AttEXplore: Attribution for Explanation with model parameters eXploration

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Jason Xue, Flora D Salim

THE UNIVERSITY OF SYDNEY

UNSW SYDNEY

ICLR

Introduction to DNN Challenges

- The critical role of DNNs in high-stakes domains.
- The need for reliability and interpretability.
- The complexity of interpreting non-linear, complex models.

What is AttEXplore?

- AttEXplore is a method integrating transferable attack techniques with attribution for DNNs.
- The goal of AttEXplore to provide more accurate and robust feature representation.
- The more important features are, the more likely they are to influence model decisions.

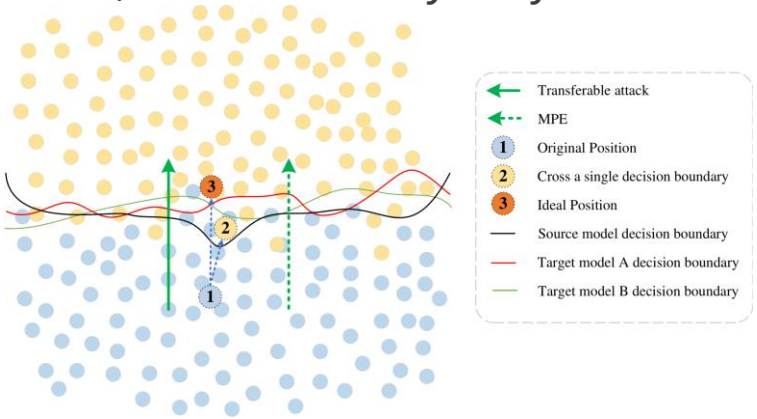

Figure 1: Decision boundaries

Model Parameter Exploration and Methods

- Definition of Model Parameter Exploration (MPE): Investigating how slight changes to model parameters or inputs affect the model's decision output.
- Example: $\mathrm{y} = w^T x$ with $w = [1, 2]$ and $x = [3, 4]$
- Methods to Explore:
  - Altering input features x (e.g., $x = [0, 4]$)
  - Modifying parameters w (e.g., $w = [0, 2]$)
- Both methods are shown to have equivalent effects on model decisions.

Applying MPE via Adversarial Methods and Attribution

- Connection of MPE with Transferable Adversarial Attacks:
  - Input transformations mimic MPE, aiming to cross decision boundaries.
- Attribution with MPE (AttEXplore):
  - Novel approach: Nonlinear integration path formula:

$$A = \int \triangle x^t \odot g(x^t) dt$$

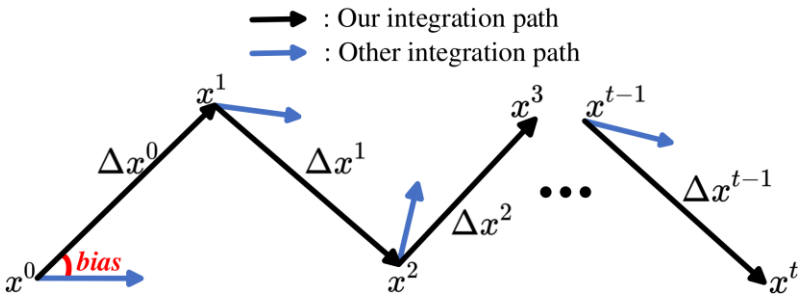

Figure 2: Our nonlinear integration path

Frequency-based Input Feature Alterations Method

- Applying DCT and IDCT for exploring model parameters in the frequency domain.
- Equations:

$$x_{f_i}^t = IDCT\left(DCT\left(x^t + N(0,1)\cdot\frac{\epsilon}{255}\right) * N(1,\sigma)\right)$$

$$\triangle x^t = \eta \cdot sign\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial L(x_{f_i}^t, y)}{\partial x_{f_i}^t}\right)$$

- Our algorithm satisfies Axioms of Sensitivity and Implementation Invariance

Experiments

- Dataset: ImageNet
- Models: Inception-v3, ResNet-50, and VGG16
- Baselines: AGI, BIG, DeepLIFT, GIG, EG, IG, Fast-IG, SM, SG, Grad-CAM
- Metrics: Insertion Score, Deletion Score, INFD score

