A. Proofs

A.1. Proof of Corollary 1

Before we prove Corollary 1, We first introduce the following lemma.

Lemma 1. *For $0 < q < p$, the following inequality holds:*

$$\|\boldsymbol{x}\|_q \leq d^{\frac{1}{q}-\frac{1}{p}} \|\boldsymbol{x}\|_p \tag{13}$$

where $\boldsymbol{x} \in \mathbb{R}^d$.

Proof. Consider $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, using the Hölder's Inequality that for m, n satisfying $\frac{1}{m} + \frac{1}{n} = 1$,

$$\sum_i |u_i||v_i| \leq \left(\sum_i |u_i|^m\right)^{\frac{1}{m}} \left(\sum_i |v_i|^n\right)^{\frac{1}{n}}. \tag{14}$$

If we take $|u_i| = |x_i|^q$, $v_i = 1$, $m = \frac{p}{q}$ and $n = \frac{p}{p-q}$, we get

$$\sum_i |x_i|^q \leq \left(\sum_i |x_i|^p\right)^{\frac{q}{p}} d^{\frac{p-q}{p}} \tag{15}$$

By taking the power of $\frac{1}{q}$ on both sides, we have

$$\left(\sum_i |x_i|^q\right)^{\frac{1}{q}} \leq \left(\sum_i |x_i|^p\right)^{\frac{1}{p}} d^{\frac{1}{q}-\frac{1}{p}} \tag{16}$$

which concludes the proof. □

Corollary 1. *Given a twice-differentiable classifier $f : \mathbb{R}^d \to \mathbb{R}^k$, and its attribution g^y on label y, assume that g^y is locally linear within the neighborhood of $\boldsymbol{x}$, $\mathcal{B}_\varepsilon(\boldsymbol{x}) = \{\boldsymbol{x} + \boldsymbol{\delta} | \|\boldsymbol{\delta}\|_p \leq \varepsilon\}$, then for all perturbations $\|\boldsymbol{\delta}\|_p \leq \varepsilon$ that $p > 2$, $\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2 \leq d^{\frac{1}{2}-\frac{1}{p}} \xi_{max} \varepsilon$, where ξ_{max} is the largest singular value of $H = \nabla g^y(\boldsymbol{x})$.*

Proof. Using Lemma 1, we have $\|\boldsymbol{\delta}\|_2 \leq d^{\frac{1}{2}-\frac{1}{p}} \|\boldsymbol{\delta}\|_p$. Similar to the proof of Theorem 1,

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2^2 \leq \lambda_{max} \|\boldsymbol{\delta}\|_2^2 \leq \lambda_{max} \left(d^{\frac{1}{2}-\frac{1}{p}} \|\boldsymbol{\delta}\|_p\right)^2 \leq \lambda_{max} \left(d^{\frac{1}{2}-\frac{1}{p}} \varepsilon\right)^2 \tag{17}$$

Therefore,

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2 \leq d^{\frac{1}{2}-\frac{1}{p}} \xi_{max} \varepsilon \tag{18}$$

□

A.2. Proof of Theorem 2

Theorem 2. *Given a twice-differentiable classifier f, its attribution on label y, g^y, and the gradient $H = \nabla g^y$, assume that g^y is locally linear within the neighborhood of $\boldsymbol{x}$, $\mathcal{B}_\varepsilon(\boldsymbol{x}) = \{\boldsymbol{x} + \boldsymbol{\delta} | \|\boldsymbol{\delta}\|_\infty \leq \varepsilon\}$, then for all perturbations $\|\boldsymbol{\delta}\|_\infty \leq \varepsilon$,*

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2 \leq \varepsilon \sqrt{\sum_{i,j} |P_{ij}|}. \tag{7}$$

where $P = HH^\top$ and the equality is taken at $\boldsymbol{\delta} = (\pm\varepsilon, \ldots, \pm\varepsilon)^\top$.

Proof. Recall that under the local linearity assumption,

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2^2 \leq \boldsymbol{\delta}^\top P \boldsymbol{\delta} = \sum_{i,j} P_{ij} \delta_i \delta_j. \tag{19}$$

Since $P_{ij} \leq |P_{ij}|$ and $\delta_i \delta_j \leq \|\boldsymbol{\delta}\|_\infty^2 \leq \varepsilon^2$ for all i, j, we can easily prove the theorem that

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2^2 \leq \varepsilon^2 \sum_{i,j} |P_{ij}|. \tag{20}$$

□

A.3. Proof of Proposition 1

Proposition 1. *Denote the gradient-based attribution satisfying the completeness axiom of $\boldsymbol{x}$ on ground truth label y by $g^y(\boldsymbol{x})$, and the attribution on a different label y' by $g^{y'}(\boldsymbol{x})$. Given the perturbation $\boldsymbol{\delta}$, assume that g^y is locally linear within the neighborhood of $\boldsymbol{x}$, $\mathcal{B}_\varepsilon(\boldsymbol{x}) = \{\boldsymbol{x}+\boldsymbol{\delta} \| \|\boldsymbol{\delta}\|_p \leq \varepsilon\}$, the classification result of $\boldsymbol{x}+\boldsymbol{\delta}$ does not change from y to y' if*

$$\left(\left(\nabla g^{y'}(\boldsymbol{x}) - \nabla g^y(\boldsymbol{x})\right)\Delta\right)^\top \boldsymbol{\delta} < f_y(\boldsymbol{x}) - f_{y'}(\boldsymbol{x}), \tag{8}$$

where Δ is an all one vector, $\Delta = (1,\dots,1)^\top \in \mathbb{R}^d$.

Proof. Recall that we denote the gradient-based attribution satisfying the completeness axiom of $\boldsymbol{x}$ on target label y by $g^y(\boldsymbol{x})$, *e.g.*, integrated gradients. Similarly, we denote the attribution on a different label y' by $g^{y'}(\boldsymbol{x})$. Given the perturbation $\boldsymbol{\delta}$, according to the above assumption, we can write that

$$g^y(\boldsymbol{x}+\boldsymbol{\delta}) = g^y(\boldsymbol{x}) + \nabla g^y(\boldsymbol{x})^\top \boldsymbol{\delta} \tag{21}$$

Similarly, the approximation of $g^{y'}(\boldsymbol{x}+\boldsymbol{\delta})$ is given by:

$$g^{y'}(\boldsymbol{x}+\boldsymbol{\delta}) = g^{y'}(\boldsymbol{x}) + \nabla g^{y'}(\boldsymbol{x})^\top \boldsymbol{\delta} \tag{22}$$

