Tangentially Aligned Integrated Gradients for User-Friendly Explanations

Lachlan Simpson[1,*], Federico Costanza[2], Kyle Millar[3], Adriel Cheng[1,3], Cheng-Chew Lim[1] and Hong Gunn Chew[1]

[1]School of Electrical and Mechanical Engineering, The University of Adelaide, Australia
[2]School of Computer and Mathematical Sciences, The University of Adelaide, Australia
[3]Information Sciences Division, Defence Science and Technology Group, Australia

Abstract

Integrated gradients is prevalent within machine learning to address the black-box problem of neural networks. The explanations given by integrated gradients depend on a choice of base-point. The choice of base-point is not a priori obvious and can lead to drastically different explanations. There is a longstanding hypothesis that data lies on a low dimensional Riemannian manifold. The quality of explanations on a manifold can be measured by the extent to which an explanation for a point lies in its tangent space. In this work, we propose that the base-point should be chosen such that it maximises the tangential alignment of the explanation. We formalise the notion of tangential alignment and provide theoretical conditions under which a base-point choice will provide explanations lying in the tangent space. We demonstrate how to approximate the optimal base-point on several well-known image classification datasets. Furthermore, we compare the optimal base-point choice with common base-points and three gradient explainability models.

Keywords

Explainable AI, XAI, Integrated Gradients, Manifold Hypothesis.

1. Introduction

Deep learning provides state-of-the-art solutions to a wide array of computer vision tasks [1]. The accuracy of deep learning comes with the trade-off of interpretability [2]. A fundamental problem of deep learning is how a model reached a prediction [3]. Post hoc gradient explainability models address the black-box problem by providing an attribution of the input features to the prediction of neural network under analysis [4]. Several gradient explainability methods exist with the underlying assumption that analysis of the model's gradient highlights features with greatest impact on a prediction [5, 6].

Several metrics have been proposed to measure the quality of explainability models. In [4, 7], the authors propose the Lipschitz constant of an explainability model as a measure of explainability quality. Other works consider the extent to which an explainability model approximates the underlying neural network as a measure of quality. These metrics do not consider the user's perception of the explanations. Following from Ganz et al.'s [8] notion of perceptually aligned gradients of a neural network, Brodt et al. [9] introduce perceptually aligned explanations. Brodt el al. [9] measure how perceptually aligned an explanation is by the extent to which an explanation lies in the tangent space of the manifold. Brodt el al. [9]'s measure of tangential explanations relies on the manifold hypothesis. The manifold hypothesis is the notion that data lies on a low dimensional Riemannian manifold [10, 11, 12, 8, 13, 14].

The tangent space captures the features of an image that can be changed whilst remaining in the distribution of images. The intuition is if an explanation lies in the tangent space of the image, the explanation will contain meaningful components of the image [9]. Brodt et al. [9] demonstrate their hypothesis on several gradient explainability models on well-known computer vision datasets. Brodt el al. [9] further demonstrate tangentially aligned explanations are robust to adversarial attacks.

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland
*Corresponding author.
lachlan.simpson@adelaide.edu.au (L. Simpson)

Integrated gradients (IG) [6] is a popular explainability method employed in a wide array of computer vision tasks [15]. IG relies on a hyper-parameter known as the base-point. The choice of base-point fundamentally alters the explanation provided [16]. Base-point selection is domain dependent and chosen heuristically. The zero vector, however, is a prevalent choice in computer vision, NLP and graph machine learning [6, 17, 18]. Several works have investigated different choices of base-point, however, none are able to determine a correct choice [19]. In this work we investigate the conditions under which a choice of base-point will provide perceptually aligned explanations.

The contributions of this work are twofold:

1. We provide sufficient conditions for when integrated gradient explanations are tangentially aligned. We extend these results to any base-point attribution method.

2. We provide a framework to choose a base-point point which provides meaningful explanations to the user. We compare our method with three gradient explainability models and IG with common base-points. We demonstrate that our base-point choice provides better tangential alignment and consequently more meaningful explanations. We validate our approach on four well-known computer vision datasets.

The remainder of this work is structured as follows: Section 2 provides related work and background. Section 3 investigates theoretical conditions for tangential alignment of base-point attribution methods. Section 4 calculates base-points for tangential alignment of IG on four well known datasets. We compare tangential IG with four common base-point choices and three gradient explainability models. We conclude in Section 5 with a discussion for future works.

2. Related Work and Background

2.1. Tangentially Aligned Integrated Gradients Explanations

Post hoc explainability models are methods for providing an attribution for the features that influence the output of a neural network. Post hoc explainability is a step towards addressing the black-box problem [6].

Base-point attribution methods (BAM) [20] are a specific class of post hoc explainability models. A BAM is a function

$$A\colon M \times M \times \mathcal{F}(M) \to \mathbb{R}^d \tag{1}$$

$$(x, x', F) \mapsto A(x, x', F) \tag{2}$$

where, $M \subset \mathbb{R}^d$ is a manifold, $\mathcal{F}(M)$ denotes the set of neural networks on M and $x, x' \in M$ are an input and a base-point, respectively.

