Overview

Motivation: A better understanding of the decision-making process of a CNN is required to provide hints for improving it. This allows us to uncover and understand failure cases, limits of the model, and shortcomings of the training data.

Fine-grained visual explanation method (FGVis):

- Peacock
- Model
- Target class
- CNN
- FGVis highlights in detail the evidence on which a model bases its decisions
- Contributions:
  - A method (FGVis) to generate fine-grained explanations in the image space.
  - A novel technique for defending against adversarial evidence, which does not depend on human-tuned parameters.
- Interpretable and class discriminative explanations, visualizing detailed evidence.

Defending against adversarial evidence

Drawback of perturbation based methods: Adversarial evidence, i.e., faulty evidence due to artifacts introduced in the optimization of the explanation.

Deflection against adversarial evidence:
- Idea: The features in an explanation should be subset of the image features.
- Corresponding optimization constraint:
  - \( 0 \leq \mathbf{h}_t(e_t) \leq \mathbf{h}_{t}(x) \), if \( \frac{dh_t(x)}{dx} > 0 \); otherwise:
  - \( \mathbf{h}_t(e_t) \): Activation of the \( t \)-th neuron in the \( t \)-th layer.
  - The constraint is applied after each nonlinearity-layer (e.g.: ReLU-Layer).
- Implemented via gradient clipping:

\[
\mathbf{e}_t' = \mathbf{e}_t - \lambda \cdot \left[ \frac{\partial \mathbf{h}_t(e_t)}{\partial \mathbf{e}_t} - \mathbf{h}_t(x) \right],
\]

Novel adversarial defense:
- Our defense prevents the hallucination of adversarial evidence.
- Our defense does not depend on human-tuned parameters and enables an explanation which is both fine-grained and preserves the characteristics of the image.

3 Perturbation based explanation methods

An explanation \( e^* \) is computed by perturbing the input image \( x \):

\[
\text{Mask } \mathbf{m}^*_b \text{ based perturbation: } e^*_b = x \cdot \mathbf{m}^*_b. \quad ^c \cdot \text{ Target class of the explanation.}
\]

Preserving explanation:

\[
\mathbf{x} = \arg \min_{\mathbf{x}} \mathbf{m} \cdot \mathbf{x} - \text{CNN}(\mathbf{x}) e^*_b.
\]

Deleting explanation:

\[
\mathbf{x} = \arg \max_{\mathbf{x}} (\mathbf{1} - \mathbf{m}) \cdot \mathbf{x} - \text{CNN}(\mathbf{x}) e^*_b.
\]

FGVis generates the most fine-grained explanation mask

4 Experiments

Qualitative comparison with other methods

- Backpropagation based methods [1,2,3]
- Activation based method [4]
- Perturbation based methods [5,6, ours]

FGVis generates the most fine-grained explanation mask

5 Experiments

Class discriminative / fine-grained

- Soccer ball
- Norwegian elkhound
- Chair-frame
- Tennis ball
- School bus

- FGVis produces discriminative explanations even when objects partially overlap
- Drawback of perturbation based methods:
  - Novelty technique to defend against adversarial evidence
  - Method (FGVis) to generate fine-grained explanations in the image space.
  - A novel technique for defending against adversarial evidence, which does not depend on human-tuned parameters.
- Interpretable and class discriminative explanations, visualizing detailed evidence.

6 References