According to the completeness axiom, given an all one vector $\Delta = (1,\dots,1)^\top$, we have

$$\Delta^\top g^y(\boldsymbol{x}) = f_y(\boldsymbol{x}). \tag{23}$$

Consider the perturbation $\boldsymbol{\delta}$, if $\boldsymbol{\delta}$ does not change the label of $\boldsymbol{x}$ from y to y', then $f_{y'}(\boldsymbol{x}+\boldsymbol{\delta}) < f_y(\boldsymbol{x}+\boldsymbol{\delta})$, *i.e.*,

$$\Delta^\top g^{y'}(\boldsymbol{x}+\boldsymbol{\delta}) < \Delta^\top g^y(\boldsymbol{x}+\boldsymbol{\delta}), \tag{24}$$

which gives

$$\Delta^\top g^{y'}(\boldsymbol{x}) + \Delta^\top \nabla g^{y'}(\boldsymbol{x})^\top \boldsymbol{\delta} < \Delta^\top g^y(\boldsymbol{x}) + \Delta^\top \nabla g^y(\boldsymbol{x})^\top \boldsymbol{\delta}. \tag{25}$$

By rearranging the above inequality, we have

$$\left(\left(\nabla g^{y'}(\boldsymbol{x}) - \nabla g^y(\boldsymbol{x})\right)\Delta\right)^\top \boldsymbol{\delta} < f_y(\boldsymbol{x}) - f_{y'}(\boldsymbol{x}). \tag{26}$$

□

A.4. Proof of Corollary 2

Corollary 2. *Given a twice-differentiable classifier $f : \mathbb{R}^d \to \mathbb{R}^k$ and its attribution g^y on label y, for all perturbations $\|\boldsymbol{\delta}\|_p \leq \varepsilon$, if the Euclidean distance of $g^y(\boldsymbol{x}+\boldsymbol{\delta})$ and $g^y(\boldsymbol{x})$ is upper bounded by $T(\varepsilon;\boldsymbol{x})$, and $0 \leq T(\varepsilon;\boldsymbol{x}) \leq \|g^y(\boldsymbol{x})\|_2$, then their cosine distance (D_c) is upper bounded by*

$$D_c(g^y(\boldsymbol{x}+\boldsymbol{\delta}), g^y(\boldsymbol{x})) \leq 1 - \sqrt{1 - \frac{T(\varepsilon;\boldsymbol{x})^2}{\|g^y(\boldsymbol{x})\|_2^2}}. \tag{10}$$

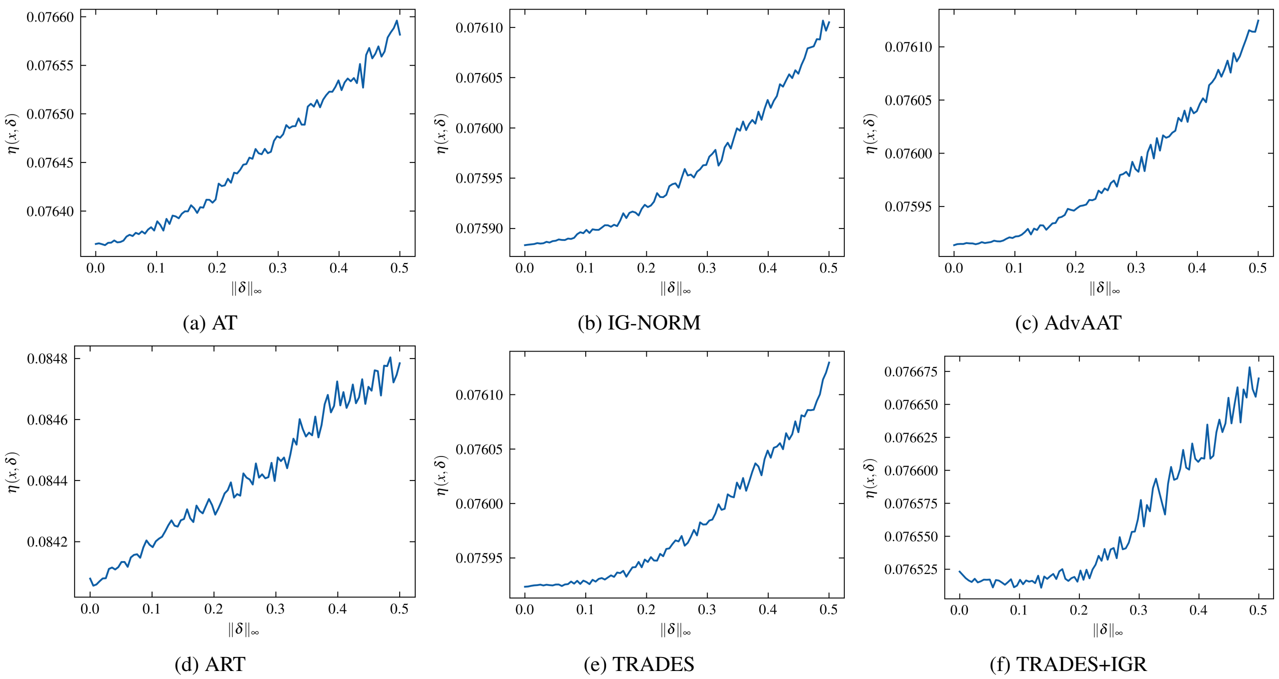
Proof. The corollary can be proved using the geometric property (see Fig. 1a) that

$$\sin(g^y(\boldsymbol{x}+\boldsymbol{\delta}), g^y(\boldsymbol{x})) \leq \frac{T(\varepsilon;\boldsymbol{x})}{\|g^y(\boldsymbol{x})\|_2}, \tag{27}$$

and,

$$\begin{aligned}
\operatorname{cosd}(g^y(\boldsymbol{x}+\boldsymbol{\delta}), g^y(\boldsymbol{x})) &= 1 - \cos(g^y(\boldsymbol{x}+\boldsymbol{\delta}), g^y(\boldsymbol{x})) &(28)\\
&= 1 - \sqrt{1 - \sin^2(g^y(\boldsymbol{x}+\boldsymbol{\delta}), g^y(\boldsymbol{x}))} &(29)\\
&\leq 1 - \sqrt{1 - \frac{T(\varepsilon;\boldsymbol{x})^2}{\|g^y(\boldsymbol{x})\|_2^2}} &(30)
\end{aligned}$$

□



(a) AT

(b) IG-NORM

(c) AdvAAT

(d) ART

(e) TRADES

(f) TRADES+IGR

Figure 3. Values of η for different $\|\boldsymbol{\delta}\|_\infty$ computed from CIFAR-10 using integrated gradients. The magnitudes are ranging from 0.07 to 0.09 and are negligible comparing with the average norm of attributions which is 3.47 on CIFAR-10.