We will further restrict the space of BAM functions to path methods, and we will generalise the definition of path methods to be independent of coordinates. Given a closed interval $I := [a, b] \subset \mathbb{R}$, a path $\gamma\colon I \to M$ and a unit vector $v \in \mathbb{R}^d$, the component of a path method $A^\gamma : M \times M \times \mathcal{F}(M) \to \mathbb{R}^d$ in the direction of v is defined as

$$A^\gamma_v(x, x', F) = \int_a^b \langle \nabla F(\gamma(t)), v\rangle \langle \gamma'(t), v\rangle \mathrm{d}t. \tag{3}$$

In this way, for a given orthonormal basis $\{v_1, \ldots, v_d\}$ of $\mathbb{R}^d$, A^γ is expressed as

$$A^\gamma(x, x', F) = \sum_{i=1}^{d} A^\gamma_{v_i}(x, x', F) v_i. \tag{4}$$

Particularly, for the standard orthonormal basis $\{e_1, \ldots, e_d\}$ of $\mathbb{R}^d$, we obtain the usual definition

$$A^\gamma_{e_i}(x, x', F) = \int_a^b \frac{\partial F}{\partial x_i}(\gamma(t)) \frac{\partial \gamma_i}{\partial t}(t) \mathrm{d}t. \tag{5}$$

The prominent path method, integrated gradients [6] is a path method where γ is taken to be the straight line between points $x, x' \in M$. For any pair of points $x, x' \in M$, a neural network $F \in \mathcal{F}(M)$, and a unit vector v, integrated gradients of the v component of x is defined to be:

$$\mathrm{IG}_v(x, x', F) := \langle x - x', v\rangle \int_0^1 \langle \nabla F(x' + t(x - x')), v\rangle \mathrm{d}t. \tag{6}$$

Letting $\mathrm{I} : M \times M \times \mathcal{F}(M) \to \mathbb{R}^n$ be the map defined by

$$\mathrm{I}(x, x', F) := \int_0^1 (\nabla F)(x' + t(x - x'))\mathrm{d}t, \tag{7}$$

integrated gradients can be expressed succinctly in the standard orthonormal basis of $\mathbb{R}^d$ as

$$\mathrm{IG}(x, x', F) = (x - x') \odot \mathrm{I}(x, x', F), \tag{8}$$

where $\odot$ denotes the Hadamard product.

Several metrics have been proposed to measure the quality of explainability models. In [7, 4], Lipschitzness is proposed as a measure of explainability quality. Other works consider the extent an explainability model approximates the neural network as a measure of quality. Brodt et al. [9] propose the extent to which an explanation lies in the tangent space of the manifold as a measure of explanation quality. Attributions which lie in tangent space were demonstrated to constitute the meaningful features that contribute to a prediction [8, 9]. Orthogonal attributions were closer to random noise. The hypothesis that tangential explanations provide meaningful explanations is validated on several image classification datasets and a user study [9]. Here tangentially aligned explanations is formalised.

For the reminder of this work we will consider $\mathbb{R}^d$, equipped with its standard inner product $\langle \cdot, \cdot \rangle$, and we will let $M \subset \mathbb{R}^d$ be a manifold of dimension $n < d$. We will also write $\langle \cdot, \cdot \rangle$ for the restriction of the inner product of $\mathbb{R}^d$ to M, such that $(M, \langle \cdot, \cdot \rangle)$ is an embedded Riemannian submanifold of $(\mathbb{R}^d, \langle \cdot, \cdot \rangle)$. We will denote the tangent space of M at a point x by T_xM which, in the context of data manifolds, will consist of all $v \in \mathbb{R}^d$ such that $x + v$ is "close" to M, with $\|v\|_2$ small [9]. Lastly, making use of the inner product of $\mathbb{R}^d$, for each $x \in M$ we have orthogonal direct sum decomposition $T_x\mathbb{R}^d = T_xM \oplus T_xM^\perp$,where

$$T_xM^\perp := \{u \in T_x\mathbb{R}^d \ : \ \langle u, v\rangle = 0, \ \forall v \in T_xM\}. \tag{9}$$

We will let $\pi_x : T_x\mathbb{R}^d \to T_xM$ denote the natural projection from $T_x\mathbb{R}^d$ to T_xM defined by

$$\pi_x(v) = \sum_{\ell=1}^{n} \langle v, \tau_\ell\rangle \tau_\ell, \tag{10}$$

where, $\{\tau_1, \dots \tau_n\}$ is an orthonormal basis for T_xM. We define the map $\mu_x : T_x\mathbb{R}^d \to [0, 1]$, given by

$$\mu_x(v) := \frac{\|\pi_x v\|_2^2}{\|v\|_2^2}, \qquad v \in T_x\mathbb{R}^d. \tag{11}$$

The map defined in Equation (11) provides us a measure of "how much" of a vector lies in the tangent space of M at x, i.e. a vector v is in T_xM if and only if $\mu_x(v) = 1$ and, on the other hand, v will be in $T_xM^\perp$ if and only if $\mu_x(v) = 0$, which can be observed directly from its definition. Moreover, letting $\pi_x^\perp : T_x\mathbb{R}^d \to T_xM^\perp$ denote the natural projection and, noting that,

$$v = \pi_x v + \pi_x^\perp v, \tag{12}$$

$$\|v\|_2^2 = \|\pi_x v\|_2^2 + \|\pi_x^\perp v\|_2^2, \tag{13}$$

we can express μ_x as

$$\mu_x(v) := \frac{\|\pi_x v\|_2^2}{\|\pi_x v\|_2^2 + \|\pi_x^\perp v\|_2^2}, \qquad v \in T_x\mathbb{R}^d. \tag{14}$$

Minimising the norm of the projection into $T_xM^\perp$ provides a framework to ensure tangential alignment.

