

A PROOFS

Lemma 1. (Paulavičius & Žilinskas (2006)) For L -Lipschitz function $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\|_q \quad (11)$$

where $L = \max \{\|\nabla f(\mathbf{x})\|_p : \mathbf{x} \in S\}$ is Lipschitz constant. Thus, $\|\nabla f(\mathbf{x})\|_p \leq L$.

Proof. Refer to Paulavičius & Žilinskas (2006) for the proof. \square

Theorem 1. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a upper bounded attribution function, and $\boldsymbol{\eta} \stackrel{U}{\sim} \mathcal{B}(\mathbf{0}; r)$. Let h be the smoothed version of g as defined in (2). Then, for all $\tilde{\mathbf{x}} \in \{\mathbf{x} + \boldsymbol{\delta} \mid \|\boldsymbol{\delta}\|_2 \leq \epsilon\}$, we have $\cos(h(\mathbf{x}), h(\tilde{\mathbf{x}})) \geq T$, where

$$T = \frac{\|h(\mathbf{x})\|_2}{\sqrt{\|h(\mathbf{x})\|_2^2 + M^2 V_U^2 / V_S^2}} \quad (6)$$

Here, M is the upper bound of g . V_S is the volume of the ℓ_2 -ball $\mathcal{B}(\mathbf{0}; r)$, and V_U is the volume of the union of the two sampling space centered at \mathbf{x} and $\tilde{\mathbf{x}}$ minus their intersection.

Proof. As defined in Eqn. (2)

$$h(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\eta} \sim \mathcal{B}(\mathbf{0}; r)} [g(\mathbf{x} + \boldsymbol{\eta})] = \frac{1}{V_S} \int_{\boldsymbol{\eta} \sim \mathcal{B}(\mathbf{0}; r)} g(\mathbf{x} + \boldsymbol{\eta}) d\boldsymbol{\eta} \quad (12)$$

where V_S is the volume of the ℓ_p -ball with radius r . Similarly, let $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \in \mathbb{R}^d$ is a vector and $\|\boldsymbol{\delta}\|_2 \leq \epsilon$. Then, we have

$$h(\tilde{\mathbf{x}}) = \frac{1}{V_S} \int_{\boldsymbol{\eta} \sim \mathcal{B}(\mathbf{0}; r)} g(\tilde{\mathbf{x}} + \boldsymbol{\eta}) d\boldsymbol{\eta} \quad (13)$$

We note that when $\boldsymbol{\eta} \sim \mathcal{B}(\mathbf{0}; r)$, $\mathbf{x} + \boldsymbol{\eta} \sim \mathcal{B}(\mathbf{x}; r)$ and $\tilde{\mathbf{x}} + \boldsymbol{\eta} \sim \mathcal{B}(\tilde{\mathbf{x}}; r)$. We then rewrite $h(\mathbf{x})$ and $h(\tilde{\mathbf{x}})$ as follows:

$$h(\mathbf{x}) = \underbrace{\frac{1}{V_S} \int_{\mathbf{x} \sim \mathcal{B}(\mathbf{x}; r) \setminus \mathcal{B}(\tilde{\mathbf{x}}; r)} g(\mathbf{x}) d\mathbf{x}}_{R_1} + \underbrace{\frac{1}{V_S} \int_{\mathbf{x} \sim \mathcal{B}(\tilde{\mathbf{x}}; r) \cap \mathcal{B}(\mathbf{x}; r)} g(\mathbf{x}) d\mathbf{x}}_{R_2} \quad (14)$$

and

$$h(\tilde{\mathbf{x}}) = \underbrace{\frac{1}{V_S} \int_{\mathbf{x} \sim \mathcal{B}(\tilde{\mathbf{x}}; r) \cap \mathcal{B}(\mathbf{x}; r)} g(\mathbf{x}) d\mathbf{x}}_{R_2} + \underbrace{\frac{1}{V_S} \int_{\mathbf{x} \sim \mathcal{B}(\tilde{\mathbf{x}}; r) \setminus \mathcal{B}(\mathbf{x}; r)} g(\mathbf{x}) d\mathbf{x}}_{R_3} \quad (15)$$