B. Analysis of local linearity assumption

B.1. Evaluation of local linearity assumption of attribution functions

The theories of this work are based on the local linearity assumption that $g^y(\boldsymbol{x})$ is linear within $\mathcal{B}_\varepsilon(\boldsymbol{x}) = \{\boldsymbol{x}+\boldsymbol{\delta} | \|\boldsymbol{\delta}\|_p \leq \varepsilon\}$. It is worth noting that such local linearity is a valid assumption for smooth functions, which can be achieved by both adversarial and attributional robust methods. Adversarial defense methods look for locally linearity functions to reduce the impact of adversarial attacks [22, 35]. Similarly, attributional defense methods train for smooth gradients to defend against attribution attacks [33]. It is also a common practice in related literature [7, 11, 16, 27, 38] to make similar assumptions.

Furthermore, the validity of this assumption also depends on the size of $\boldsymbol{\delta}$. The perturbation $\boldsymbol{\delta}$ is restricted within a small ℓ_p ball around $\boldsymbol{x}$ to ensure that the perturbed images are visually indistinguishable comparing to its original counterpart. The maximum allowable size ε for $\boldsymbol{\delta}$ is relatively small compared with the intensity range of the original image. When $\boldsymbol{\delta}$ is small, the remainder of the Taylor series of $g^y(\boldsymbol{x})$ is negligible and the local linearity assumption is valid. As shown in Figure 3, the value of $\eta(\boldsymbol{x},\boldsymbol{\delta}) = \|g^y(\boldsymbol{x}) - g^y(\boldsymbol{x}+\boldsymbol{\delta}) - \boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x})\|_2$ is small and negligible when $\|\boldsymbol{\delta}\|_\infty$ is small.

B.2. Generalization of Theorem 1

Theorem 3. *Given a twice-differentiable classifier $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$, and its attribution g^y on label y, denote the Taylor series of $g^y(\boldsymbol{x}+\boldsymbol{\delta})$ as $g^y(\boldsymbol{x}) + \boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x}) + R_1(\boldsymbol{x})$. If $-(c-1)\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x}) \preceq R_1(\boldsymbol{x}) \preceq (c-1)\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x})$ for a constant $c \geq 1$, where $\preceq$ refers to element-wise less than or equal to, then for all perturbations $\|\boldsymbol{\delta}\|_2 \leq \varepsilon$,*

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2 \leq c\xi_{max}\varepsilon,$$

where ξ_{max} is the largest singular value of $H = \nabla g^y(\boldsymbol{x})$.

Proof. Based on the Taylor series of $g^y(\boldsymbol{x})$ and the above condition, we have

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2^2 \leq \|\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x}) + (c-1)\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x})\|_2^2 = c^2 \boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x}) \nabla g^y(\boldsymbol{x})^\top \boldsymbol{\delta} \tag{31}$$

$$= c^2 \frac{\boldsymbol{\delta}^\top}{\|\boldsymbol{\delta}\|_2} P \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2} \cdot \|\boldsymbol{\delta}\|_2^2 \tag{32}$$

$$\leq c^2 \lambda_{max} \|\boldsymbol{\delta}\|_2^2 \leq c^2 \lambda_{max} \varepsilon^2 \tag{33}$$

where λ_{max} is the largest eigenvalue of $P = HH^\top = \nabla g^y(\boldsymbol{x})\nabla g^y(\boldsymbol{x})^\top$, and $\boldsymbol{v}_{max}$ is the corresponding eigenvector. The equality in Eq. 33 is achieved when $\boldsymbol{\delta}$ is $\varepsilon \boldsymbol{v}_{max}$ or $-\varepsilon \boldsymbol{v}_{max}$. Since the singular values of H are equal to the square root of the eigenvalues of P, then,

$$\|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2 \leq c\sqrt{\lambda_{max}}\varepsilon = c\xi_{max}\varepsilon. \tag{34}$$

□

This is a generalized version of Theorem 1 that is applicable for all twice-differentiable classifiers. Under local linearity assumption, $R_1(\boldsymbol{x}) = 0$, which means $c = 1$, the result coincides with the original version of Theorem 1.

B.3. Derivation of Eq. (11)

By Taylor expansion, $g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x}) = \boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x}) + R_1(\boldsymbol{x})$, where R_1 is the first order Taylor remainder. Thus, we have

$$\|R_1(\boldsymbol{x})\|_2 \geq \|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2 - \|\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x})\|_2 \tag{35}$$

Take $c = \frac{\|R_1(\boldsymbol{x})\|_2}{\|\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x})\|_2} + 1$,

$$\|\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x})\|_2 + \|R_1(\boldsymbol{x})\|_2 = c\|\boldsymbol{\delta}^\top \nabla g^y(\boldsymbol{x})\|_2, \tag{36}$$

and it would be the worst-case for the linear assumption when $\boldsymbol{\delta} = \varepsilon \boldsymbol{v}_{max}$. By taking $\varepsilon \boldsymbol{v}_{max}$ as $\boldsymbol{\delta}$, $\|R_1(\boldsymbol{x})\|_2$ can be estimated by

$$\max\left\{0, \|g^y(\boldsymbol{x}+\varepsilon\boldsymbol{v}_{max}) - g^y(\boldsymbol{x})\|_2 - \|\varepsilon\boldsymbol{v}_{max}^\top \nabla g^y(\boldsymbol{x})\|_2\right\}. \tag{37}$$

Since $\|g^y(\boldsymbol{x}+\varepsilon\boldsymbol{v}_{max}) - g^y(\boldsymbol{x})\|_2 - \|\varepsilon\boldsymbol{v}_{max}^\top \nabla g^y(\boldsymbol{x})\|_2 \leq \|R_1(\boldsymbol{x})\|_2$. Putting Eq. (37) into c and using the result in Eq. (6), we have

$$\begin{align} c &= \max\left\{0, \frac{\|g^y(\boldsymbol{x}+\varepsilon\boldsymbol{v}_{max}) - g^y(\boldsymbol{x})\|_2 - \|\varepsilon\boldsymbol{v}_{max}^\top \nabla g^y(\boldsymbol{x})\|_2}{\xi_{max}\varepsilon}\right\} + 1 \tag{38}\\ &= \max\left\{1, \frac{\|g^y(\boldsymbol{x}+\varepsilon\boldsymbol{v}_{max}) - g^y(\boldsymbol{x})\|_2}{\xi_{max}\varepsilon}\right\}. \tag{39}\end{align}$$

C. Analysis of attribution gradients

C.1. The gradient of integrated gradients

We provide the justification showing that the gradient of IG is diagonal-dominated. Consider that

$$\mathrm{IG}(\boldsymbol{x})_i = x_i \times \frac{1}{m}\sum_{\alpha=1}^{m} \frac{\partial f(\frac{\alpha}{m}\boldsymbol{x})}{\partial x_i} \tag{40}$$

and

$$\nabla \mathrm{IG}(\boldsymbol{x})_{ij} = \frac{\partial \mathrm{IG}(\boldsymbol{x})_i}{\partial x_j} \tag{41}$$

