Explainable Adversarial Attacks on Coarse-to-Fine Classifiers

Akram Heidarizadeh [1] Connor Hatfield [1] Lorenzo Lazzarotto [2] HanQin Cai [3,4] George Atia [1,4]

[1]Dept. of Electrical and Computer Engineering, University of Central Florida, Orlando FL, USA [2]School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil

[3]Dept. of Statistics and Data Science, University of Central Florida, Orlando FL, USA [4]Dept. of Computer Science, University of Central Florida, Orlando FL, USA

Overview

- **Challenge:** Most traditional adversarial attacks such as DeepFool [5], PGD [3] and FGSM [2] focus on fooling the model but offer little to no explainability, making it difficult to understand how perturbations affect decisions.
- Hierarchical classifiers are largely unexplored in adversarial research.

- **Goal:** Our goal is to introduce an explainable adversarial attack that not only fools hierarchical classifiers but also provides insights into decision making process.

Coarse-to-Fine (C2F) Model Formulation

- M is the number of coarse classes and $[M] := \{1, 2, \ldots, M\}$.
- M_i is the number of fine classes associated with the i-th coarse label.
- **Coarse level:** $C : \mathbb{R}^N \to [M]$ assigns x to a coarse class such that:

$$C(x) = \operatorname{argmax}_{i\in[M]} C_i(x).$$

- **Fine level:** $F^i : \mathbb{R}^N \to [M_i]$ is the i-th fine classifier function. The finer class is obtained as:

$$F^i(x) = \operatorname{argmax}_{j\in[M_i]} F^i_j(x).$$

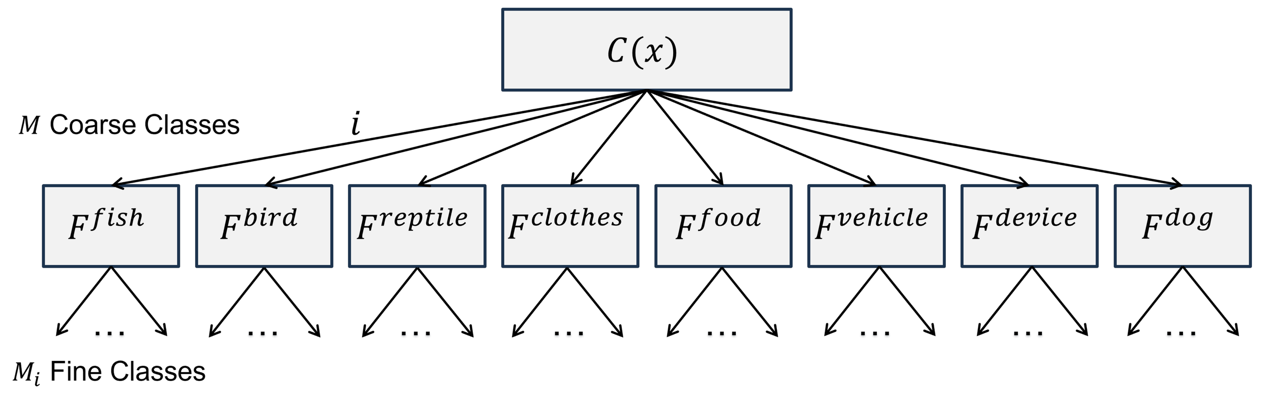

Figure 1. A coarse-to-fine classification model.

Layer-wise Relevance Propagation (LRP)

- LRP is a technique to determine the **contribution** of each pixel of the input data to the final **decision** [1].
- **Output layer:** The relevance is defined as: $R_i^L = \delta_{i,c}$, where $\delta_{i,c}$ (Kronecker delta) sets $R_i^L = 1$ when $i = c$ and $R_i^L = 0$ otherwise.
- **Intermediate layers:** The relevance scores are backpropagated using z+ rule:

$$R_i^l = \sum_j \frac{a_i^l (W^l)^+_{ij}}{\sum_k a_k^l (W^l)^+_{kj}} R_j^{l+1},$$

- **Input layer:** The relevance scores are calculated using the $z\beta$ rule [4]:

$$LRP_f(x;c) := R_i^0 = \sum_j \frac{a_i^0 W_{ij}^0 - l_i (W^0)^+_{ij} - h_i (W^0)^-_{ij}}{\sum_k (a_i^0 W_{kj}^0 - l_i (W^0)^+_{kj} - h_i (W^0)^-_{kj})} R_j^1,$$

