Latent Surface Attribution

Akanksha Das

Indian Statistical Institute Kolkata

October 10, 2025

Contents

What is Feature Attribution?

Feature Attribution methods explain machine learning model predictions by assigning importance scores to input features.

Formal Definition

Given a model $f : \mathbb{R}^n \to \mathbb{R}$ and input $\mathbf{x} \in \mathbb{R}^n$, feature attribution computes:

$$\phi(\mathbf{x}) = (\phi_1, \phi_2, \ldots, \phi_n)$$

where ϕ_i quantifies the contribution of feature x_i to the output $f(\mathbf{x})$.

Applications

- **Model Debugging**: Identify spurious correlations and biases
- **Scientific Discovery**: Reveal important biomarkers or genetic factors
- **Trust and Verification**: Build user confidence in model decisions
- **Model Refinement**: Can be used to refine attentions

Integrated Gradients

Integrated Gradients (IG) is a path-based feature attribution method that satisfies desirable axiomatic properties (completeness, sensitivity, implementation invariance).

Method

Given a model F, input x, and baseline x', IG computes attributions by integrating gradients along a straight-line path:

$$\mathrm{IG}_j(x, x') := (x_j - x'_j) \times \int_{t=0}^{1} \frac{\partial F(x' + t(x - x'))}{\partial x_j} dt$$

Limitations

- Linear path accumulates noise in feature visualizations
- Vulnerable to adversarial attributional attacks
- Path may deviate from natural data manifold

Manifold Integrated Gradients (MIG)

Manifold Integrated Gradients (MIG) addresses IG's limitations by integrating gradients along geodesics of the data manifold.

Key Innovation

- Uses VAE to learn Riemannian data manifold structure
- Computes geodesics (shortest paths) in latent space
- Integrates gradients along manifold-respecting paths.

Advantages over IG

- **Perceptually Aligned**: Produces cleaner, more intuitive feature visualizations
- **Robust**: More resistant to targeted attributional attacks
- **Manifold-Aware**: Path respects intrinsic data geometry
- **Smooth Interpolations**: Avoids saturation regions and noisy gradients