If $i \neq j$, then

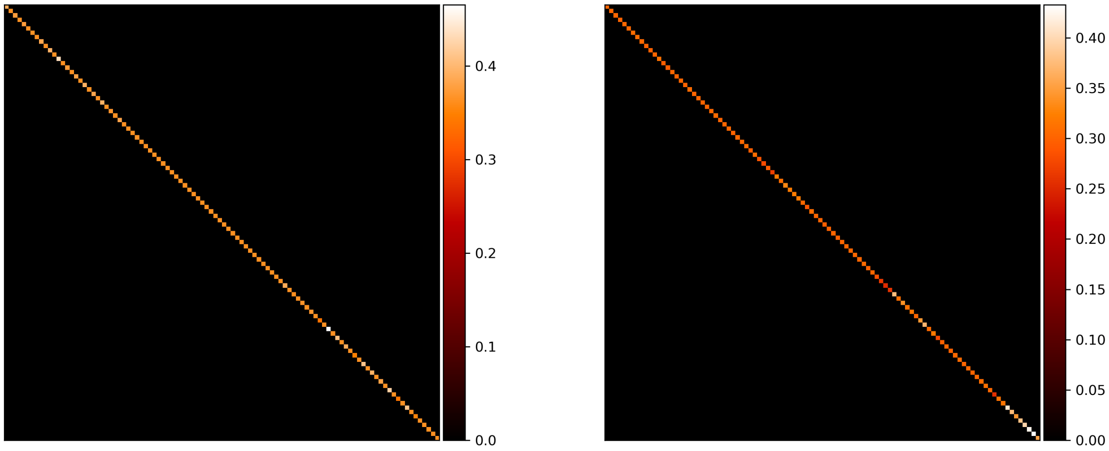
$$\frac{\partial \mathrm{IG}(\boldsymbol{x})_i}{\partial x_j} = x_i \cdot \frac{1}{m}\sum_{\alpha=1}^{m} \frac{\partial^2 f(\frac{\alpha}{m}\boldsymbol{x})}{\partial x_i \partial x_j} \times \frac{\alpha}{m} \tag{42}$$

If $i = j$, then

$$\frac{\partial \mathrm{IG}(\boldsymbol{x})_i}{\partial x_j} = \frac{1}{m}\sum_{\alpha=1}^{m} \frac{\partial f(\frac{\alpha}{m}\boldsymbol{x})}{\partial x_j} + x_i \cdot \frac{1}{m}\sum_{\alpha=1}^{m} \frac{\partial^2 f(\frac{\alpha}{m}\boldsymbol{x})}{\partial x_i \partial x_j} \times \frac{\alpha}{m} \tag{43}$$

Denote that $H_{ij}^{(\alpha)} = \frac{\partial^2 f(\frac{\alpha}{m}\boldsymbol{x})}{\partial x_i \partial x_j}$, *i.e.*, $H^{(\alpha)}$ is the Hessian matrix of $f(\frac{\alpha}{m}\boldsymbol{x})$. Thus

$$\frac{\partial \mathrm{IG}(\boldsymbol{x})_i}{\partial x_j} = \begin{cases} \frac{1}{m}\sum_{\alpha=1}^{m} \nabla f(\frac{\alpha}{m}\boldsymbol{x}) + x_i \cdot \frac{\alpha}{m^2} H_{ij}^{(\alpha)}, & i = j \\ x_i \cdot \sum_{\alpha=1}^{m} \frac{\alpha}{m^2} H_{ij}^{(\alpha)}, & i \neq j \end{cases} \tag{44}$$



(a)

(b)

Figure 4. The first 100 dimensions of gradient attribution generated from (a) MNIST and (b) Fashion-MNIST.

In matrix form,

$$\nabla \text{IG} = \text{diag}\left(\frac{1}{m}\sum_{\alpha=1}^{m}\nabla f(\frac{\alpha}{m}\boldsymbol{x})\right) + [\boldsymbol{x}, \cdots, \boldsymbol{x}] \otimes \frac{\alpha}{m^2}\sum_{\alpha=1}^{m} H^{(\alpha)} \tag{45}$$

If we use softplus as an activation function, *i.e.*, $g(\boldsymbol{x}) = \frac{1}{\beta}\log(1+\exp(\beta\boldsymbol{x}))$, then,

$$g''(\boldsymbol{x}) = \frac{\beta e^{\beta \boldsymbol{x}}}{(e^{\beta \boldsymbol{x}}+1)^2} \tag{46}$$

and

$$\lim_{\beta\to\infty} g''(\boldsymbol{x}) = 0 \tag{47}$$

As $\beta \to \infty$, $H^{(\alpha)}$ will tend to 0, and the second term in Eq. 45 will tend to 0. At the same time, if we choose the number of steps in IG, m larger, $\frac{\alpha}{m^2}$ will converge to 0 faster than $\frac{1}{m}$. Therefore, ∇IG will be diagonal-dominated.

C.2. Additional visualization of attribution gradients

We provide the first 100-dimensions heatmaps of absolute values of attribution gradients, *i.e.*, gradients of IG, on MNIST and Fashion-MNIST in addition to CIFAR-10 presented in Fig. 1b. Moreover, the complete heatmaps for all the three datasets are also presented. As observed in Figs. 4 to 7, the matrices of attribution gradients are diagonal-dominant.

D. Additional experimental results

D.1. Additional results of upper bound on more models without the label constraint

In this subsection, we evaluate the proposed upper bound without the label constraint for the other models, apart from TRADES+IGR in the paper. The perturbation size is chosen to be 0.1 for all evaluations. As in Sec. 5, we use Theorem 1 and 2 to compute $T_e = \xi_{max}\varepsilon$ and extend it to T_c using Eq. 10. The modified upper bound $T'_e = c\xi_{max}\varepsilon$ is also provided to address the inaccurate Taylor approximation (less than 1%). $\widehat{T}_e$ and $\widehat{T}_c$ are computed from the corresponding average attribution differences. The results are given in Table 6. It is shown that the sample distances under both Euclidean and cosine metrics are bounded by T'_e and T_c as expected. All the distortion caused by the attacks *i.e.*, $\widehat{T}_e$ and $\widehat{T}_c$ are smaller than T'_e and T_c.