2.2. Base-point Selection for Integrated Gradients

The attribution of IG depends on the base-point chosen. Base-point selection is domain dependent and chosen heuristically. Here we review common base-point choices as provided by [21].

1. **Zero**. Here the base-point for all points is a constant zero vector

$$\alpha^{\text{zero}} = 0. \tag{15}$$

   in general the zero base-point can be any constant vector.

2. **Maximum Distance**. For a given input $x \in M$, α is defined as the point in M of maximum distance from x i.e.

$$\alpha_x^{\max} = \underset{y \in M}{\operatorname{argmax}} \|x - y\|_p. \tag{16}$$

   Usually $p = 1$ or 2.

3. **Uniform**. We sample uniformly over a valid range of M

$$\alpha_i^{\text{uniform}} \sim U(\min_i, \max_i). \tag{17}$$

4. **Gaussian**. A Gaussian filter is applied to the input x.

$$\alpha^{\text{Gaussian}} = \sigma \cdot v + x, \tag{18}$$

   where, $v_i \sim \mathcal{N}(0, 1)$ and $\sigma \in \mathbb{R}$. We require the α^{Gaussian} is still within the data distribution so $\alpha^{\text{Gaussian}} \to \alpha^{\text{Uniform}}$ as $\sigma \to \infty$ [21].

The zero base-point (Equation 15) will not highlight the aspects of the image which may be important if the object of interest contains black pixels [21, 22]. To address the issue of a constant base-point missing important features maximum distance (Equation 16) was proposed in [21]. Maximum distance takes the furthest point (in ℓ_p distance) from the input image such that the base-point does not contain important information of the input. Another alternative is to sample a base-point from a distribution such as uniform (Equation 17) or Gaussian (Equation 18) [21, 23]. Despite the various choices of base-point we demonstrate none of the aforementioned base-points provide perceptually aligned explanations.

Zaher et al. [24] propose Manifold Integrated Gradients (MIG). MIG replaces the straight line in IG with a geodesic such that the attribution lies in the Riemannian manifold. Whilst MIG addresses the problem of IG not conforming to the geometry of the data. MIG does not resolve the issue of base-point choice nor does MIG ensure that the attribution lies in the tangent space of the manifold.

3. Optimising the Base-point for Tangentially Aligned Explanations

Throughout this section, we will study the map defined in Equation 11, to identify possible choices of base-points for the attribution given by a BAM to be tangent to M at a point. To be precise, for a given BAM, we want to find $\alpha \in M$ such that the map

$$x' \mapsto \mu_x(A(x, x', F)) \tag{19}$$

attains its maximum and, particularly, when this maximum value is equal to 1. We note that $\alpha = x$ is always a solution, however, we will always require $\alpha \neq x$ for non-trivial solutions.

Definition 1. *Let $A : M \times M \times \mathcal{F}(M) \to \mathbb{R}^d$ be a BAM and $x, \alpha \in M$, $F \in \mathcal{F}(M)$. A is tangentially aligned at x, with base-point α, if $\mu_x(A(x, \alpha, F)) = 1$.*

In the remainder of this section $x \in M$ and $F \in \mathcal{F}(M)$ will be fixed, unless otherwise stated. Letting $\pi_x^\perp : T_x\mathbb{R}^d \to T_xM^\perp$ denote the natural projection and defining the maps

$$H_x : M \to T_xM^\perp, \quad H_x(x') := \pi_x^\perp A(x, x', F) \tag{20}$$

and

$$E_x : M \to \mathbb{R}, \quad E_x(x') := \frac{1}{2}\|H_x(x')\|_2^2, \tag{21}$$

we can characterise tangentially aligned BAM explanations with the following theorem.

