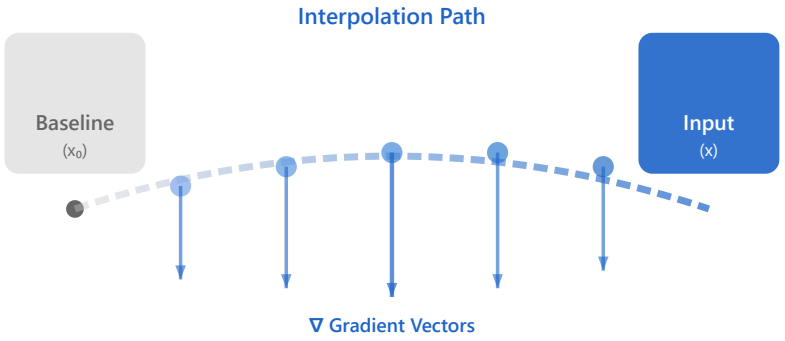
Integrated Gradients



Attribution Formula

$$IG(x) = (x - x_0) \times \int_{\alpha=0}^{\alpha=1} \nabla F(x_0 + \alpha(x - x_0))\, d\alpha$$

Sum of gradients along path

∫ Path Integration

Integrating gradients along interpolation path from baseline

🔍 Baseline Comparison

Comparing with neutral reference input (e.g., black image)

✓ Axiom Satisfaction

Meeting sensitivity and implementation invariance axioms

💪 Robust Attribution

More stable and reliable than simple gradient methods