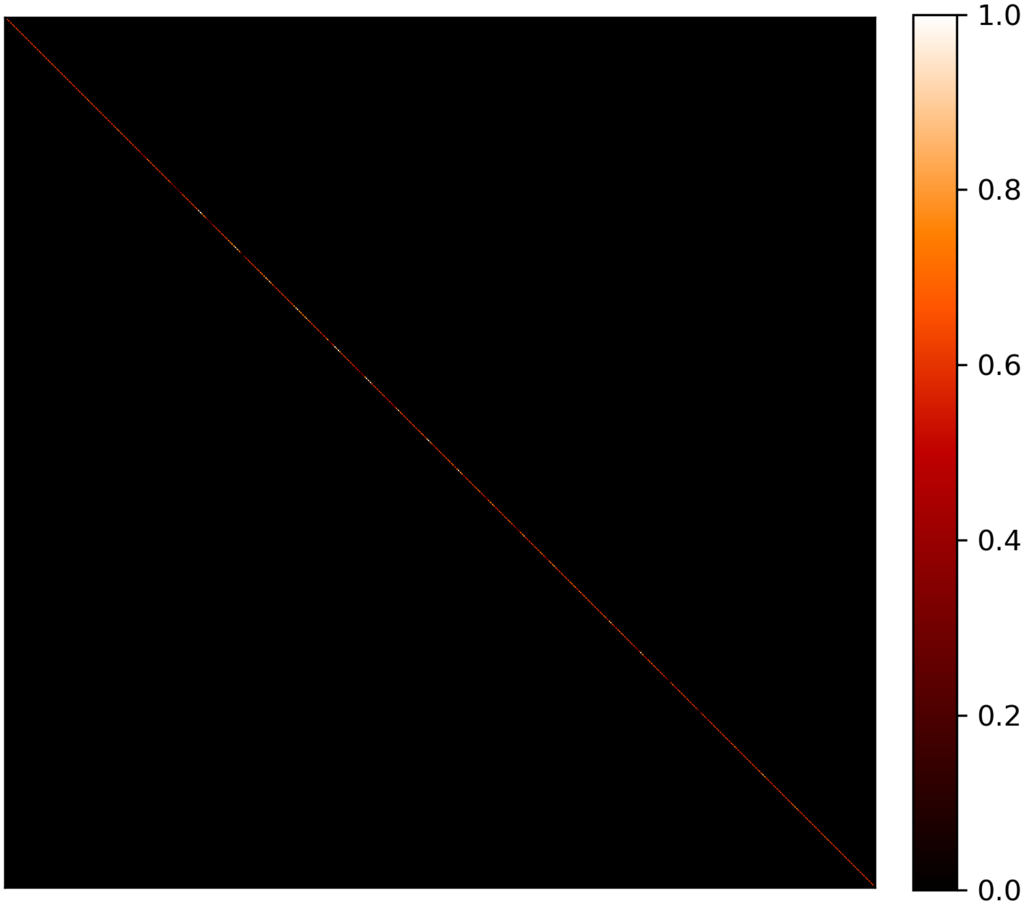


Figure 5. The full heatmap of attribution gradients of MNIST in size 784×784.

Table 6. Evaluation of upper bounds without the label constraint. The cosine distance values $\widehat{T}_c$ and $\widehat{T}'_c$ are converted to degrees for easier comparison.

	SM					Input*gradient					IG				
ℓ_2	$\widehat{T}_e$	T_e	T'_e	$\widehat{T}_c$	T_c	$\widehat{T}_e$	T_e	T'_e	$\widehat{T}_c$	T_c	$\widehat{T}_e$	T_e	T'_e	$\widehat{T}_c$	T_c
AT	0.44	0.94	0.98	9.19	14.87	0.07	0.63	0.63	1.17	4.34	0.04	0.25	0.25	2.73	4.77
IG-NORM	0.03	0.70	0.79	4.33	9.06	0.03	0.50	0.52	1.40	4.75	0.01	0.16	0.16	1.65	4.37
AdvAAT	0.30	1.83	1.83	11.24	20.44	0.08	0.66	0.67	1.84	3.79	0.04	0.24	0.24	0.28	3.82
ART	0.18	0.79	0.81	10.88	14.21	0.09	0.92	0.97	0.83	6.06	0.07	0.23	0.23	0.59	4.21
TRADES	0.11	0.76	0.76	10.01	18.40	0.05	0.48	0.48	1.19	3.20	0.03	0.17	0.17	1.91	3.87
ℓ_∞															
AT	0.55	1.27	-	23.47	30.18	0.63	0.73	-	9.28	61.03	0.41	0.76	-	26.62	45.32
IG-NORM	0.42	0.70	-	25.16	32.60	0.21	0.70	-	6.88	42.94	0.20	0.48	-	21.63	35.30
AdvAAT	0.64	1.83	-	25.20	31.25	0.07	0.74	-	7.79	45.16	0.23	0.52	-	28.73	39.40
ART	0.49	1.01	-	23.81	35.17	0.27	0.79	-	10.21	48.30	0.31	0.67	-	31.01	35.64
TRADES	0.39	0.75	-	22.40	29.10	0.33	0.69	-	9.17	52.63	0.23	0.50	-	22.98	36.38

D.2. Ablation study of upper bound using different ε

In this subsection, we provide more experimental results of the proposed bound on MNIST, Fashion-MNIST and CIFAR-10 in both ℓ_2 and ℓ_∞ cases under label constraint. More specifically, for MNIST and Fashion-MNIST, we additionally