Theorem 1. *Let $A : M \times M \times \mathcal{F}(M) \to \mathbb{R}^d$ be a BAM and $x, \alpha \in M$, $F \in \mathcal{F}(M)$. Then A is tangentially aligned at x, with base-point α, if and only if $H_x(\alpha) = 0$ or, equivalently, if $E_x(\alpha) = 0$.*

Proof. It is immediate from the definitions of H_x and E_x, since they are the projection to $T_xM^\perp$ of A and a multiple of its norm, respectively. □

Choosing an orthonormal basis

$$\{\tau_1, \ldots, \tau_n, \nu_{n+1}, \ldots, \nu_d\} \tag{22}$$

of $T_x\mathbb{R}^d$ such that $\{\tau_i\}_{i=1}^n$ and $\{\nu_i\}_{i=n+1}^d$ are orthonormal basis of T_xM and $T_xM^\perp$, respectively, we observe that

$$H_x(x') \;=\; A(x, x', F) - \sum_{i=1}^{n} \langle A(x, x', F), \tau_i\rangle \tau_i = \sum_{i=n+1}^{d} \langle A(x, x', F), \nu_i\rangle \nu_i \tag{23}$$

and

$$E_x(x') = \frac{1}{2}\sum_{i=n+1}^{d} \langle A(x, x', F), \nu_i\rangle^2. \tag{24}$$

Therefore, any choice of a basis for $T_x\mathbb{R}^d$, adapted to the splitting of $T_x\mathbb{R}^d$ into tangent and normal spaces of M at x, will provide us with with a system of equations to test for tangentially aligned explanations.

Theorem 1 provides us with a necessary condition that a base-point must satisfy to obtain a tangentially aligned explanation. To observe this, suppose that there exists $\alpha \in M$ such that $A(x, \alpha, F)$ is tangentially aligned. Then, by Theorem 1, $E_x(\alpha) = 0$ and since $E_x(x') \geq 0$ for all $x' \in M$, it is in fact a global minimum of E_x and, consequently, $(\nabla E_x)(\alpha) = 0$. Moreover, its Hessian matrix $\mathrm{Hess}E_x$ is positive definite at α.

To simplify notation, in what follows we will denote the partial derivatives with respect to x_i and x_i' by ∂_i and ∂_i', respectively.

Corollary 2. *It is a necessary condition for $A(x, \alpha, F)$ to be tangentially aligned, that*

$$\langle H_x(\alpha), (\partial_i' H_x)(\alpha)\rangle = 0, \tag{25}$$

for all $i = 1, \ldots, d$.

Proof. If $A(x, \alpha, F)$ is tangentially aligned, then $(\nabla E_x)(\alpha) = 0$, which is equivalent to $(\partial_i' E_x)(\alpha) = 0$ for all $i = 1, \ldots, d$. It follows from the definition of E_x that:

$$\langle H_x(\alpha), (\partial_i' H_x)(\alpha)\rangle = \frac{1}{2}\partial_i' \langle H_x, H_x\rangle|_\alpha = (\partial_i' E_x)(\alpha) = 0. \tag{26}$$

for all $i = 1, \ldots, d$, as claimed. □

In order to find conditions for the Hessian matrix of E_x to be positive definite, we will make use of *Geršgorin circle theorem* [25] to find bounds for the eigenvalues of Hess E_x. For a given complex $n \times n$ matrix A, its i-th *Geršgorin disk* is the closed disk $G_i(A) := D(A_{ii}, R_i) \subset \mathbb{C}$, where the radius is given by the formula

$$R_i = \sum_{j \in J_i} |A_{ij}|, \quad J_i = \{1, \ldots, i-1, i+1, \ldots, n\}. \tag{27}$$

Lemma 3. *Let A be a real symmetric matrix such that $A_{ii} > R_i$ for all i, then A is positive definite.*

Lemma 3 follows immediately from [25]. The following theorem is an immediate consequence of Corollary 2 and of Lemma 3 applied to HessE_x.

Theorem 4. *It is a sufficient condition for $A(x, \alpha, F)$ to be tangentially aligned, that for all i*

$$\langle H_x(\alpha), (\partial_i' H_x)(\alpha) \rangle = 0 \tag{28}$$

and that

$$(\text{Hess } E_x)(\alpha)_{ii} > R_i(\alpha), \tag{29}$$

where $R_i(\alpha)$ denotes the radius of the i-th Geršgorin disk of $(\text{Hess}E_x)(\alpha)$.

4. Numerical Analysis

In this section we approximate tangential base-point choices on four well-known datasets in computer vision: MNIST [26], Fashion-MNIST [27], CIFAR10 and FER2013 [28]. We demonstrate that the four common base-point choices defined in Section 2.2 consistently provide explanations that are not well aligned with the tangent space. We further demonstrate tangentially aligned IG provides higher tangentially aligned explanations than three gradient explainability models: Gradient [29], Smooth Grad (SG) [5] and Input*Gradient (I*G) [30].

4.1. Approximating the Tangent and Normal Space

Following [9] the tangent space is approximated via a convolutional autoencoder. As discussed in [31] if we consider the decoder, $\text{dec} : L \to M$, as a map from the latent space L to the manifold M, then the Jacobian of the decoder is a linear map from the tangent spaces of L and M