Hence,

$$h(\tilde{\mathbf{x}}) = h(\mathbf{x}) - R_1 + R_3 \quad (16)$$

Denote $av = R_3 - R_1$, where v is a unit vector in the same direction of $R_3 - R_1$ and $a = \|R_3 - R_1\|_2$ is a scalar with the same magnitude of $R_3 - R_1$. Then, we have

$$\cos(h(\mathbf{x}), h(\tilde{\mathbf{x}})) = \frac{h(\mathbf{x})^\top}{\|h(\mathbf{x})\|_2} \left(\frac{h(\mathbf{x}) + av}{\|h(\mathbf{x}) + av\|_2} \right) \quad (17)$$

Note that the attribution $g(\mathbf{x})$ is upper bounded by M , specifically, $\|g(\mathbf{x})\|_2 \leq M$, for some constant M . Thus, we can derive that

$$a = \|R_3 - R_1\|_2 \quad (18)$$

$$= \left\| \frac{1}{V_S} \left(\int_{\mathbf{x} \sim \mathcal{B}(\tilde{\mathbf{x}}; r) \setminus \mathcal{B}(\mathbf{x}; r)} g(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{x} \sim \mathcal{B}(\mathbf{x}; r) \setminus \mathcal{B}(\tilde{\mathbf{x}}; r)} g(\mathbf{x}) d\mathbf{x} \right) \right\|_2 \quad (19)$$

$$\leq \frac{1}{V_S} \left(\left\| \int_{\mathbf{x} \sim \mathcal{B}(\tilde{\mathbf{x}}; r) \setminus \mathcal{B}(\mathbf{x}; r)} g(\mathbf{x}) d\mathbf{x} \right\|_2 + \left\| \int_{\mathbf{x} \sim \mathcal{B}(\mathbf{x}; r) \setminus \mathcal{B}(\tilde{\mathbf{x}}; r)} g(\mathbf{x}) d\mathbf{x} \right\|_2 \right) \quad (20)$$

$$\leq \frac{1}{V_S} \left(\int_{\mathbf{x} \sim \mathcal{B}(\tilde{\mathbf{x}}; r) \setminus \mathcal{B}(\mathbf{x}; r)} \|g(\mathbf{x})\|_2 d\mathbf{x} + \int_{\mathbf{x} \sim \mathcal{B}(\mathbf{x}; r) \setminus \mathcal{B}(\tilde{\mathbf{x}}; r)} \|g(\mathbf{x})\|_2 d\mathbf{x} \right) \quad (21)$$

$$\leq \frac{1}{V_S} \left(\int_{\mathbf{x} \sim \mathcal{B}(\tilde{\mathbf{x}}; r) \setminus \mathcal{B}(\mathbf{x}; r)} M d\mathbf{x} + \int_{\mathbf{x} \sim \mathcal{B}(\mathbf{x}; r) \setminus \mathcal{B}(\tilde{\mathbf{x}}; r)} M d\mathbf{x} \right) \quad (22)$$

$$= M \times \frac{V_{\mathcal{B}(\mathbf{x}; r) \setminus \mathcal{B}(\tilde{\mathbf{x}}; r) \cup \mathcal{B}(\tilde{\mathbf{x}}; r) \setminus \mathcal{B}(\mathbf{x}; r)}}{V_S} = M \frac{V_U}{V_S} \quad (23)$$

Thus, the lower bound of $\cos(h(\mathbf{x}), h(\tilde{\mathbf{x}}))$ can be found by solving the optimization problem²

$$\begin{aligned} \min_{\mathbf{v}} \quad & \frac{h(\mathbf{x})^\top}{\|h(\mathbf{x})\|} \left(\frac{h(\mathbf{x}) + a\mathbf{v}}{\|h(\mathbf{x}) + a\mathbf{v}\|} \right) \\ \text{s.t.} \quad & \|\mathbf{v}\| = 1 \\ & a \leq M \frac{V_U}{V_S} \end{aligned} \quad (24)$$