Experiments

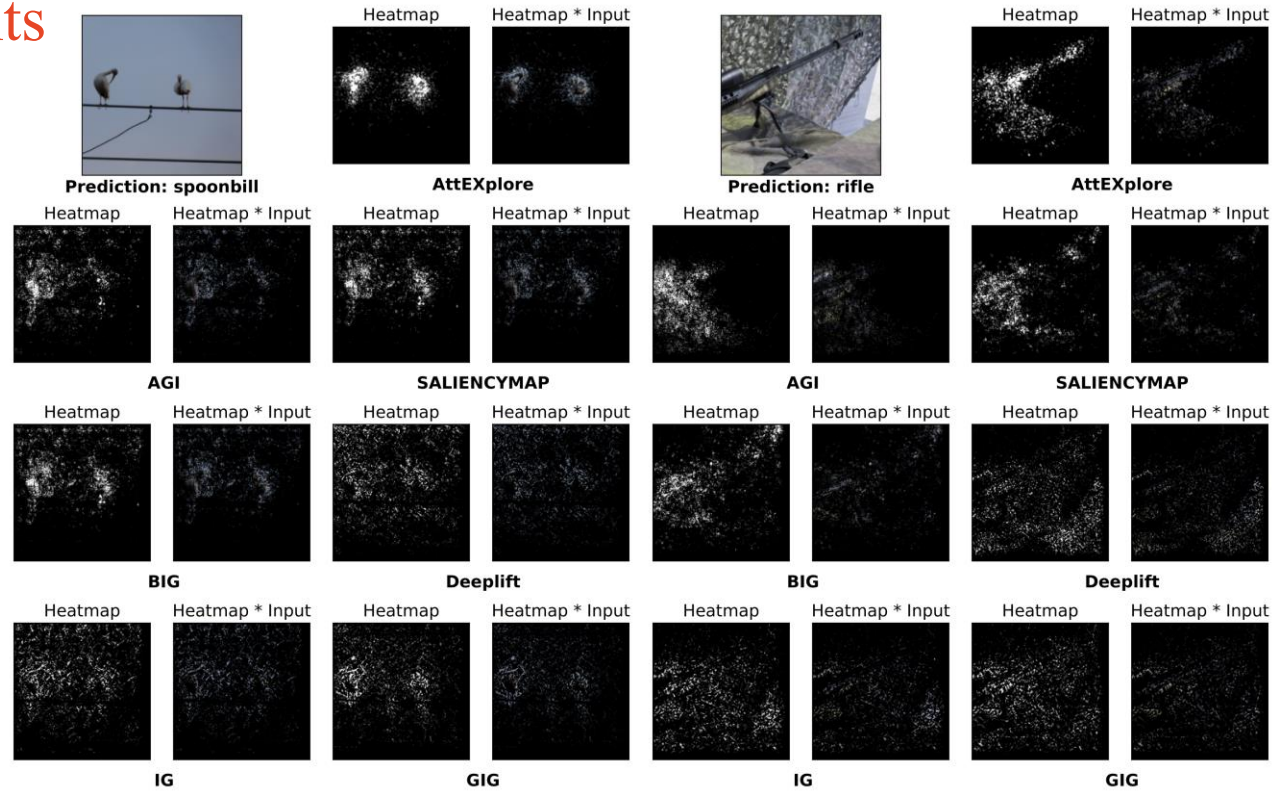

Figure 1: Visualization Results of our AttEXplore and Other Competitive Methods

Experiments

Methods	Inception-v3		ResNet-50		VGG-16	
	INS	DEL	INS	DEL	INS	DEL
Grad-CAM	0.4496	0.1084	0.2541	0.0942	0.3169	0.0841
BIG	0.3563	0.0379	0.2272	0.0415	0.1762	0.0303
SaliencyMap	0.3974	0.0422	0.256	0.048	0.2089	0.0323
DeepLift	0.216	0.0314	0.1246	0.0256	0.0827	0.0157
GIG	0.2584	0.0239	0.1308	0.0184	0.0859	0.0142
EG	0.2364	0.1656	0.256	0.2178	0.1959	0.1797
Fast-IG	0.146	0.0338	0.0889	0.0315	0.0623	0.0213
IG	0.2268	0.0284	0.1136	0.0247	0.0701	0.0173
SG	0.301	0.023	0.2357	0.0202	0.1423	0.015
AGI	0.4243	0.0439	0.3796	0.0465	0.2585	0.0319
AttEXplore	0.4732	0.0297	0.4197	0.0293	0.3186	0.0226

Table 1: Insertion&Deletion score comparison of AttEXplore and other competitive baselines

Method	FPS
BIG	3.3798
AGI	0.8818
IG	19.7461
SG	19.4942
GIG	2.2814
AttEXplore	47.2805

Table 2: FPS Results for Analysis of Time Complexity

Conclusion

1. We uncover, for the first time, the decision boundary exploration approaches of attribution and transferable attacks are consistent.
2. We propose a novel attribution algorithm by performing Attribution for Explanation with Model Parameter Exploration based on transferable attacks, named AttEXplore.
3. We conduct extensive experiments to verify the effectiveness of our AttEXplore.
4. We release the code of AttEXplore at: https://github.com/LMBTough/ATTEXPLORE


Thanks you