$$J_{\text{dec}}(x) : T_x L \to T_{\text{dec}(x)} M. \tag{30}$$

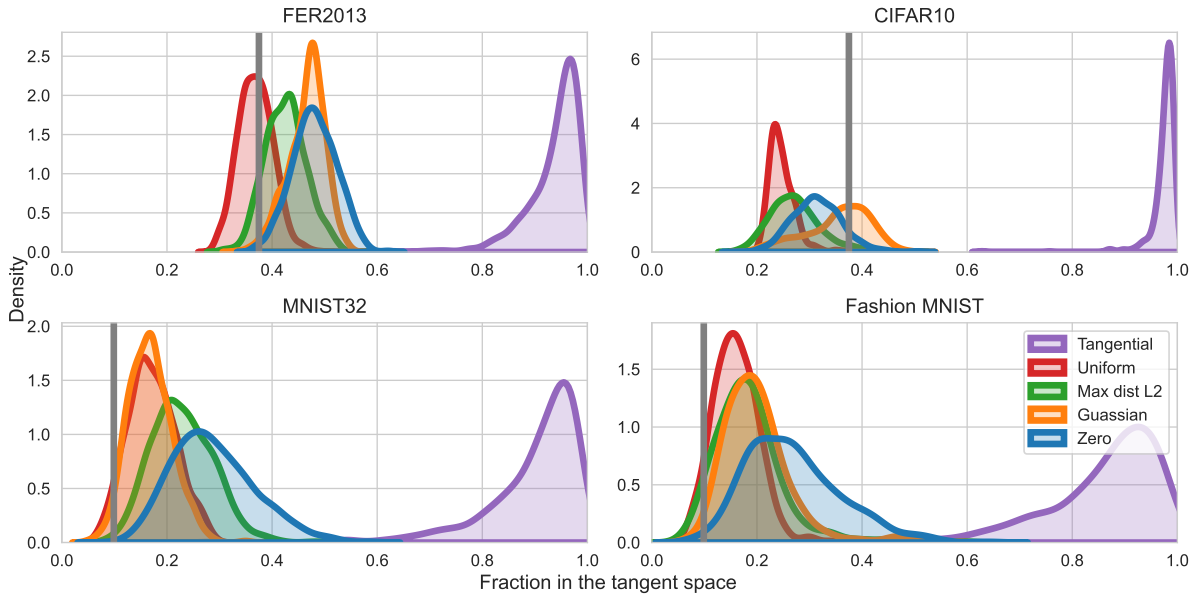
The Jacobian of the decoder can be computed via back-propagation [31]. The tangent space of M is spanned by the gradient of dec [9]. For our work we require the normal space $T_x M^{\perp}$. Given a basis for the tangent space $\{\tau_1, \ldots, \tau_n\}$, one can compute a basis for the normal space by

$$\text{Null}\left(\tau_1, \ldots, \tau_n\right), \tag{31}$$

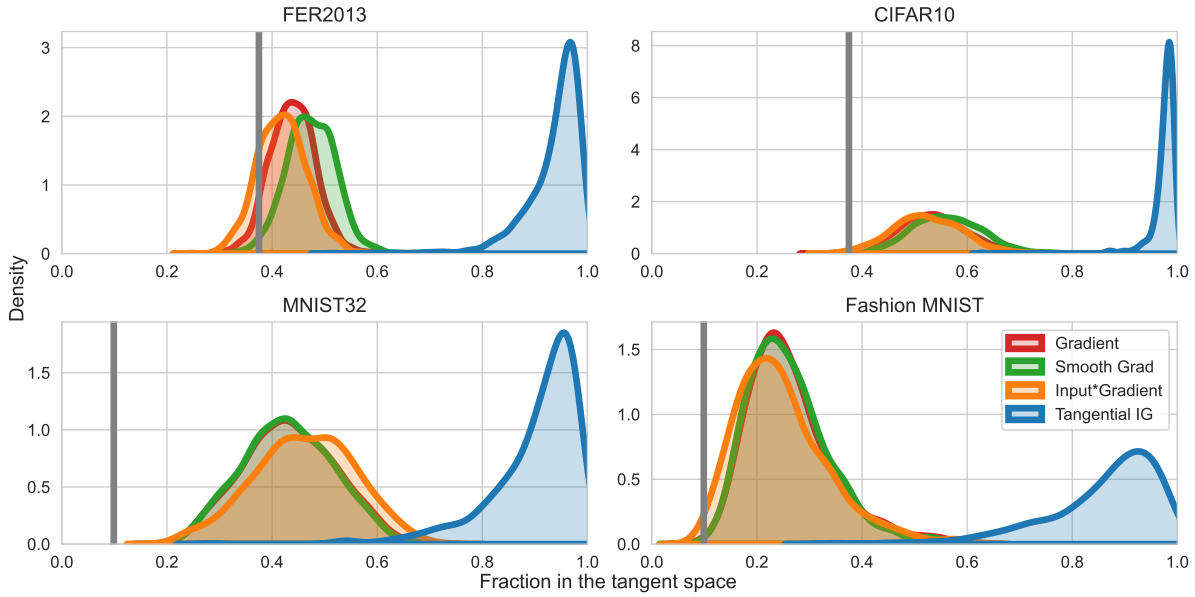
where one considers the basis of the tangent space as a matrix.

4.2. Experimental Setup

We utilise the implementation of [9] to generate the tangent space with a convolutional autoencoder and train a CNN for classification. The convolutional autoencoder has two convolutional layers with pooling followed by a fully connected layer with ReLU activation. A two layer CNN of kernel size 3 with dropout and Relu activation is used to perform image classification. Using the parameters of [9], $n = \dim(T_x M) = 144$ for CIFAR10 and FER2013 and for MNIST32 and Fashion $n = 10$. Explainability models are produced with the PyTorch library Captum.ai [32].