Visualizing World lines and World sheets

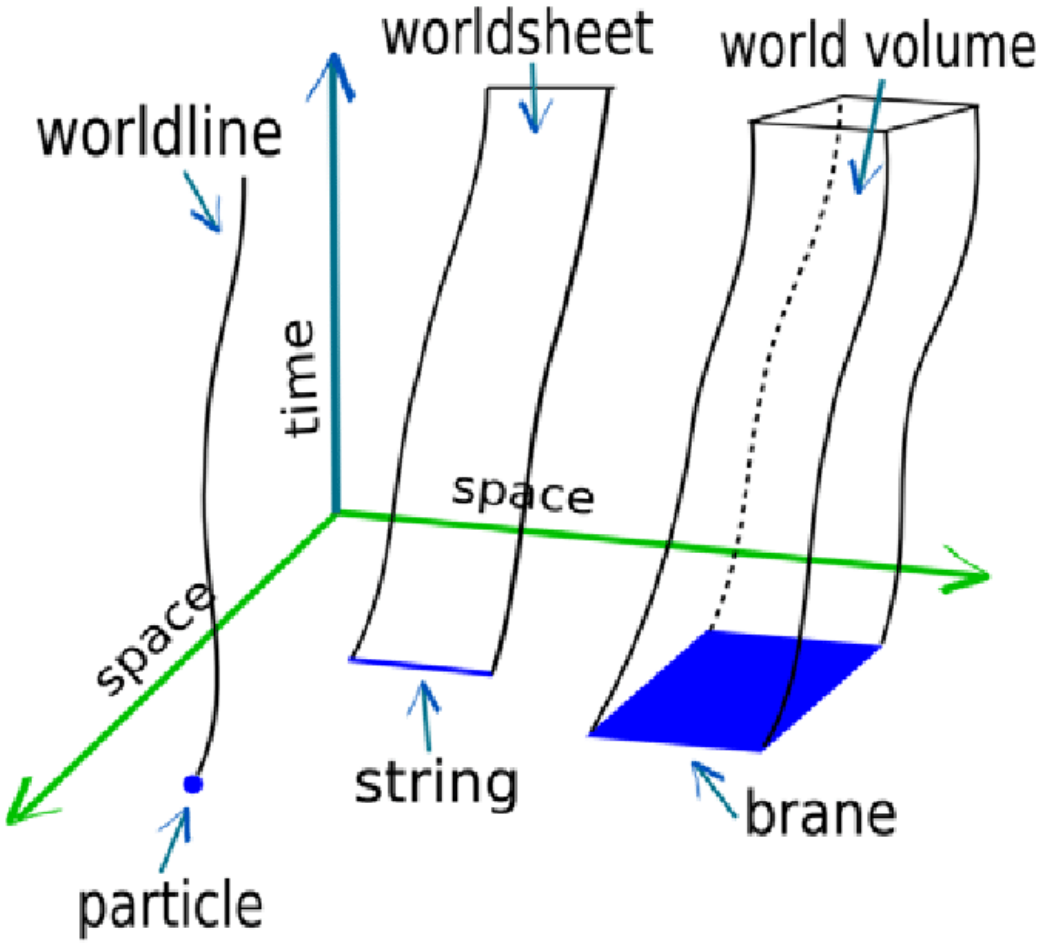

Illustration of a string worldsheet vs a world line traced by a particle

String Theory Inspiration – Nambu–Goto Action

- Defined as:

$$S_{\mathrm{NG}} = -T \int d^2\sigma \, \sqrt{-\det g_{\alpha\beta}}$$

- $T =$ string tension; $\sigma^\alpha = (\tau, \sigma)$ are worldsheet coordinates.
- $g_{\alpha\beta}$ is the **induced metric**:

$$g_{\alpha\beta} = \partial_\alpha X^\mu \partial_\beta X^\nu \eta_{\mu\nu}$$

- $X^\mu(\tau, \sigma)$: embedding of the worldsheet in spacetime.
- $\eta_{\mu\nu}$: Minkowski metric.

Weighted Nambu–Goto Action with Scalar Field

- Consider the action:

$$S = \int d\sigma \, d\tau \, \phi(x(\sigma, \tau)) \sqrt{\det g}$$

- This is a generalization of the Nambu–Goto action, where $\phi(x)$ is a scalar field defined on spacetime, pulled back to the worldsheet.
- Appears in string theory when the string couples to a background field, like the dilaton.
- The action computes a *weighted surface area*. Regions with higher ϕ are costlier for the surface to traverse.

Attribution Score Functional over a 2D Surface

Let $z : (\sigma, \tau) \mapsto \mathbb{R}^d$ be a latent 2D surface, and $D : \mathbb{R}^d \to \mathbb{R}^n$ a decoder mapping latent codes to input space. We define the attribution score as:

$$S_{\text{attr}}[z] = \int d\sigma \, d\tau \, \|\nabla f(D(z(\sigma, \tau)))\|_1 \sqrt{\det g}$$

Where:

- $f : \mathbb{R}^n \to \mathbb{R}$: scalar model output.
- $D(z(\sigma, \tau)) \in \mathbb{R}^n$: decoded surface in input space.
- $\nabla f(D(z)) \in \mathbb{R}^n$: gradient of f with respect to σ and τ.
- $\|\nabla f(D(z))\|_1$: L_1 norm of $\nabla f(D(z))$
- $J = \frac{\partial D(z(\sigma,\tau))}{\partial(\sigma,\tau)} \in \mathbb{R}^{n \times 2}$: Jacobian of the composed map $D \circ z$.
- $g = J^\top J \in \mathbb{R}^{2 \times 2}$: induced metric on the surface.

Attribution Score Functional over a 2D Surface

- $S_{\text{attr}}[z]$ measures the total attribution mass flowing through a 2D surface.
- Reduces to standard path-based attribution when $z(\sigma, \tau)$ collapses to a curve.
- The metric captures the geometry of the latent space i.e. the distribution of the training data.
- The derivative of the model captures the sensitivity of the model on the 2D surface.

Weighted Nambu-Goto vs. Attribution Score

Weighted Nambu–Goto	**Attribution Functional**
Worldsheet $z(\sigma, \tau) \subset \mathbb{R}^D$	Latent surface $z(\sigma, \tau) \in \mathbb{R}^d$
Target space $\mathbb{R}^D$	Input space $\mathbb{R}^n$
Embedding $X^\mu(\sigma, \tau)$	Decoder map $D(z(\sigma, \tau))$
Induced metric $g_{\alpha\beta}$	$g_{\alpha\beta} = J^\top J$, with $J = \frac{\partial D(z)}{\partial(\sigma,\tau)}$
Scalar Dilation Field $\phi(X)$	$\|\nabla f(x)\|_1$: gradient norm of the model

Decoder-Induced Geometry: Inspiration from MIG

- Manifold Integrated Gradients (MIG) introduces a latent curve $z(t) \in \mathbb{R}^d$, mapped to input space via a decoder $D(z(t)) \in \mathbb{R}^n$.
- The decoder induces a Riemannian geometry on the curve through the pullback metric:

$$g = J^\top J \in \mathbb{R}^{1\times 1}, \quad J = \frac{dD(z)}{dt} \in \mathbb{R}^{n\times 1}$$

- This scalar metric reflects how changes in latent space translate to displacements in input space.
- In our method, we use the same principle: a latent map $z(\sigma, \tau) \in \mathbb{R}^d$ is decoded into input space via $D(z) \in \mathbb{R}^n$.