Since $h(\mathbf{x})$ and $h(\tilde{\mathbf{x}})$ form a spherical cone, we can decompose \mathbf{v} by $\mathbf{v} = \cos \theta \mathbf{v}_\parallel + \sin \theta \mathbf{v}_\perp$, where \mathbf{v}_\parallel and \mathbf{v}_\perp are two orthogonal unit vectors such that $h^\top(\mathbf{x})\mathbf{v}_\perp = 0$ and $\mathbf{v}_\parallel = h(\mathbf{x})/\|h(\mathbf{x})\|$. Then, the optimization problem can be rewritten as

$$\min \quad \mathbf{v}_\parallel^\top \left(\frac{h(\mathbf{x}) + a(\cos \theta \mathbf{v}_\parallel + \sin \theta \mathbf{v}_\perp)}{\|h(\mathbf{x}) + a(\cos \theta \mathbf{v}_\parallel + \sin \theta \mathbf{v}_\perp)\|} \right) \quad (25)$$

$$\Rightarrow \min \quad \mathbf{v}_\parallel^\top \left(\frac{\|h(\mathbf{x})\| \mathbf{v}_\parallel + a(\cos \theta \mathbf{v}_\parallel + \sin \theta \mathbf{v}_\perp)}{\| \|h(\mathbf{x})\| \mathbf{v}_\parallel + a(\cos \theta \mathbf{v}_\parallel + \sin \theta \mathbf{v}_\perp) \|} \right) \quad (26)$$

$$\Rightarrow \min \quad \frac{(\|h(\mathbf{x})\| + a \cos \theta) \mathbf{v}_\parallel^\top \mathbf{v}_\parallel + a \sin \theta \mathbf{v}_\parallel^\top \mathbf{v}_\perp}{\sqrt{(\|h(\mathbf{x})\| + a \cos \theta)^2 \mathbf{v}_\parallel^\top \mathbf{v}_\parallel + (a \sin \theta)^2 \mathbf{v}_\perp^\top \mathbf{v}_\perp}} \quad (27)$$

$$\Rightarrow \min \quad \frac{\|h(\mathbf{x})\| + a \cos \theta}{\sqrt{(\|h(\mathbf{x})\| + a \cos \theta)^2 + (a \sin \theta)^2}} \quad (28)$$

Since $h(\mathbf{x})$ is known for a given sample, the optimization problem can be written as follows by taking $\|h(\mathbf{x})\| = c$:

$$\begin{aligned} \min \quad & \frac{c + a \cos \theta}{\sqrt{(c + a \cos \theta)^2 + (a \sin \theta)^2}} \\ \text{s.t.} \quad & a \leq M \frac{V_U}{V_S} \end{aligned} \quad (29)$$

We now consider the Lagrange function of the optimization problem:

$$\mathcal{L}(x, \theta, \lambda) = \frac{c + a \cos \theta}{\sqrt{(c + a \cos \theta)^2 + (a \sin \theta)^2}} - \lambda(a - M \frac{V_U}{V_S}) \quad (30)$$

Taking the derivative of \mathcal{L} with respect to a and θ and setting them to zero, we have

$$\frac{\partial}{\partial a} \mathcal{L} = \frac{1}{T^2} \left(T \cos \theta - \frac{1}{T} (c \cos \theta + 2a) \times (c + a \cos \theta) \right) - \lambda = 0 \quad (31)$$

² $\|\cdot\|$ in the following content denotes the ℓ_2 -norm unless otherwise specified.

and

$$\frac{\partial}{\partial \theta} \mathcal{L} = \frac{1}{T^2} \left(-a \sin \theta \cdot T + \frac{1}{T} (c^2 a \sin \theta + ca^2 \sin \theta \cos \theta) \right) = 0 \quad (32)$$

where $T = \sqrt{(c + a \cos \theta)^2 + (a \sin \theta)^2}$. Solving the above equations, we have

$$\cos \theta = 0 \quad \text{or} \quad a = 0 \quad (33)$$

where $a = 0$ reaches the maximum and $\cos \theta = 0$ is the minimum. Therefore, the lower bound of $\cos(h(\mathbf{x}), h(\tilde{\mathbf{x}}))$ is

$$\cos(h(\mathbf{x}), h(\tilde{\mathbf{x}})) \geq \frac{c}{\sqrt{c^2 + (M \frac{V_U}{V_S})^2}} = \frac{\|h(\mathbf{x})\|}{\sqrt{\|h(\mathbf{x})\|^2 + (MV_U/V_S)^2}} \quad (34)$$

□

Corollary 1. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a bounded attribution function, and $\eta \stackrel{U}{\sim} \mathcal{B}(\mathbf{x}; r)$. Let h be the smoothed version of g as defined in (2).