(a) Base-point choices.



(b) Gradient models.

Figure 1: Kernel density estimate plot of the fraction of the explanation in the tangent space with **(a)** different base-point choices and **(b)** different gradient explainability models. The fraction of explanation in the tangent space is measured with μ_x (Equation 11). The vertical line represents the expected fraction a random vector lies in the tangent space $\approx \sqrt{n/d}$, where $n = \dim(T_x M)$ and $d = \dim(M)$. On CIFAR10 and FER2013, $n = 144$. On MNIST32 and Fashion-MNIST, $n = 10$.

4.3. Complexity Analysis

The problem of finding a base-point that gives tangentially aligned explanations can be phrased as

$$\alpha_x^* = \underset{x \neq \alpha}{\operatorname{argmin}} \, E_x(\alpha). \tag{32}$$

If we suppose $M \subseteq \mathbb{R}^d$ is compact, then by Weierstrass's theorem such an α^* exists [33]. The continuity of E_x follows from the continuity of IG and norms. The condition $x \neq \alpha$ is required to ensure non-trivial

solutions. A solution to the optimisation problem in Equation 32 can be approximated via gradient-descent. We note that zero, Gaussian and Uniform base-points have constant time $\mathcal{O}(1)$ complexity. Maximum ℓ_2 distance is $\mathcal{O}(|D|)$ where $|D|$ is the number of points in the dataset. Calculating a tangential base-point has complexity $\mathcal{O}(\varepsilon)$, where ε is the number of iterations in gradient descent to solve Equation 32. IG with base-point α_x^* from Equation 32 will be referred to as tangentially aligned IG.

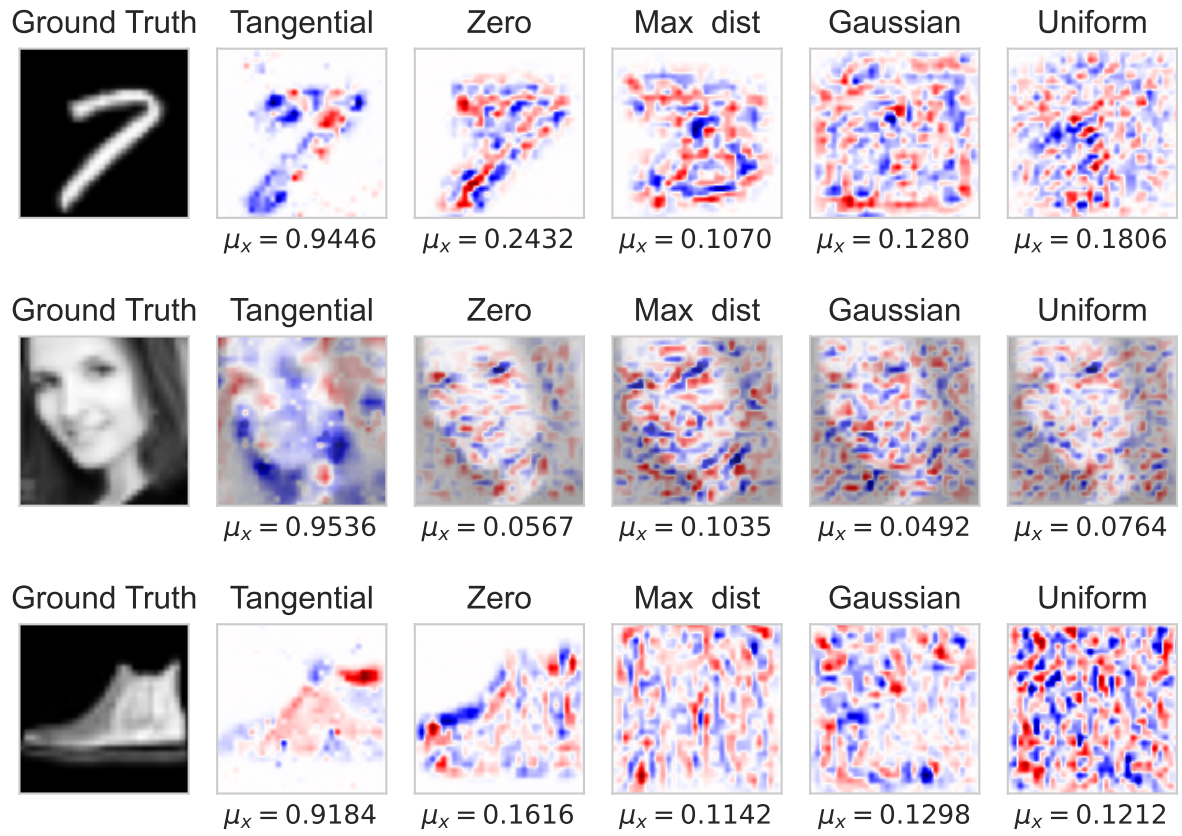


Figure 2: Attributions of IG with differing base-point choice on example points from MNIST, FER2013, and Fashion-MNIST. The fraction of explanation in the tangent space is denoted by μ_x (Equation 11).