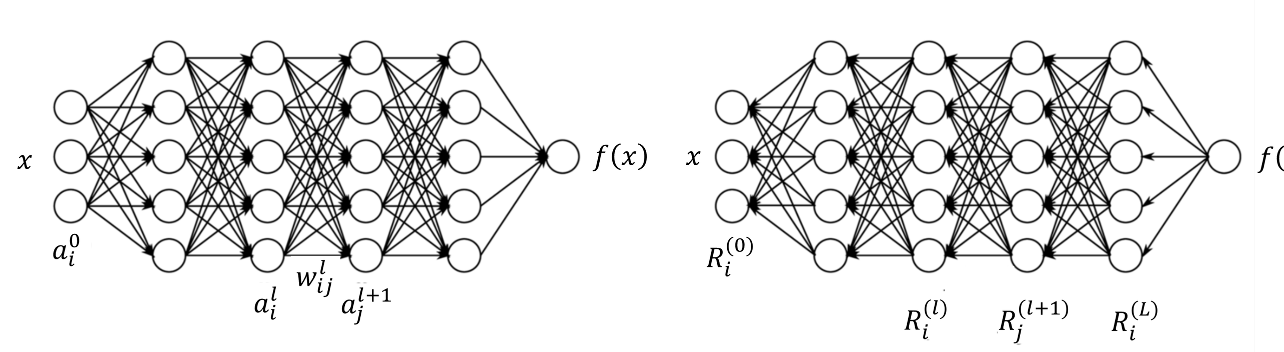

Figure 2. Multilayer neural network annotated with the different variables describing weight connections and activation vectors. Left: forward pass. Right: backward pass.

LRP Attack Formulation

- We propose an **explainable** adversarial attack for **Coarse-to-Fine** classifiers by using LRP to guide perturbation toward the most relevant features.
- Our algorithm is designed to craft perturbations that specifically **disrupt the DNN's attention** and alter its decision-making process at both **Coarse** and **Fine level** attacks.

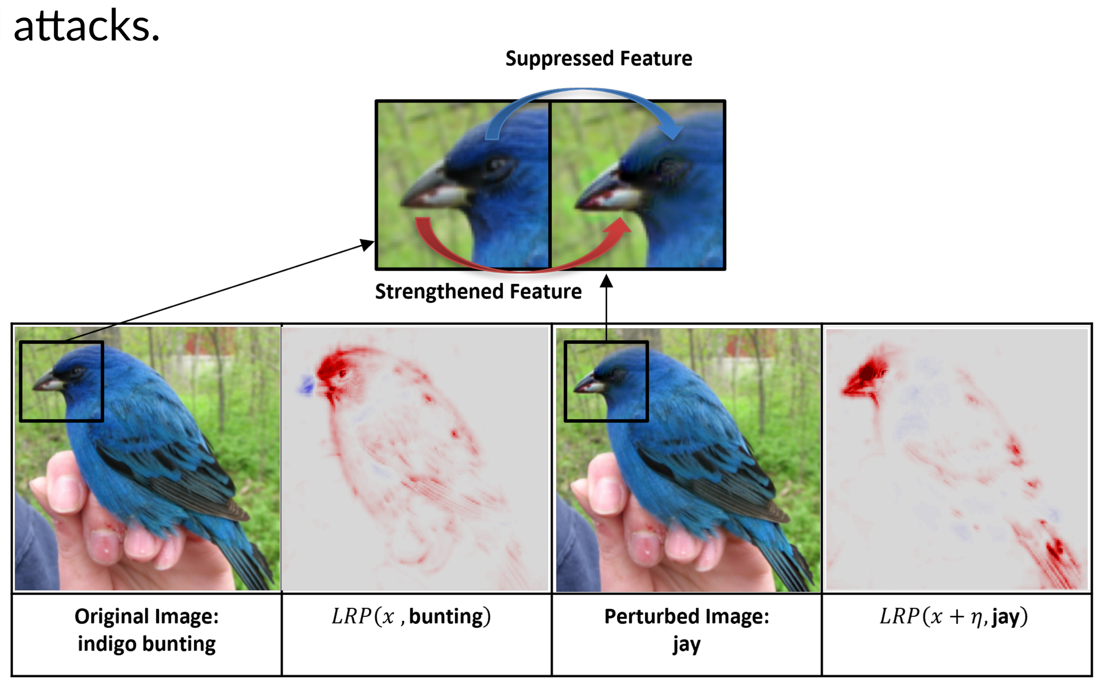

Figure 3. Strengthened and suppressed features alter classifier perception, highlighting the impact of explainable adversarial attacks (LRPF).