Decoder-Induced Geometry: Inspiration from MIG

- The decoder again induces a Riemannian metric on the latent domain, now 2D:

$$g = J^\top J \in \mathbb{R}^{2\times 2}, \quad J = \frac{\partial D(z)}{\partial(\sigma, \tau)} \in \mathbb{R}^{n\times 2}$$

- As in MIG, the metric captures how geometric structure in latent space is stretched or distorted in input space.
- This metric g enters the attribution score via $\sqrt{\det g}$, measuring local surface area in input space.
- Thus, our use of decoder-induced geometry — via $g = J^\top J$ — is directly inspired by MIG's interpretation of latent-space attribution.

Extracting Feature Attributions from the Attribution Functional

Attribution functional:

$$S_{\text{attr}}[z] = \int d\sigma\, d\tau\, \|\nabla f(D(z(\sigma,\tau)))\|_1 \sqrt{\det g(\sigma,\tau)}$$

- For each input feature $i \in \{1,\ldots,n\}$, define a pointwise attribution density:

$$a_i(\sigma,\tau) = \left|\frac{\partial f}{\partial x_i}(D(z(\sigma,\tau)))\right| \cdot \sqrt{\det g(\sigma,\tau)}$$

- The total attribution to feature i is the surface integral:

$$S_{attr}(i) = \int a_i(\sigma,\tau)\, d\sigma\, d\tau$$

- This gives a feature-wise decomposition of the global score $S_{\text{attr}}[z] = \sum_i S_{attr}(i)$.

Local Attribution Surface

Purpose: Construct a 2D surface $z(\sigma, \tau) \in \mathbb{R}^d$ in latent space around $D^{-1}(x_0)$(x_0 : the input of interest) over which attribution is computed:

$$S_{\text{attr}}[z] = \int \|\nabla f(D(z(\sigma, \tau)))\|_1 \sqrt{\det(J^\top J)}\, d\sigma\, d\tau$$

Why constraints are necessary:

- To ensure the surface remains **localized around the input** of interest.
- To guarantee the surface lies on the **latent manifold** (so its image under D is meaningful in input space).
- To enforce a **fixed geometric scale** (e.g., fixed area), allowing attribution magnitude to be comparable across surfaces.

Constraints on the Local Attribution Surface

Let $z : [0,1]^2 \to \mathbb{R}^d$ be a smooth latent surface. We impose the following constraints:

- **Centering constraint:**

$$z\left(\tfrac{1}{2}, \tfrac{1}{2}\right) = z_0 \quad \text{where } D(z_0) = x_0$$

- **Manifold constraint:**

$$z(\sigma, \tau) \in \mathcal{L}(\text{Latent Space Manifold}), \quad D(z(\sigma, \tau)) \in \mathcal{M} \subseteq \mathbb{R}^n$$

- **Fixed area constraint:**

$$\int \sqrt{\det(J^\top J)}\, d\sigma\, d\tau = A_0$$

where $J = \frac{\partial D(z)}{\partial(\sigma,\tau)}$ is the Jacobian of the surface in input space.

These constraints ensure a stable and geometrically meaningful domain for attribution computation.

Axioms Satisfied by Path Attribution Methods

Path Attribution Methods are uniquely characterized by satisfying the following axioms:

- **Sensitivity (a):** If two inputs differ in one feature and the output differs, then that feature must receive non-zero attribution.
- **Sensitivity (b):** If the function f does not depend on a particular input variable x_i, then the attribution to x_i must be zero.
- **Implementation Invariance:** Two functionally equivalent networks (i.e., identical input-output mappings) must yield identical attributions.
- **Linearity:** For any $f = af_1 + bf_2$, attributions decompose linearly:

$$\text{Attr}_f(x) = a \cdot \text{Attr}_{f_1}(x) + b \cdot \text{Attr}_{f_2}(x)$$