4.4. Comparison of Different Base-points with Tangential Integrated Gradients

For each dataset IG is applied to the CNN with the base-points defined in Section 2.2 and the fraction of each explanation is calculated via Equation 11. To calculate each base-point in Section 2.2 we use the implementation provided by [21].

For each point we approximate the solution to Equation 32 to provide tangential explanations on each dataset. To approximate the solution to Equation 32 over all points we use the same learning rate and number of iterations. Some points may require a different learning rate and number of iterations to achieve higher tangential alignment. We leave this to future work. In Figure 1a we have the distributions of the fraction in the tangent space on FER2013, CIFAR10, MNIST32 and Fashion MNIST. We see in Figure 1a that approximating solutions to Equation 32 consistently provides explanations with high tangential alignment.

We see in Figure 1a that the uniform base-point provides explanations consistently close to the normal space; followed by maximum ℓ_2 distance and Gaussian. We note that on FER2013 and CIFAR10 the Gaussian base-point performs better than zero, uniform, and maximum ℓ_2 distance. The better performance of a Gaussian base-point is likely due to the smoothing parameter σ defined in Section 2.2. It is the goal of future work to determine the impact of σ on tangential alignment. The vertical lines

in Figure 1a indicate the expectation a random vector will lie in the tangent space. The expectation is approximately $\sqrt{n/d}$, where n and d are the dimensions of the tangent space approximation and manifold, respectively [9]. An explanation is therefore sufficiently aligned with the tangent space when that fraction in the tangent space is greater than $\sqrt{n/d}$. We see in Figure 1a that standard base-point choices on CIFAR10 are significantly below the vertical line. It is the goal of future work to determine if the dimension of the tangent space of CIFAR10 of $n = 144$ or the parameter of the Gaussian base-point impacts the tangential alignment of IG on CIFAR10.

We provide in Figure 2, example integrated gradient explanations for a point on MNIST32, FER2013, and Fashion-MNIST with differing base-point choice. We see that our method provides tangentially aligned explanations with $\mu_x > 0.91$ for all datasets. The tangentially aligned integrated gradient attributions are clear and perceptually aligned with the object to classify in the image. We see in Figure 2 that uniform, maximum ℓ_2 distance, and Gaussian are consistently random noise.

4.5. Comparison of Gradient Explainability Models with Tangential Integrated Gradients

In this section we compare tangentially aligned integrated gradients with three common gradient explainability models: Gradient, Smooth Grad and Input *Gradient. The aforementioned gradient explainability models do not require a base-point choice. We demonstrate that tangentially aligned integrated gradients significantly improves upon integrated gradients. The gradient explainability models for a given model are defined as follows:

1. **Gradient** The gradient of a model f at $x \in \mathbb{R}^d$ for class i is defined as:

$$\mathrm{grad(x)_i} \coloneqq \frac{\partial \mathrm{f(x)_i}}{\partial \mathrm{x}}. \tag{33}$$

2. **Smooth Grad** We define Smooth Grad with n samples and standard deviation σ as:

$$\mathrm{SmoothGrad}(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(x+a), \tag{34}$$

   where, $a \sim \mathcal{N}(0, \sigma^2)$. Following [9] we take $\sigma = 0.02$ and $n = 25$.

3. **Input*Gradient** Input*Gradient is defined as:

$$\mathrm{Input} * \mathrm{Gradient} \coloneqq x \odot \frac{\partial f(x)_i}{\partial x}. \tag{35}$$

In Figure 1b we provide density plots of the fraction an attribution is in the tangent space for: Gradient, Smooth Grad, Input*Gradient and tangentially aligned integrated gradients. We see in Figure 1b that tangentially aligned integrated gradients provides attributions consistently in the tangent space, out-performing the aforementioned gradient explainability models. In Figures 1a and 1b Gradient, Smooth Grad and Input*Gradient provide better tangential alignment than the common base-point choices provided in Section 2.2 on MNIST and CIFAR10. On Fashion-MNIST we see that the zero base-point choice provides comparable performance with Gradient, Smooth Grad and Input*Gradient. In Figures 1a and 1b we see that on FER2013, Gradient, Smooth Grad and Input*Gradient perform similarly to Gaussian, maximum ℓ_2 distance and zero base-point choices for integrated gradients. All gradient models on FER2013 outperform the uniform base-point choice for integrated gradients. We see in Figures 1a and 1b, Gradient, Smooth Grad, and Input*Gradient tend to out-perform Integrated gradients with standard the standard base-point choices. Tangential integrated gradients out-performs the aforementioned gradient explainability models and standard base-point choices.

5. Conclusions and Future Work

In this work we investigated how to choose base-points for IG that provide tangentially aligned explanations. We provided theoretical conditions for a base-point to provide tangentially aligned explanations for any BAM. We demonstrated how to numerically approximate the base-point which provides tangentially aligned explanations and validated this approach on several well-known image classification datasets. In future work we seek to further investigate the theoretical conditions a base-point must have to provide tangential explanations.