Fooling the Coarse Level

The goal is to

$$C(x+\eta) \neq C(x).$$

We define original and adversarial coarse labels as

$$r_{\text{org}} = C(x), r_{\text{adv}} = \underset{i\in[M]\setminus r_{\text{org}}}{\operatorname{argmax}}\ C_i(x)\,.$$

To redirect the coarse classifier's attention from r_{org} to r_{adv}, the loss function for the LRP Coarse-level attack (LRPC) is defined as:

$$\mathcal{L}_C = \|LRP_C(x+\eta; r_{\text{org}})^+\|_p - \|LRP_C(x+\eta; r_{\text{adv}})^+\|_p \\ - \|LRP_C(x+\eta; r_{\text{org}})^-\|_p + \|LRP_C(x+\eta; r_{\text{adv}})^-\|_p\,.$$

Fooling the Fine Level

The goal is to

$$F^{r_{\text{org}}}(x+\eta) \neq F^{r_{\text{org}}}(x), \text{ while } C(x+\eta) = C(x).$$

We define original and adversarial fine labels as

$$f_{\text{org}} := F^{r_{\text{org}}}(x), f_{\text{adv}} = \underset{j\in[M_{r_{\text{org}}}]\setminus f_{\text{org}}}{\operatorname{argmax}}\ F_j^{r_{\text{org}}}(x)\,.$$

Then, we define a loss function for the LRP Fine-level attack (LRPF):

$$\mathcal{L}_F = \|LRP_{F^{r_{\text{org}}}}(x+\eta; f_{\text{org}})^+\|_p - \|LRP_{F^{r_{\text{org}}}}(x+\eta; f_{\text{adv}})^+\|_p \\ - \|LRP_{F^{r_{\text{org}}}}(x+\eta; f_{\text{org}})^-\|_p + \|LRP_{F^{r_{\text{org}}}}(x+\eta; f_{\text{adv}})^-\|_p.$$

Experimental Setup

- **Dataset:** 393 out of 1,000 ImageNet (ILSVRC2012) classes selected for the C2F classifier; 80% for **training**, 20% for **validation**; evaluated on **VGG-16**.
- **C2F framework:** We use a C2F classifier with $M = 8$ coarse categories: **{fish, bird, reptile, clothes, food, vehicle, electrical device, dog}**, which are further classified by separate **fine-level** classifiers.

Results

Explainability-Perceptibility Tradeoff

Our attack outperforms traditional methods in providing clearer interpretation without compromising attack imperceptibility.