- (i) Given a predefined threshold $T \in [0, 1]$, then for all $\|\delta\|_2 \leq \epsilon$, we have $\cos(h(\mathbf{x}), h(\mathbf{x} + \delta)) \geq T$, where

$$\epsilon = 2r \sqrt{1 - I_Z^{-1} \left(\frac{d+1}{2}, \frac{1}{2} \right)}. \quad (8)$$

- (ii) Given a predefined threshold $T \in [0, 1]$ and the maximum perturbation size $\epsilon \geq 0$, the smoothed attribution satisfies $\cos(h(\mathbf{x}), h(\tilde{\mathbf{x}})) \geq T$ for all $\tilde{\mathbf{x}} \in \{\mathbf{x} + \delta \mid \|\delta\|_2 \leq \epsilon\}$ when $r \geq R$, where

$$R = \frac{\epsilon}{2} \left(1 - I_Z^{-1} \left(\frac{d+1}{2}, \frac{1}{2} \right) \right)^{-\frac{1}{2}}. \quad (9)$$

$I_z^{-1}(a, b)$ is the inverse of the regularized incomplete beta function, and Z is defined as

$$Z = 1 - \frac{\|h(\mathbf{x})\|_2}{2M} \left(\frac{1}{T^2} - 1 \right) \quad (10)$$

Proof. Corollary 1 can be obtained by fixing T and taking r as unknown, and fixing T and taking ϵ as unknown, respectively. We can first derive that

$$I_{(2rh-h^2)/r^2} \left(\frac{d+1}{2}, \frac{1}{2} \right) = 1 - \frac{\|h(\mathbf{x})\|_2}{2M} \sqrt{\frac{1}{T^2} - 1} = Z \quad (35)$$

Using the inverse of the regularized incomplete beta function, i.e., $x = I_y^{-1}(a, b)$, and $h = r - \epsilon/2$, we have

$$I_Z^{-1} \left(\frac{d+1}{2}, \frac{1}{2} \right) = (2rh - h^2)/r^2 = 1 - \frac{\epsilon^2}{4r^2} \quad (36)$$

The results in Corollary can then be solved accordingly. □

B IMPLEMENTATION DETAILS

In the experiments, we implemented the ℓ_2 attribution attack adapted from [Ghorbani et al. (2019)]. The attack uses top- k intersection version as the loss function. Following previous works, we choose $k = 100$ for MNIST and $k = 1000$ for CIFAR-10. The number of iterations in PGD-like attack is 200, and the step size is 0.1. As mentioned in the main content, we do not implement the attack on ImageNet since the attribution attacks are not scalable to large size images. In the following parts of this section, we provide more details of evaluations in the experiments.

B.1 ATTRIBUTION METHODS

We used saliency maps (SM) and integrated gradients (IG) in the evaluation sections. These two methods are defined as follows:

- Saliency maps: $\text{SM}(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$.
- Integrated gradients: $\text{IG}(\mathbf{x}) = (\mathbf{x} - \mathbf{x}') \times \int_{\alpha=0}^1 \frac{\partial f(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial \mathbf{x}} d\alpha$.

The SmoothSM and SmoothIG are the smoothed versions of SM and IG, respectively.

B.2 EVALUATION METRICS

Given original attribution $g(\mathbf{x})$ and perturbed attribution $g(\tilde{\mathbf{x}})$, we use top-k intersection, Kendall's rank correlation (Ghorbani et al., 2019) and cosine similarity (Wang & Kong, 2022) to evaluate their differences.