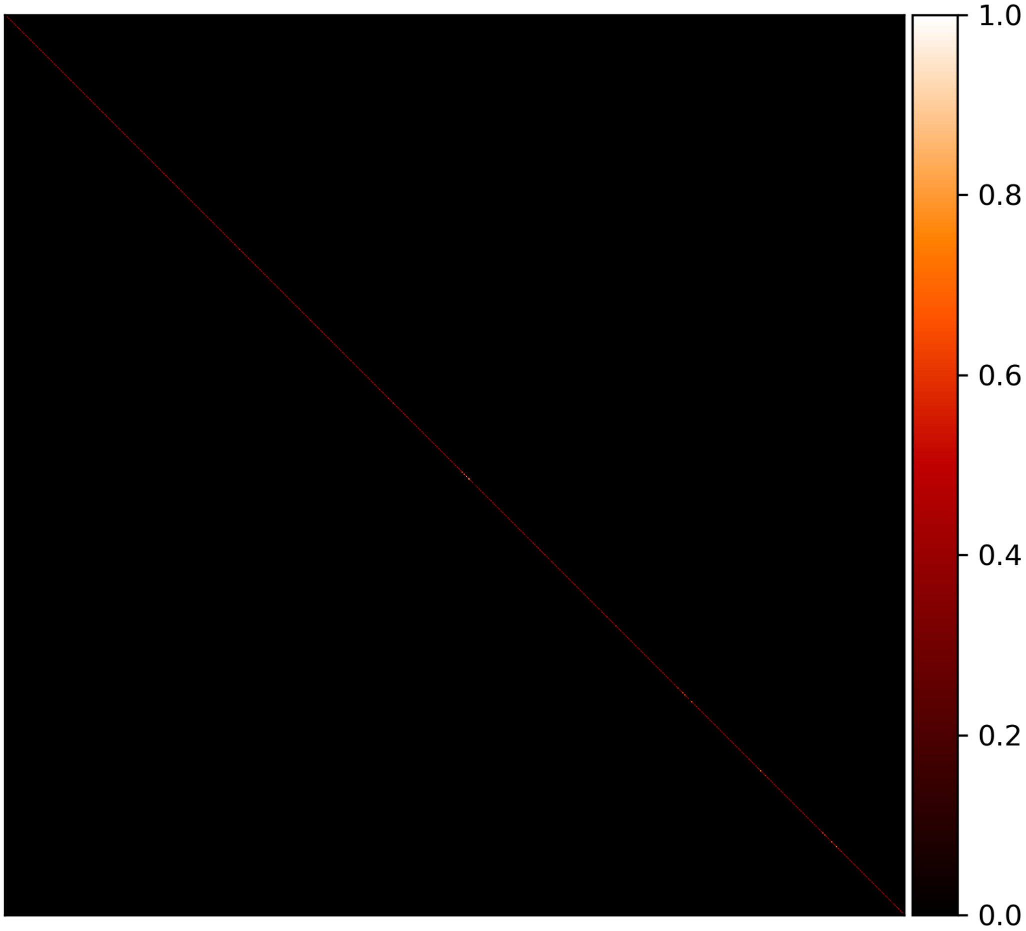


Figure 6. The full heatmap of attribution gradients of Fashion-MNIST in size 784 × 784.

provide results of $\varepsilon = 0.1$ and $\varepsilon = 0.2$ in ℓ_2 case, and $\varepsilon = 0.01$ and $\varepsilon = 0.03$ in ℓ_∞ case. For CIFAR-10, we provide $\varepsilon = 0.2$ and $\varepsilon = 0.3$ for ℓ_2 case, and $\varepsilon = 4/255$ and $\varepsilon = 8/255$ in ℓ_∞ case. The results are presented in Tabs. 7 and 8. For ℓ_2 constrained case, we also provide the modified upper bound T'_e as in Sec. 5 since the Taylor approximations are inaccurate occasionally $(0 \sim 6\%)$. For all tested ε, it is noticed that the theoretical bounds bound the sample Euclidean and cosine distance above. In some cases, the means of T_e and T'_e are the same because T_e bound $\widehat{T}_e$ well and the c in Eq. (11) equals to 1 for T'_e. As in Sec. 5, for ℓ_∞ case, we do not present the results of T'_e, because T_e has bounded all $\widehat{T}_e$ above.

D.3. Evaluation of upper bounds under ℓ_2-norm and ℓ_∞-norm constraints on larger size images.

The proposed method is also scalable to larger size images. In this subsection, we provide experimental results on Flower [20], which contains images of size of $128 \times 128 \times 3$, and a subset of ImageNet [5] containing 5,000 randomly chosen images with size of $224 \times 224 \times 3$. We choose $\varepsilon = 0.1$ for ℓ_2 and $\varepsilon = 8/255$ for ℓ_∞ cases to compute the theoretical upper bounds T_e and T_c, as well as the modified bound T'_e, as introduced in Sec. 5. The sample distance $\widehat{T}_e$ and $\widehat{T}_c$ are computed from the mean of distances between perturbed and original attributions, where PGD-20 is used as ℓ_2 attack and 200-step IFIA is used as ℓ_∞ attack. In paricular, since the baseline attribution robustness methods do not scale up to ImageNet, we only provide results using standard training and adversarial training to illustrate the scalability of our method. The results are presented in Tabs. 9 and 10.

We notice that the theoretical bounds are all valid for larger size images, where all angular and modified Euclidean bound effective bound the maximum discrepancy of perturbed attributions. It worths noting that the computation costs of the values for the upper bound in ℓ_2-norm constrained case become heavier for high-dimensional images due to the computation of eigenvalues for large matrices. For ℓ_∞-norm case, these eigenvalue computations have been avoided. We will study the

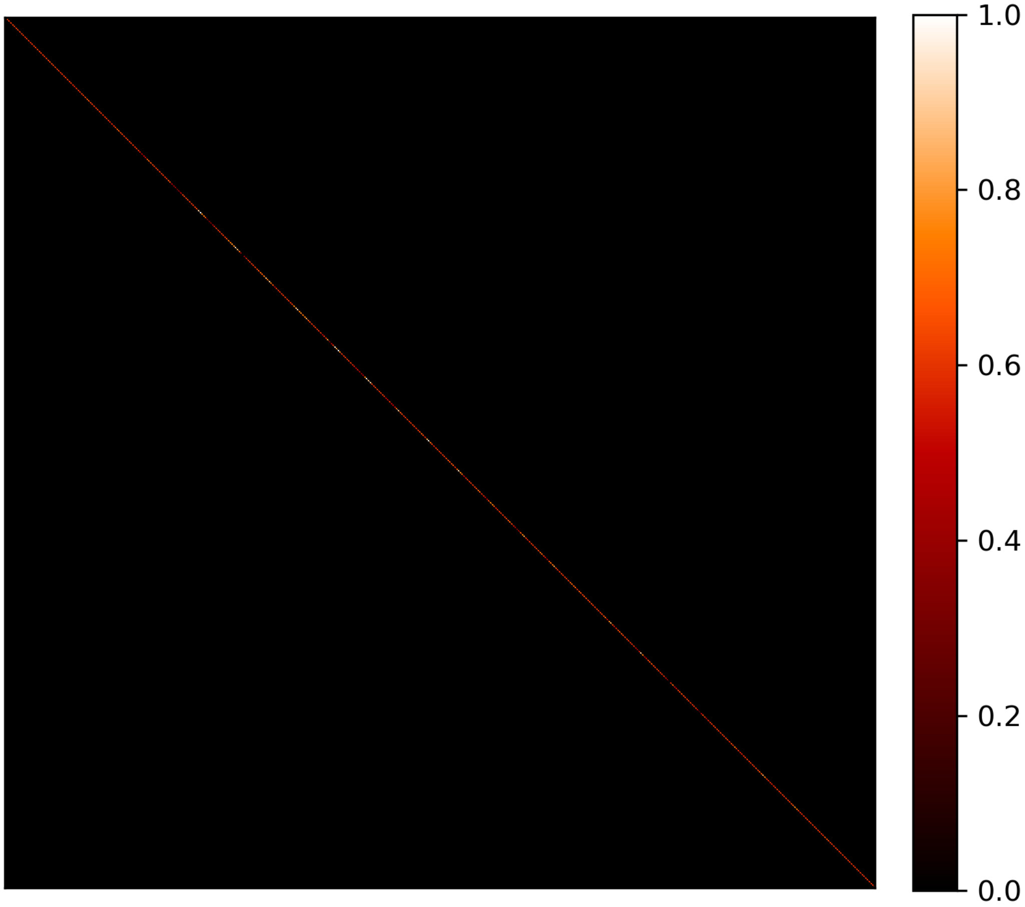


Figure 7. The full heatmap of attribution gradients of CIFAR-100 in size 3072 × 3072.