- **Completeness:** The total attribution equals the difference in outputs:

$$\sum_{i=1}^{n} \text{Attr}_i(x) = f(x) - f(x')$$

Axiom Satisfied by Latent Surface Attribution

Axiom	**Latent Surface Attribution**
Sensitivity (a)	satisfies
Sensitivity (b)	satisfies
Implementation Invariance	satisfies
Linearity	no ; $\text{Attr}_f(x) \leq \lvert a\rvert \cdot \text{Attr}_{f_1}(x) + \lvert b\rvert \cdot \text{Attr}_{f_2}(x)$
Completeness	no ; $S_{\text{attr}}[z] = \sum_i S_{attr}(i)$

Salient Features

- Removes the requirement of **baselines** like in path attribution methods.
- The 2D surface makes the attribution **local** and **comparable**.
- Derived from the Mathematically grounded **Nambu Goto** Action.
- Instead of calculating Attributions at a point, it calculates it in the neighbourhood of a point thus **mitigates pointwise noise** by aggregating over a continuous neighborhood.
- Does not use **principle of least action** hence we do not need to solve a second order differential equation to calculate path.

Salient Features

- The baseline may be far from the input hence the attributions may integrate over a lot of distant points along the way that is avoided by our **hyper local 2D surface**.
- Most of **axioms** (except **linearity**) of Path Attribution Techniques are satisfied by our method.
- While **completeness** is not explicitly enforced, our method indirectly satisfies it by summing feature wise saliency contributions across a fixed-area surface centered at the input. This aggregation reflects the total local sensitivity of the model, capturing output variation in a neighborhood rather than approximating the net prediction difference with the baseline.

Implementation Details : The Latent Surface

Surface Definition: We define a circular latent surface $z(\sigma, \tau) \in \mathbb{R}^d$ centered at the latent code $z_0 = E(x_0)$:

$$z(\sigma, \tau) = z_0 + r\,[v_1 \cdot \cos(2\pi\sigma)\cos(\pi\tau) + v_2 \cdot \sin(2\pi\sigma)\cos(\pi\tau)]$$

How v_1, v_2 are chosen:

- We sample 1000 posterior vectors around z_0 from $q(z|x_0)$.
- We flatten them and apply **PCA** to obtain the top 2 principal directions v_1, v_2.
- These directions capture the highest local variance in the latent distribution.

Why This Surface?

- Ensures surface lies in a high-density region of the posterior.
- Locally captures model behavior in the most sensitive directions.
- Circular parametrization ensures smoothness and symmetry.
- The scalar r is chosen so that the surface area is fixed.

Implementation Details: Evaluation Metrics

AUC-ROC (Area Under ROC Curve):

- Measures the discriminative power of attributions.
- Sorted pixels by attribution $\rightarrow$ successively perturbed $\rightarrow$ model prediction monitored.
- AUC reflects how quickly the model prediction collapses when salient features are removed.

SIC (Sufficiency-informed Completeness):

- Inspired by the completeness axiom.
- Measures how well the most salient regions explain the model output.
- SIC AUC: Integrates model confidence over retained salient regions.
- Two types:
  - **Aggregated SIC AUC**: Single curve across the dataset
  - **Mean Individual SIC AUC**: Averaged per sample

Comparison Table: Latent Surface Attribution vs MIG

Metric	(Ours)	MIG	Percentage Increase
SIC AUC	**0.639**	0.546	17.03
AUC-ROC	**0.871**	0.702	24.07
Surface Area	0.3	N/A (Path)	–
Dimensionality	2D Surface	1D Path	–
Baseline Required	No	Yes	–

Table: Evaluation and methodological comparison between Latent Surface Attribution and MIG.

Future Directions

- We can find a **global 2D surface** of a fixed area that maximizes attribution score and test if training with only those input data reduces the prediction model accuracy and by how much.
- Instead of a 2D surface, we can define a surface with dimensions same as that of the Latent Space and define attribution as a **density function** which can be integrated over a surface of the latent space to get its corresponding attribution score.
- Finding a general framework that connects **String Theoretic Actions** with Attribution Methods.

Future Directions

- **Attribution-aware training:** Use surface attribution fields as regularizers to promote smoother, more interpretable model behavior during training.
- **Surface ensembles:** Investigate averaging attribution scores over multiple stochastic or learned surfaces centered at a point to increase robustness and reduce variance.
- **Cross-domain generalization:** Apply this framework to modalities beyond vision (e.g., Recommender Systems, NLP etc) by learning appropriate surface parametrizations in structured latent spaces.

Thank You

Akanksha Das

Advisor — Dr. Niloy Ganguly