- Top-k intersection measures the proportion of k largest features that overlap between $g(\mathbf{x})$ and $g(\tilde{\mathbf{x}})$.
- Kendall's rank correlation measures the proportion of pairs of features that have the same order in $g(\mathbf{x})$ and $g(\tilde{\mathbf{x}})$: $\frac{2}{d(d-1)} \sum_{i=1}^d \sum_{j=i+1}^d \mathbf{1}_{\{g(\mathbf{x})_i > g(\mathbf{x})_j\}} \mathbf{1}_{\{g(\tilde{\mathbf{x}})_i > g(\tilde{\mathbf{x}})_j\}}$.
- Cosine similarity measures the cosine of the angle between $g(\mathbf{x})$ and $g(\tilde{\mathbf{x}})$: $\frac{g(\mathbf{x})^\top g(\tilde{\mathbf{x}})}{\|g(\mathbf{x})\| \|g(\tilde{\mathbf{x}})\|}$.

B.3 BASELINE METHODS

We compare with the following adversarial and attributional robust models:

IG-NORM (Chen et al., 2019)

$$\text{CE}(f(\mathbf{x}), y) + \lambda \max_{\tilde{\mathbf{x}} \in \mathcal{B}_\varepsilon(\mathbf{x})} \|\text{IG}(\mathbf{x}, \tilde{\mathbf{x}})\|_1 \quad (37)$$

TRADES (Zhang et al., 2019)

$$\text{CE}(f(\tilde{\mathbf{x}}), y) + \beta \text{KL}(f(\mathbf{x}) \| f(\tilde{\mathbf{x}})) \quad (38)$$

IGR (Wang & Kong, 2022)

$$\text{CE}(f(\tilde{\mathbf{x}}), y) + \beta \text{KL}(f(\mathbf{x}) \| f(\tilde{\mathbf{x}})) + \lambda (1 - \cos(\text{IG}(\mathbf{x}), \text{IG}(\tilde{\mathbf{x}}))) \quad (39)$$

Here CE denotes the cross-entropy loss and KL denotes the Kullback-Leibler divergence.

C ADDITIONAL EXPERIMENTS

C.1 TEST ON MONTE CARLO ESTIMATION

Note that the bound given by Theorem 1 is deterministic. In this section, we provide a probabilistic bound for the attribution robustness. Specifically, we want to find the value of t such that $\Pr(T \leq t) = 1 - \alpha$, where T is defined in Eqn. (6) and α is the significance level. Recall that T is defined as follows:

$$T = \frac{\|h(\mathbf{x})\|_2}{\sqrt{\|h(\mathbf{x})\|_2^2 + c}} \quad (40)$$

where $c = M^2 V_U^2 / V_S^2$. If we denote that $Q = \|h(\mathbf{x})\|_2$, then we have

$$\Pr(T \leq t) = \Pr\left(\frac{Q}{\sqrt{Q^2 + c}} \leq t\right) = \Pr\left(Q^2 \leq \frac{ct^2}{1 - t^2}\right) \quad (41)$$

Table 4: Evaluation of center smoothing on attributions

ϵ_1	0.1	0.2	0.3	0.4	0.5
SmoothSM	1.207	1.729	1.843	1.907	1.998

Note that we used Monte Carlo Integration to calculate the integral in $h(\mathbf{x})$, which estimates $h(\mathbf{x})$ by sampling $\boldsymbol{\eta}$ from \mathcal{B} , i.e.,

$$\hat{h}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N g(\mathbf{x} + \boldsymbol{\eta}_i), \quad \boldsymbol{\eta}_i \sim \mathcal{B}. \quad (42)$$

Note that $\hat{h}(\mathbf{x})$ is an unbiased estimator of $h(\mathbf{x})$, i.e. $\mathbb{E}[\hat{h}(\mathbf{x})] = h(\mathbf{x})$. The estimator almost surely converges to $h(\mathbf{x})$ as $N \rightarrow \infty$, i.e. $\lim_{N \rightarrow \infty} \hat{h}(\mathbf{x}) = h(\mathbf{x})$ almost surely. By the Central Limit Theorem, the estimator $\hat{h}(\mathbf{x})$ has the following asymptotic distribution,

$$\hat{h}(\mathbf{x}) \stackrel{a.s.}{\sim} \mathcal{N}(h(\mathbf{x}), D), \quad (43)$$