Table 7. Evaluation of ℓ_2-norm upper bound with the label constraint on MNIST, Fashion-MNIST and CIFAR-10 using different ε.

	$\widehat{T}_e$	T_e	T'_e	$\widehat{T}_c$(deg)	T_c(deg)	$\widehat{T}_e$	T_e	T'_e	$\widehat{T}_c$(deg)	T_c(deg)
MNIST			$\varepsilon = 0.1$					$\varepsilon = 0.2$		
AT	0.0856	0.3074	0.3101	4.6026	14.3020	0.1176	0.4611	0.4617	5.9845	29.6082
IG-NORM	0.1436	0.5776	0.5776	3.9514	14.6430	0.2094	0.8664	0.8679	5.4824	30.3707
AdvAAT	0.0938	0.7182	0.7193	2.1315	13.8325	0.1346	1.0773	1.1013	2.8725	28.5660
ART	0.2031	0.6538	0.6542	6.4244	13.9011	0.2302	0.9807	0.9993	8.5982	28.7175
TRADES	0.2159	1.0120	1.0812	3.4791	14.1049	0.3281	1.5180	1.5211	4.9429	29.1695
TRADES+IGR	0.2171	0.9928	1.0101	3.4171	14.0621	0.3032	1.4892	1.4892	4.5166	29.0745
Fashion-MNIST			$\varepsilon = 0.1$					$\varepsilon = 0.2$		
AT	0.1080	0.1400	0.1401	16.7770	26.6451	0.1413	0.2100	0.2119	21.3901	63.7570
IG-NORM	0.1232	0.3578	0.3578	8.9312	17.8256	0.1771	0.5367	0.5371	12.5177	37.7516
AdvAAT	0.1500	0.3470	0.3533	7.3499	19.0014	0.1984	0.5205	0.5209	9.4643	40.6308
ART	0.2057	0.2774	0.2775	11.6920	19.9515	0.2343	0.4161	0.4161	13.4216	43.0352
TRADES	0.0797	0.1926	0.1987	10.5544	24.7845	0.1050	0.2889	0.2889	13.8358	56.9729
TRADES+IGR	0.0672	0.0906	0.0906	11.3338	17.9020	0.0879	0.1359	0.1510	14.7998	37.9358
CIFAR-10			$\varepsilon = 0.2$					$\varepsilon = 0.3$		
AT	0.0607	0.5064	0.5064	3.7975	9.5783	0.0858	1.2661	1.2661	5.2981	24.5816
IG-NORM	0.0123	0.3164	0.3164	1.4311	8.7679	0.0592	0.7910	0.7910	6.9460	22.4006
AdvAAT	0.0300	0.4772	0.4775	1.7094	7.6575	0.0548	1.1933	1.1933	3.0553	19.4588
ART	0.0501	0.4556	0.4699	3.1004	8.4476	0.0718	1.1391	1.1420	6.3493	21.5468
TRADES	0.0360	0.3468	0.3468	3.9435	7.7550	0.0528	0.8671	0.8780	5.7514	19.7151
TRADES+IGR	0.0395	0.3384	0.3385	4.1222	7.6942	0.0577	0.8460	0.8460	5.9201	19.5551

Table 8. Evaluation of upper bounds under ℓ_∞-norm constraint and label constraint on MNIST, Fashion-MNIST and CIFAR-10 with different ε.

	$\widehat{T}_e$	T_e	$\widehat{T}_c$(deg)	T_c(deg)	$\widehat{T}_e$	T_e	$\widehat{T}_c$(deg)	T_c(deg)
MNIST	$\varepsilon = 0.01$				$\varepsilon = 0.03$			
AT	0.0556	0.1550	2.9408	7.1839	0.0888	0.4651	4.2516	22.0345
IG-NORM	0.1005	0.2409	2.8745	6.0632	0.1710	0.7228	4.4179	18.4742
AdvAAT	0.0608	0.4398	1.4264	5.0839	0.1280	1.3195	2.4883	15.4170
ART	0.0767	0.5644	2.8025	10.3833	0.3617	1.6931	9.3505	32.7312
TRADES	0.1634	0.4443	2.7539	6.3323	0.3193	1.3330	4.7523	19.3224
TRADES+IGR	0.1744	0.4077	2.7731	5.1333	0.2932	1.2232	4.2425	15.5702
Fashion-MNIST	$\varepsilon = 0.01$				$\varepsilon = 0.03$			
AT	0.0516	0.0560	6.5146	9.4467	0.1043	0.1680	16.4165	29.4979
IG-NORM	0.0611	0.1113	4.7737	8.3315	0.1137	0.3339	8.1315	25.7661
AdvAAT	0.0987	0.1841	5.3706	8.1184	0.1616	0.5523	7.9204	25.0658
ART	0.0660	0.1443	6.6582	9.2791	0.3946	0.4329	23.0589	28.9294
TRADES	0.0509	0.0907	7.0612	9.0233	0.0804	0.2721	10.8579	28.0672
TRADES+IGR	0.0363	0.0505	7.1214	8.0541	0.0716	0.1515	12.1090	24.8550
CIFAR-10	$\varepsilon = 4/255$				$\varepsilon = 8/255$			
AT	0.0894	0.1200	6.0843	6.4041	0.1549	0.2400	10.5129	12.8901
IG-NORM	0.0388	0.0750	4.5743	5.2004	0.0700	0.1501	8.1882	10.4443
AdvAAT	0.0776	0.0817	2.2657	5.7139	0.0959	0.1635	3.8595	11.4857
ART	0.0722	0.1056	4.3010	5.2445	0.1281	0.2113	8.4555	10.5337
TRADES	0.0539	0.0784	3.6093	5.3381	0.0909	0.1569	9.3571	10.7232
TRADES+IGR	0.0589	0.0821	3.8230	5.1622	0.0978	0.1643	9.5879	10.3668