Acknowledgments

The Commonwealth of Australia (represented by the Defence Science and Technology Group) supports this research through a Defence Science Partnerships agreement. Lachlan Simpson is supported by a scholarship from the University of Adelaide.

References

[1] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).

[2] C. Zednik, Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence, Philosophy & Technology 34 (2021) 265–288.

[3] T. J. Sejnowski, The Unreasonable Effectiveness of Deep Learning in Artificial Intelligence, Proceedings of the National Academy of Sciences 117 (2020) 30033–30038.

[4] L. Simpson, K. Millar, A. Cheng, C.-C. Lim, H. G. Chew, Probabilistic Lipschitzness and the Stable Rank for Comparing Explanation Models, arXiv preprint arXiv:2402.18863 (2024).

[5] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: Removing Noise by Adding Noise, arXiv preprint arXiv:1706.03825 (2017). arXiv:1706.03825.

[6] M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks, Proceedings of the 34th International Conference on Machine Learning (ICML) 70 (2017) 3319–3328.

[7] Z. Khan, D. Hill, A. Masoomi, J. Bone, J. Dy, Analyzing Explainer Robustness via Lipschitzness of Prediction Functions, arXiv preprint arXiv:2206.12481 (2023).

[8] R. Ganz, B. Kawar, M. Elad, Do Perceptually Aligned Gradients Imply Robustness?, in: Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 10628–10648.

[9] S. Bordt, U. Uddeshya, Z. Akata, U. von Luxburg, The Manifold Hypothesis for Gradient-Based Explanations, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 3697–3702.

[10] N. Whiteley, A. Gray, P. Rubin-Delanchy, Statistical Exploration of the Manifold Hypothesis, arXiv preprint arXiv:2208.11665 (2024).

[11] C. Fefferman, S. Mitter, H. Narayanan, Testing the Manifold Hypothesis, Journal of the American Mathematical Society 29 (2016) 983–1049.

[12] I. J. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, USA, 2016.

[13] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness May Be at Odds with Accuracy, International Conference on Learning Representations (2019).

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, International Conference on Learning Representations (2018).

[15] A. Das, P. Rad, Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey, arXiv preprint arXiv:2006.11371 (2020). arXiv:2006.11371.

[16] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, B. Kim, The (Un)reliability of Saliency Methods, arXiv preprint arXiv:1711.00867 (2017). arXiv:1711.00867.

[17] P. Xenopoulos, G. Chan, H. Doraiswamy, L. G. Nonato, B. Barr, C. Silva, GALE: Globally Assessing Local Explanations, in: Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022, volume 196 of *Proceedings of Machine Learning Research (PMLR)*, 2022, pp. 322–331.

[18] B. Sanchez-Lengeling, J. Wei, B. Lee, E. Reif, P. Wang, W. Qian, K. McCloskey, L. Colwell, A. Wiltschko, Evaluating Attribution for Graph Neural Networks, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 5898–5910.

[19] D. Drakard, R. Liu, J. Yosinski, Exploring unfairness in Integrated Gradients based attribution methods, OpenReview (2022).

[20] D. Lundstrom, T. Huang, M. Razaviyayn, A Rigorous Study of Integrated Gradients Method and Extensions to Internal Neuron Attributions, Proceedings of the 39th International Conference on Machine Learning 162 (2022) 14485–14508.

[21] P. Sturmfels, S. Lundberg, S.-I. Lee, Visualizing the Impact of Feature Attribution Baselines, Distill (2020). Https://distill.pub/2020/attribution-baselines.

[22] M. Sundararajan, A. Taly, A Note About: Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values, arXiv preprint arXiv:1806.04205 (2018). arXiv:1806.04205.

[23] R. C. Fong, A. Vedaldi, Interpretable Explanations of Black Boxes by Meaningful Perturbation, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 3449–3457.

[24] E. Zaher, M. Trzaskowski, Q. Nguyen, F. Roosta, Manifold Integrated Gradients: Riemannian Geometry for Feature Attribution, arXiv preprint arXiv:2405.09800 (2024). arXiv:2405.09800.

[25] S. Geršgorin, Über die Abgrenzung der Eigenwerte einer Matrix, Bulletin de l'Académie des Sciences de l'URSS (1931) 749–754.

[26] L. Deng, The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web], IEEE Signal Processing Magazine 29 (2012) 141–142.

[27] H. Xiao, K. Rasul, R. Vollgraf, Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, arXiv preprint:1708.07747 (2017).

[28] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, Y. Bengio, Challenges in Representation Learning: A report on three machine learning contests, arXiv preprint arXiv:1307.0414 (2013).

[29] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, arXiv preprint arXiv:1312.6034 (2014). arXiv:1312.6034.

[30] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, arXiv preprint arXiv:1605.01713 (2017). arXiv:1605.01713.

[31] H. Shao, A. Kumar, P. T. Fletcher, The Riemannian Geometry of Deep Generative Models, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 428–4288.

[32] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A Unified and Generic Model Interpretability Library for PyTorch, arXiv preprint arXiv:2009.07896 (2020). arXiv:2009.07896.

[33] W. Rudin, Principles of Mathematical Analysis, McGraw Hill, 1976.