which the covariance matrix $D = \text{diag}(\sigma_{ii}^2/N)$ can be estimated by the empirical variances of $g(\mathbf{x} + \boldsymbol{\eta}_i)$. Thus, the quadratic form $Q^2 = \|\hat{h}(\mathbf{x})\|_2^2$ can be seen as generalized chi-square distributed. We can derive the cumulative distribution function of Monte Carlo estimator T_{MC} at t as the cumulative distribution function of the generalized chi-square distribution at $\frac{ct^2}{1-t^2}$, i.e.,

$$Pr(T_{MC} \leq t) = F\left(\frac{ct^2}{1-t^2}\right), \quad (44)$$

where F is the cumulative distribution function of the generalized chi-square distribution constructed from the quadratic form of Gaussian random variable with mean $h(\mathbf{x})$ and covariance D (Davies, 1980; Das & Geisler, 2021). In this work, we use the R package CompQuadForm (Duchesne & De Micheaux, 2010) to compute the cumulative distribution function. For any fixed image sample \mathbf{x} , we can validate $t_2 - t_1$ is close to 0 when $Pr(t_1 \leq T_{MC} \leq t_2) = 1 - \alpha$ by solving the following equation. For small $\alpha = 0.01$ and the number of samples $N = 100,000$, we found that the values of $t_2 - t_1$ are at scale of 10^{-4} in MNIST and CIFAR-10, and 10^{-3} in ImageNet calculated by choosing 10,000 samples from each dataset. This validates the error from Monte Carlo integral is minute and that the probabilistic bound is close to the deterministic bound.

$$F\left(\frac{ct_2^2}{1-t_2^2}\right) = 1 - \alpha/2 \quad \text{and} \quad F\left(\frac{ct_1^2}{1-t_1^2}\right) = \alpha/2. \quad (45)$$

C.2 ADDITIONAL VISUALIZATION OF THE UNIFORMLY SMOOTHED ATTRIBUTIONS

In Figure 1 (left), we have shown that the uniformly smoothed attributions have a comparable quality as the original attributions. Here more examples are provided in Figure 3 to illustrate the quality of the uniformly smoothed attributions.

C.3 EVALUATION OF CENTER SMOOTHING (KUMAR & GOLDSTEIN, 2021) ON ATTRIBUTIONS

To compare the performance with center smoothing (Kumar & Goldstein, 2021), we also implemented the same method to evaluate the certification of attributions. Specifically, we compute the bound for SmoothSM on IG-NORM using MNIST, and follow the same setting by choosing $h = 1$ and $\epsilon_1 = 0.1, 0.2, \dots, 0.5$. Directly using the cosine similarity on the method is not applicable since cosine similarity does not satisfy the triangle inequality. Following the relaxation method in Sec.4 of Kumar & Goldstein (2021), a multiplier $\gamma = 2$ is added. Besides, we use $1 - \cos \theta$ to reflect the distance metric instead of the similarity metric. The results are shown in the Table 4. It can be observed that the upper bound for $1 - \cos \theta$ is greater than 1 for all the choices of ϵ , which is trivially valid for the trigonometric function since we only consider $\cos \theta \in [0, 1]$. Thus, the upper bound provided in the aforementioned work can be too loose on our setting.

