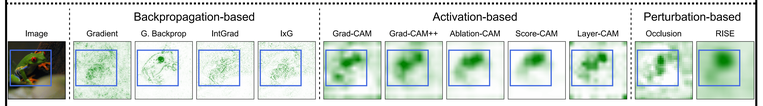
Towards Better Understanding Attribution Methods

Sukrut Rao, Moritz Böhle, Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus

Objective and Background

Perform a more systematic, fair, and holistic evaluation of attribution methods

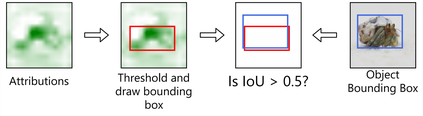


Attribution Methods

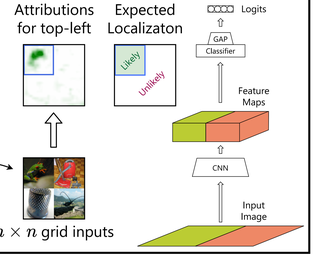
- Post-hoc attribution methods attempt to explain image classification networks
- Model is a black box ⇒ no ground truth attributions, challenging to evaluate

Evaluating Attributions: Common Metrics

Weakly Supervised Object Localization



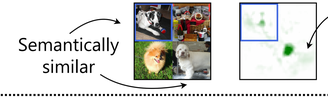
Grid Pointing Game (GridPG)



Localization Metric:

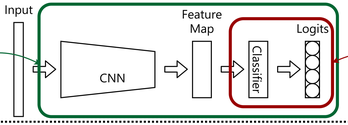
Shortcomings of Existing Evaluation Metrics

① Challenging to evaluate model-faithfulness



Region outside cell can influence model ⇒ unknown if such attributions erroneous

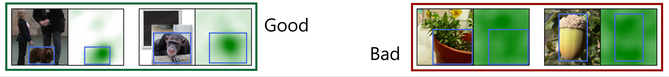
② Comparing methods that explain models at varied depths unfair



Explain full network, e.g. IxG

Explain only the classifier, e.g. Grad-CAM

③ Qualitative evaluation sensitive to the selection of examples used

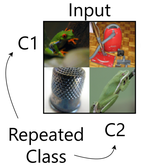


Contributions

① Full Disconnection (DiFull)

Disentangle model from attribution method

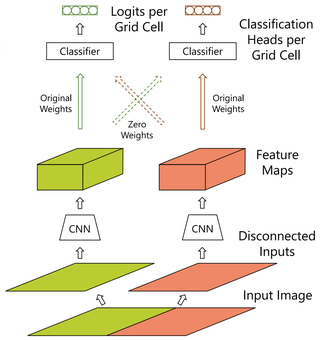
Separate classification head per grid cell, each only influenced by pixels inside cell



Classification head for C1 cannot be influenced by C2

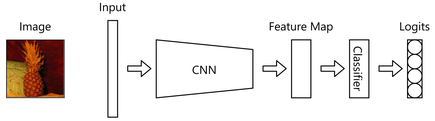
Localization Guarantee

Analogous for other cells

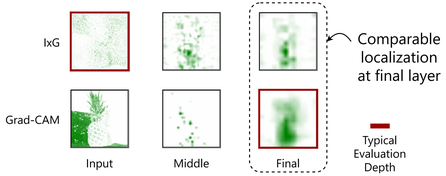


② Multi-Layer Attribution Evaluation (ML-Att)

Evaluate and compare at identical model depths



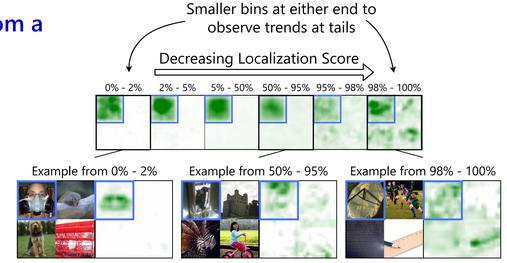
Evaluation at several depths, with focus on three — Input, Middle, Final — to understand trends in performance



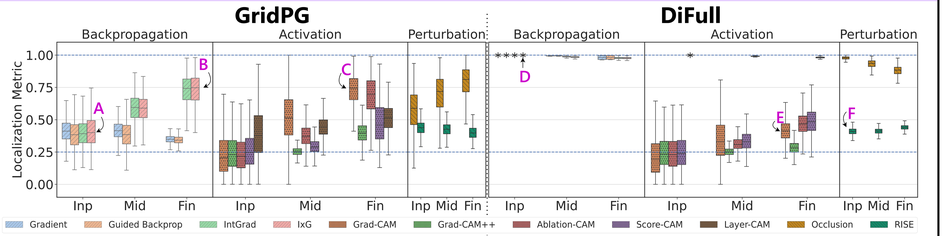
③ Aggregate Attribution Evaluation (AggAtt)

Aggregate attributions from a large number of images

1. Sort by localization score
2. Bin attribution maps
3. Aggregate and visualize mean attributions per bin



Results



① DiFull identifies strengths and failures of attribution methods, acts as sanity check

- Backpropagation-based methods localize perfectly (D), activation-based methods localize poorly (E)

② Localization improves at deeper layers on GridPG, many methods perform similarly

- Methods like IntGrad and IxG (A, B) perform similar to best-performing methods like Grad-CAM (C)

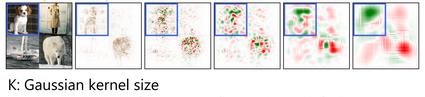
③ AggAtt visualizes performance across dataset, helps understand quantitative results

- Visualizations corroborate quantitative results and serve as a holistic visual summary (A-F)
- Unlike individual examples, shows a range of performance across different images (e.g. F)
- Shows that activation-based methods perform poorly on DiFull by incorrectly attributing repeated class (E)

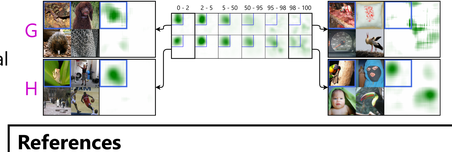
Smoothing Attributions

Post-hoc Gaussian smoothing of attributions can improve GridPG localization

- IxG (G) and IntGrad (H) at input layer perform comparably to Grad-CAM at final layer (C)
- Smoothing can show the net contribution of a local region to the logit for piece-wise linear models



K: Gaussian kernel size
Green: Positive attributions, Red: Negative attributions



References

- Gradient (Simonyan et al., 2014)
- G. Backprop (Springenberg et al., 2015)
- IntGrad (Sundararajan et al., 2017)
- IxG (Shrikumar et al., 2017)
- Grad-CAM (Selvaraju et al., 2017)
- Grad-CAM++ (Chattopadhyay et al., 2018)
- Ablation-CAM (Desai et al., 2020)
- Score-CAM (Wang et al., 2020)
- Layer-CAM (Jiang et al., 2021)
- Occlusion (Zeiler et al., 2014)
- RISE (Petsiuk et al., 2018)
- Weakly Supervised Object Localization (Cao et al., 2015)
- GridPG (Böhle et al., 2021)