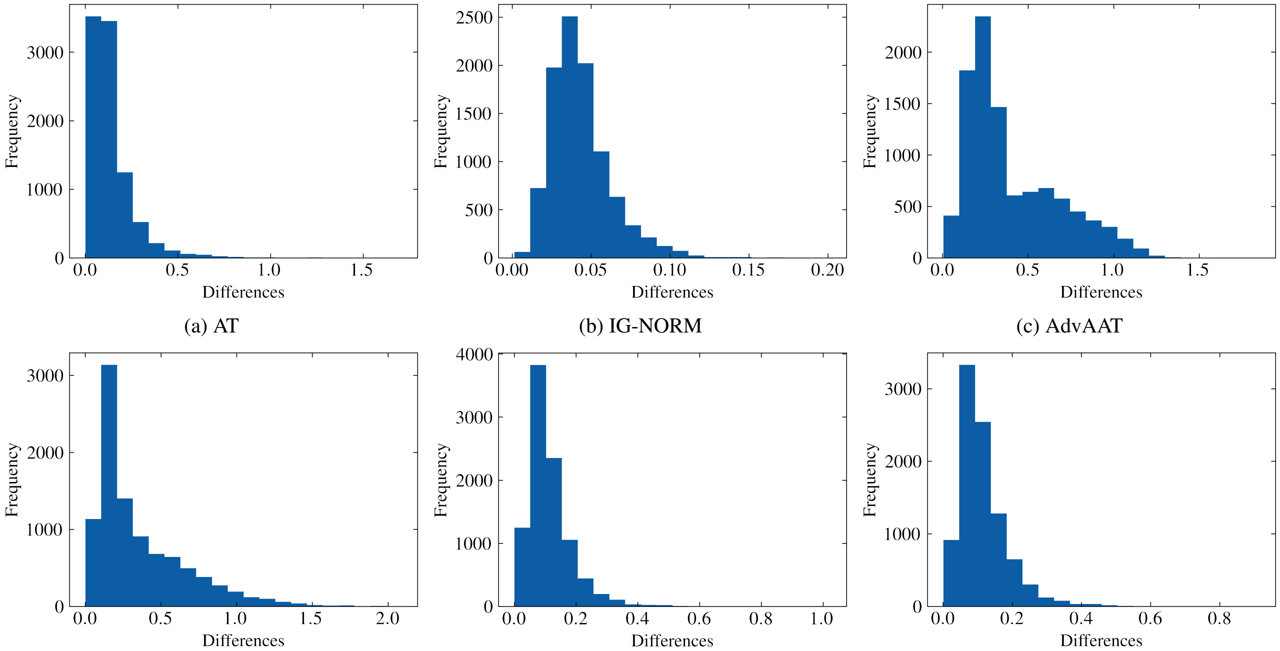
Table 9. Evaluation of upper bounds with the label constraint on Flower dataset. The numbers in the brackets indicate the percentages that attacked attribution is outside the T_e.

	ℓ_2					ℓ_∞			
	$\widehat{T}_e$	T_e	T'_e	$\widehat{T}_c$(deg)	T_c(deg)	$\widehat{T}_e$	T_e	$\widehat{T}_c$(deg)	T_c(deg)
AT	0.0170	0.0341 [2.17%]	0.0447	1.3165	1.9806	0.0238	0.4100	2.1937	13.4811
AdvAAT	0.0295	0.1424 [0.00%]	0.1424	1.5568	2.2835	0.0472	0.1025	1.4130	11.8732
TRADES	0.0220	0.0534 [0.72%]	0.0592	1.3383	3.1567	0.0182	0.1081	3.3887	11.9829
TRADES+IGR	0.0080	0.0219 [0.72%]	0.0262	0.8870	2.1255	0.0242	0.2873	1.5930	12.5584

Table 10. Evaluation of upper bounds with the label constraint on ImageNet. The numbers in the brackets indicate the percentages that attacked attribution is outside the T_e.

	ℓ_2					ℓ_∞			
	$\widehat{T}_e$	T_e	T'_e	$\widehat{T}_c$(deg)	T_c(deg)	$\widehat{T}_e$	T_e	$\widehat{T}_c$(deg)	T_c(deg)
CE	0.1049	0.1923[3.06%]	0.2365	7.7959	8.0933	0.3148	0.7399	3.8227	10.9339
AT	0.1077	0.1588[3.32%]	0.5221	3.4773	5.1797	0.1974	0.2226	0.3455	8.7333

scalability of our methods under ℓ_2-norm constraint in future work.



(a) AT

(b) IG-NORM

(c) AdvAAT

(d) ART

(e) TRADES

(f) TRADES+IGR

Figure 8. Distributions of differences between computed bounds and attribution differences from CIFAR-10.

E. Alternative formulation of upper bound to the worst-case attribution deviations

The formulation of Eq. 1 can be rewritten in an equivalent form to find the maximum ε subject to the attribution difference under certain threshold ω. Formally, the formulation can be written as

$$\begin{aligned} \max \quad & \varepsilon \\ \text{s.t.} \quad & D(g^y(\boldsymbol{x}), g^y(\boldsymbol{x}+\boldsymbol{\delta})) \leq \omega \\ & \|\boldsymbol{\delta}\|_p \leq \varepsilon \\ & \arg\max_k f_k(\boldsymbol{x}) = \arg\max_k f_k(\boldsymbol{x}+\boldsymbol{\delta}) \end{aligned} \tag{48}$$

Under the above formulation, we can use the theoretical bound derived using Eq. 1 to find the corresponding optimal ε. For the ℓ_2-norm case with or without the label constraint, when $D(\cdot,\cdot)$ is the ℓ_2 distance, the maximum ε can be computed using the upper bound $\xi_{max}\varepsilon$ derived in Theorem 1,

$$\max_{\boldsymbol{\delta}} \|g^y(\boldsymbol{x}+\boldsymbol{\delta}) - g^y(\boldsymbol{x})\|_2 = \xi_{max}\varepsilon \leq \omega \tag{49}$$

$$\Rightarrow \varepsilon \leq \frac{\omega}{\xi_{max}} \tag{50}$$

Similarly, the maximum ε when $D(\cdot,\cdot)$ is cosine distance can be derived using Corollary 2 as

$$\max_{\boldsymbol{\delta}} D_c(g^y(\boldsymbol{x}+\boldsymbol{\delta}), g^y(\boldsymbol{x})) = 1 - \sqrt{1 - \frac{\xi_{max}\varepsilon}{\|g^y(\boldsymbol{x})\|_2^2}} \leq \omega \tag{51}$$

$$\Rightarrow \varepsilon \leq \frac{\|g(\boldsymbol{x})\|_2^2}{\xi_{max}} \left(1 - (1-\omega)^2\right) \tag{52}$$

The maximum ε for the ℓ_∞ constraint case with and without the label constraint can be also derived in the same way using the relaxed upper bound in Theorem 2. Since the Kendall's rank correlation is discontinuous, researchers proposed to use cosine similarity and ℓ_p distance to measure the similarity/dissimilarity between attributions from attacked samples and original samples [3, 4, 32]. Thus, in this work, we derive the bounds for cosine similarity and Euclidean distance.