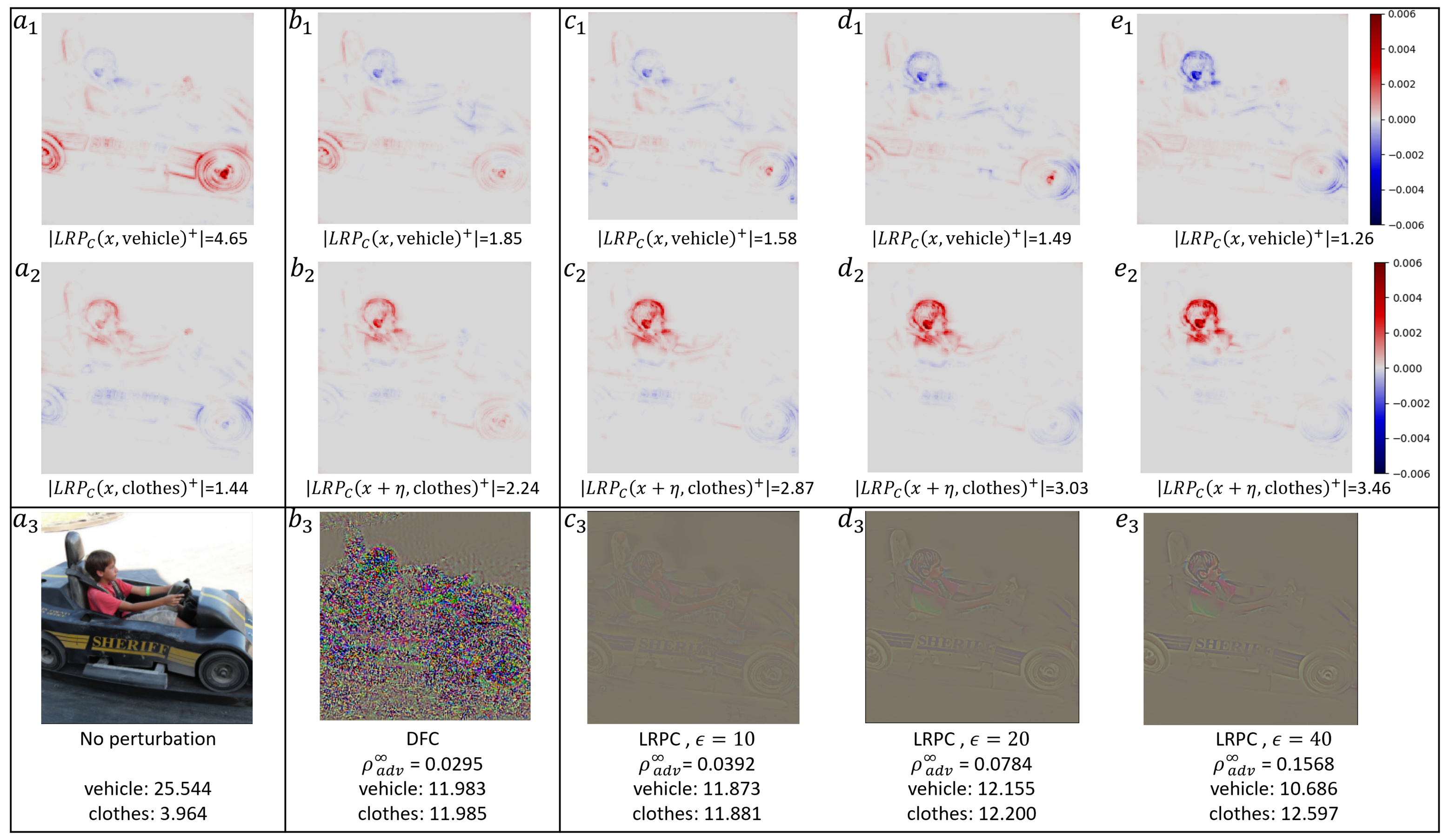

Figure 4. LRP visualizations before and after LRPC and DFC attacks. (a_1) LRP of the original coarse class and (a_2) adversarial coarse class before the attack. (a_3) Benign image. (c_1, d_1, e_1) LRP of r_{org} after LRPC attack for $\epsilon = 10, 20, 40$, compared to (b_1) for DFC. (c_2, d_2, e_2) LRP of r_{adv} after LRPC attack for $\epsilon = 10, 20, 40$, compared to (b_2) for DFC. Perturbations generated with LRPC ($\epsilon = 10, 20, 40$) are shown in (c_3, d_3, e_3), and for DFC in (b_3).

Performance Evaluation

- **Evaluation Metrics:** The average **perceptibility** of the attack:

$$\rho^p_{\text{adv}}(f) = \frac{1}{|D|} \sum_{x\in D} \frac{\|\eta\|_p}{\|x\|_p}.$$

- The **fooling ratio**, defined as the proportion of images whose labels are changed by the attack relative to the total number of images.

Table 1. Fooling ratio and perceptibility of coarse-level attacks.

Algorithm	LRPC $\epsilon = 10$	LRPC $\epsilon = 20$	LRPC $\epsilon = 40$	DFC	PGDC
ρ^2_{adv}	0.0294	0.0323	0.0405	0.0045	0.0262
ρ^1_{adv}	0.0216	0.0174	0.0195	0.0031	0.0224
ρ^∞_{adv}	0.0399	0.0778	0.1557	0.0408	0.0101
Fooling(%)	87.1	92.5	99.3	100	100

Table 2. Fooling ratio and perceptibility of fine-level attacks.

Algorithm	LRPF $\epsilon = 10$	LRPF $\epsilon = 20$	LRPF $\epsilon = 40$	DFF	PGDF
ρ^2_{adv}	0.0127	0.0145	0.0151	0.0020	0.0078
ρ^1_{adv}	0.0084	0.0079	0.0066	0.0013	0.0092
ρ^∞_{adv}	0.0241	0.0542	0.0819	0.0029	0.0035
Fooling(%)	98.7	100	100	100	95.7

- Both **LRPC** and **LRPF** achieve high fooling rates while improving **explainability**.
- Our attack **prioritizes** explainability over perceptibility, while still achieving competitive fooling rates with **controlled** perturbation levels.

References

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[4] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[5] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

This work was supported by NSF under Awards CCF-2106339 and DMS-2304489, and DARPA under Agreement No. HR0011-24-9-0427.