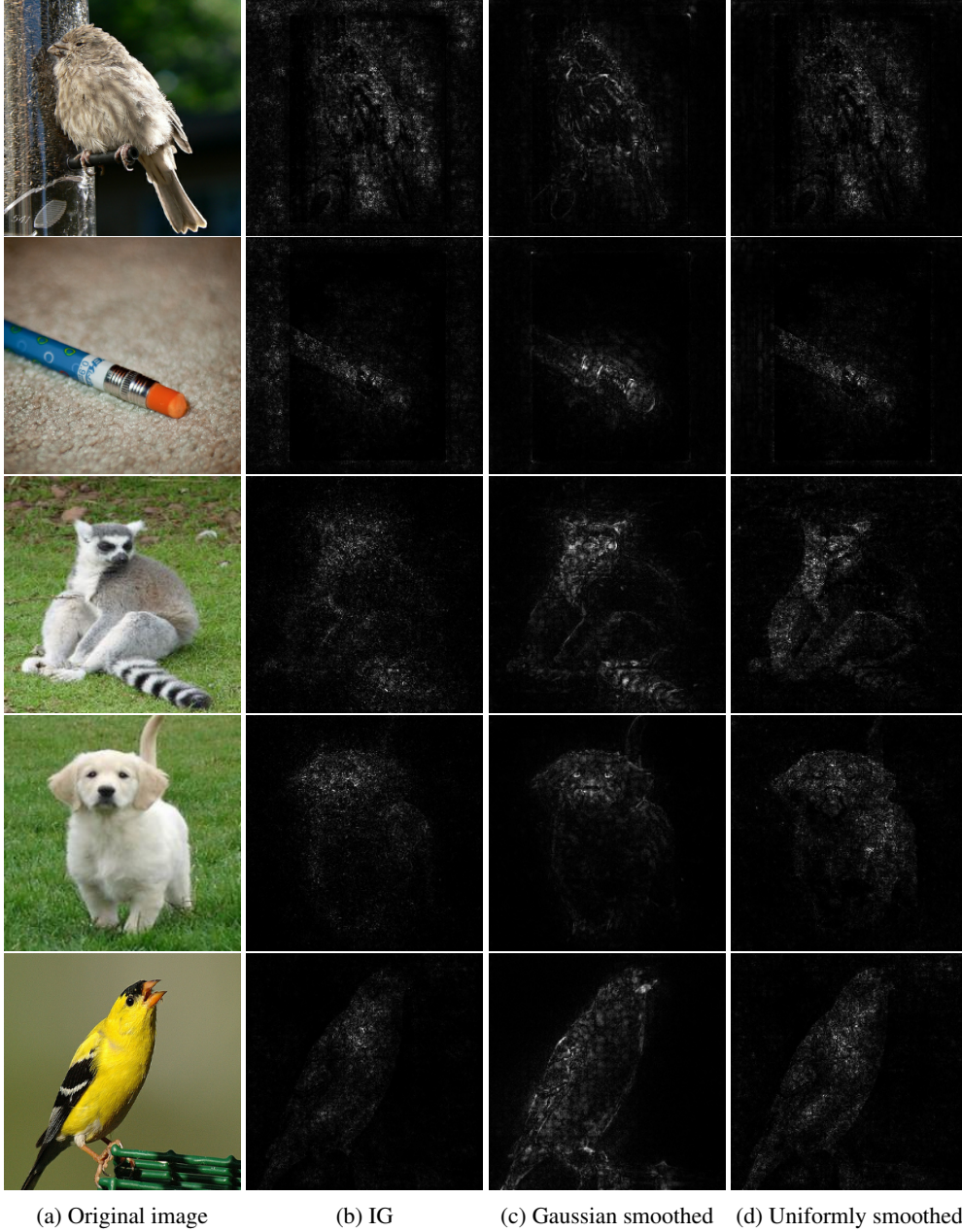


Figure 3: Additional visualization of the attribution maps of the (a) original image, (b) IG, (c) Gaussian smoothed IG, and (d) uniformly smoothed IG.

C.4 EVALUATION OF ALTERNATIVE FORMULATIONS

In Section 4.3, we introduced two alternative formulations of the proposed method that can be applied in specific scenarios. In this section, we provide additional information to report the experiments on these two formulations.

In Tables 5 to 7, which correspond to MNIST, CIFAR-10 and ImageNet, respectively, we report the computed values of the maximum allowable perturbation size. Under the size constraint, no examples can be found by the attacks against uniformly smoothed IG of a certain radius such that the cosine similarity between clean and perturbed attributions exceeds the given threshold ($T = 0.8$ and $T = 0.9$). The results are consistent with our theory. For larger radius smoothing, the maximum

Table 5: Maximum allowable perturbation size for different threshold ($T = 0.8$ and $T = 0.9$) under various choices of ℓ_2 smoothing radii r evaluated on MNIST.

