[46] Y. Li, Y. Song, and J. Luo. Improving pairwise ranking for multi-label image classification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[47] Z. Li, W. Lu, Z. Sun, and W. Xing. Improving multi-label classification using scene cues. *Multimedia Tools and Applications*, 2017. 2

[48] Dekun Lin. Probability guided loss for long-tailed multi-label image classification. 2023. 1

[49] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. 2

[50] S. Melacci, G. Ciravegna, A. Sotgiu, A. Demontis, B. Biggio, M. Gori, and F. Roli. Domain knowledge alleviates adversarial attacks in multi-label classifiers. 2021. 3

[51] J.-H. Metzen, M.-C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. *International Conference on Computer Vision*, 2019. 1

[52] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 1995. 5

[53] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *IEEEConference on Computer Vision and Pattern Recognition*, 2016. 6

[54] J. Nam, E. L. Mencía, H. J. Kim, and J. Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. *Neural Information Processing Systems*, 2017. 2

[55] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu. Improving adversarial robustness via promoting ensemble diversity. *International Conference on Machine learning*, 2019. 2

[56] T. Pang, K. Xu, Y. Dong, C. Du, N. Chen, and J. Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv*, 2020.

[57] T. Pang, X. Yang, Y. Dong, K. Xu, H. Su, and J. Zhu. Boosting adversarial training with hypersphere embedding. *Neural Information Processing Systems*, 2020. 2

[58] Nicolas Papernot, Patrick Mcdaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016. 1

[59] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. 2022. 1

[60] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *Advances in Neural Information Processing Systems*, 35:29845–29858, 2022. 2

[61] T. Ridnik, H. Lawen, A. Noy, and I. Friedman. Tresnet: High performance gpu-dedicated architecture. *ArXiv preprint arXiv:2003.13630*, 2020. 5

[62] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. 2021. 5, 6, 7

[63] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. 2023. 5, 6

[64] Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4322–4330, 2019. 2

[65] J. Cohenand E. Rosenfeld and Z. Kolter. Certified adversarial robustness via randomized smoothing. *International Conference on Machine learning*, 2019. 2

[66] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *Neural Information Processing Systems*, 2019.

[67] Nasim Shafiee and Ehsan Elhamifar. Zero-shot attribute attacks on fine-grained recognition models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 262–282. Springer, 2022. 1, 2

[68] Nitish Shukla and Sudipta Banerjee. Generating adversarial attacks in the latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 730–739, 2023. 1

[69] Q. Song, H. Jin, X. Huang, and X. Hu. Multi-label adversarial perturbations. *IEEE International Conference on Data Mining*, 2018. 1, 2, 3, 6

[70] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. 2017. 5

[71] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014. 2

[72] N. Tursynbek, A. Petiushko, and I. Oseledets. Geometry-inspired top-k adversarial perturbations. *arXiv*, 2020. 2

[73] J. Uesato, J. B. Alayrac, P. Huang, R. Stanforth, A. Fawzi, and P. Kohli. Are labels required for improving adversarial robustness? *Neural Information Processing Systems*, 2019. 2

[74] Thomas Verelst, Paul K Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. Spatial consistency loss for training multi-label classifiers from single-label annotations. 2023. 2

[75] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[76] Z. Wang, T. Chen, G. Li, G. Li, and L. Lin. Multi-label image recognition by recurrently discovering attentional regions. *IEEE International Conference on Computer Vision*, 2017. 2

[77] Y. Wu, H. Liu, S. Feng, Y. Jin, G. Lyu, and Z. Wu. Gm-mlic: Graph matching based multi-label image classification. *International Joint Conference on Artificial Intelligence*, 2021. 2

[78] Z. Wu and M. Palmer. Verbs semantics and lexical selection. *Annual Meeting on Association for Computational Linguistics*, 1994. 5

[79] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. Yuille. Improving transferability of adversarial examples with input diversity. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[80] Ming-Kun Xie, Jiahao Xiao, and Sheng-Jun Huang. Label-aware global consistency for multi-label learning with single positive labels. 2022. 1

[81] J. Xu, H. Tian, Z. Wang, Y. Wang, W. Kang, and F. Chen. Joint input and output space learning for multi-label image classification. *IEEE Transactions on Multimedia*, 2020. 2

[82] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Network and Distributed Systems Security Symposium*, 2018. 2

[83] H. Yang, J. T. Zhou, Y. Zhang, B. Gao, J. Wu, and J. Cai. Exploit bounding box annotations for multi-label object recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[84] Zhuo Yang, Yufei Han, and Xiangliang Zhang. Characterizing the evasion attackability of multi-label classifiers. 2021. 1, 2

[85] Z. Yang, Y. Han, and X. Zhang. Attack transferability characterization for adversarially robust multi-label classification. 2021. 1, 2

[86] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. *European Conference on Computer Vision*, 2020. 2

[87] R. You, Z. Guo, L. Cui, X. Long, S. Y. Bao, and S. Wen. Cross-modality attention with semantic graph embedding for multi-label classification. *AAAI Conference on Artificial Intelligence*, 2020. 2

[88] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2019. 2

[89] ML. Zhang and Z. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 2006. 2

[90] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. 2022. 2

[91] Z. Zhao, G. Chen, J. Wang, Y. Yang, F. Song, and J. Sun. Attack as defense: Characterizing adversarial examples using robustness. *arXiv*, 2021. 2

[92] Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. 2022. 1

[93] N. Zhou, W. Luo, X. Lin, P. Xu, and Z.. Zhang. Generating multi-label adversarial examples by linear programming. *International Joint Conference on Neural Networks*, 2020. 3

[94] N. Zhou, W. Luo, J. Zhang, L. Kong, and H. Zhang. Hiding all labels for multi-label images: An empirical study of adversarial examples. *International Joint Conference on Neural Networks*, 2021. 2, 7

[95] Y. Zhu, J. T. Kwok, and Z. Zhou. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 2018. 2

[96] D. Zügner, A. Akbarnejad, and S. Günnemann. Adversarial attacks on neural networks for graph data. *International Conference on Knowledge Discovery & Data Mining*, 2018. 1