$T = 0.9$	ℓ_2 radius (r)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
	Standard	0.0389	0.0951	0.1550	0.2164	0.2783	0.3404	0.4029
	IG-NORM	0.0394	0.0957	0.1557	0.2170	0.2790	0.3420	0.4067
	IGR	0.0390	0.0952	0.1552	0.2174	0.2818	0.3477	0.4163
$T = 0.8$	ℓ_2 radius (r)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
	Standard	0.0447	0.1051	0.1691	0.2345	0.3004	0.3664	0.4329
	IG-NORM	0.0448	0.1052	0.1692	0.2354	0.3037	0.3733	0.4456
	IGR	0.0452	0.1057	0.1697	0.2350	0.3010	0.3680	0.4365

Table 6: Maximum allowable perturbation size for different threshold ($T = 0.8$ and $T = 0.9$) under various choices of ℓ_2 smoothing radii r evaluated on CIFAR-10.

$T = 0.9$	ℓ_2 radius (r)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
	Standard	0.0086	0.0469	0.0885	0.1322	0.1773	0.2222	0.2683
	IG-NORM	0.0323	0.0705	0.1104	0.1510	0.1923	0.2337	0.2749
	IGR	0.0167	0.0545	0.1032	0.1586	0.2150	0.2588	0.2805
$T = 0.8$	ℓ_2 radius (r)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
	Standard	0.0128	0.0522	0.0951	0.1402	0.1866	0.2330	0.2868
	IG-NORM	0.0343	0.0742	0.1157	0.1580	0.2009	0.2439	0.2867
	IGR	0.0237	0.0693	0.1258	0.1861	0.2546	0.3090	0.3559

Table 7: Maximum allowable perturbation size for different threshold ($T = 0.8$ and $T = 0.9$) under various choices of ℓ_2 smoothing radii r evaluated on ImageNet.

ℓ_2 radius (r)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
$T = 0.9$	0.0046	0.0100	0.0152	0.0295	0.0494	0.0628	0.0768
$T = 0.8$	0.0058	0.0127	0.0196	0.0369	0.0618	0.0820	0.1040

Table 8: Empirical cosine similarity between original and perturbed smoothed attributions under various choices of ℓ_2 smoothing radius r , and the perturbation size computed in Table 5 ($T = 0.8$).

r	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Standard	0.8636	0.8522	0.8347	0.8127	0.8477	0.8603	0.8310
IG-NORM	0.8308	0.8181	0.8504	0.8728	0.8502	0.8193	0.8199
IGR	0.8231	0.8800	0.8720	0.8603	0.8362	0.8135	0.8567

Table 9: Minimum smoothing radius requires to achieve the threshold ($T = 0.8$ and $T = 0.9$) under various choices of ℓ_2 perturbation size ϵ . IG-NORM and IGR are omitted since they are not scalable to ImageNet.

		MNIST		CIFAR-10		ImageNet	
perturbation size (ϵ)		0.5	1.0	0.5	1.0	0.5	1.0
$T = 0.9$	Standard	5.1902	5.8752	5.9752	7.9504	74.6272	149.2544
	IG-NORM	5.1189	5.7699	5.6860	7.3720	/	/
	IGR	5.0265	5.6623	5.2895	6.5790	/	/
$T = 0.8$	Standard	3.8927	4.4064	5.7082	7.4164	48.2095	96.4190
	IG-NORM	3.8392	4.3274	5.4875	6.9750	/	/
	IGR	3.7699	4.2468	5.0287	6.0573	/	/

allowable perturbation size is also larger. When the threshold requirement is stricter, the maximum allowable perturbation size is smaller, which suggests weaker attacks are allowed. The method is also scalable to ImageNet, which takes around 15 seconds to compute for each sample. Moreover, we also applied attribution attacks using the same radius and maximum perturbation size ϵ , computed using Eqn. (8). Similar to the experiments in Section 5, we performed 20 attacks on each sample. We found that out of the total 200,000 attacked samples, the cosine similarities between clean and perturbed attributions were higher than the given threshold, suggesting that the computed bound is valid (see Table 8).

We also evaluate the third formulation that the minimum radius of smoothing required such that, within the given perturbation sizes, the cosine similarity between original and perturbed smoothed attributions is larger than the given threshold. In Table 9, the computed minimum radius of smoothing is reported. Similarly, we observe that the minimum radius of smoothing is larger when the threshold requirement is stricter, and when the attack is stronger. This is also consistent with our theory. We also notice that the radius for ImageNet is extremely large, which indicates that ImageNet is difficult to defend under such strict threshold requirements.