Analysis of Feature Attribution Methods in Explainable AI

Ayça Avcı (s4505972)

Principal project advisor: Prof. Dr. Ir. G. GAYDADJIEV

Second project advisor: R. BRANDT, M.Sc.

Contributions

- Giving an overview of feature attribution methods for image classification.
- Comparing and evaluating five state-of-the-art feature attribution methods using a proper evaluation framework.

Feature Attribution Methods

Selected Methods

- Vanilla Gradients
- SmoothGrad
- XRAI
- Grad-CAM
- BlurIG

Selection Criteria

- Recently developed,
- Outperforms commonly used methods (e.g. Integrated Gradients and LIME),
- Gradient-based image classification methods (faster to compute).

Evaluation Metrics

Selected Metrics

- MoRF (Deletion),
- Smoothness,
- Invariance to data labeling,
- Similarity to edge map,
- Runtime,
- Invariance to model weights.
    - Cascading Randomization
    - Independent Randomization

Selection Criteria

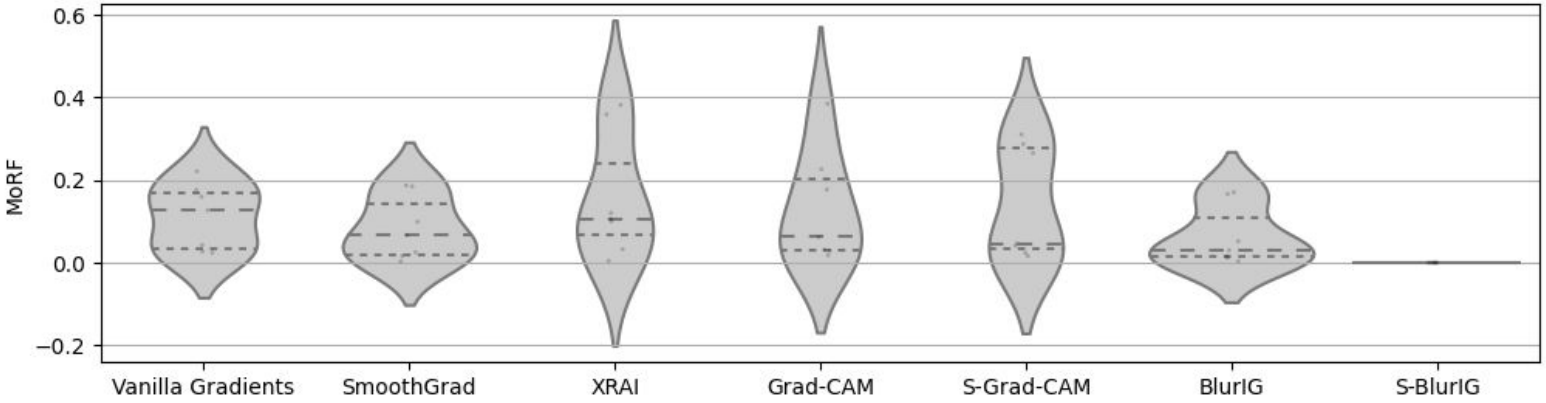
- Established metrics,
- Can be computed efficiently,
- Implementing the metrics is doable.

Experimental Design

- **Inception-v3** convolutional neural network trained on **ImageNet**
  - **82.8%** accuracy on top-1 prediction
- ImageNet validation set as a test set
  - Contains **1000 categories** (classes), including **over 3000 images**
- Evaluation metrics
  - MoRF (Deletion)
  - Smoothness
  - Similarity to edge map
  - Invariance to data labeling
  - Runtime
  - Invariance to model weights
    - Cascading Randomization
    - Independent Randomization

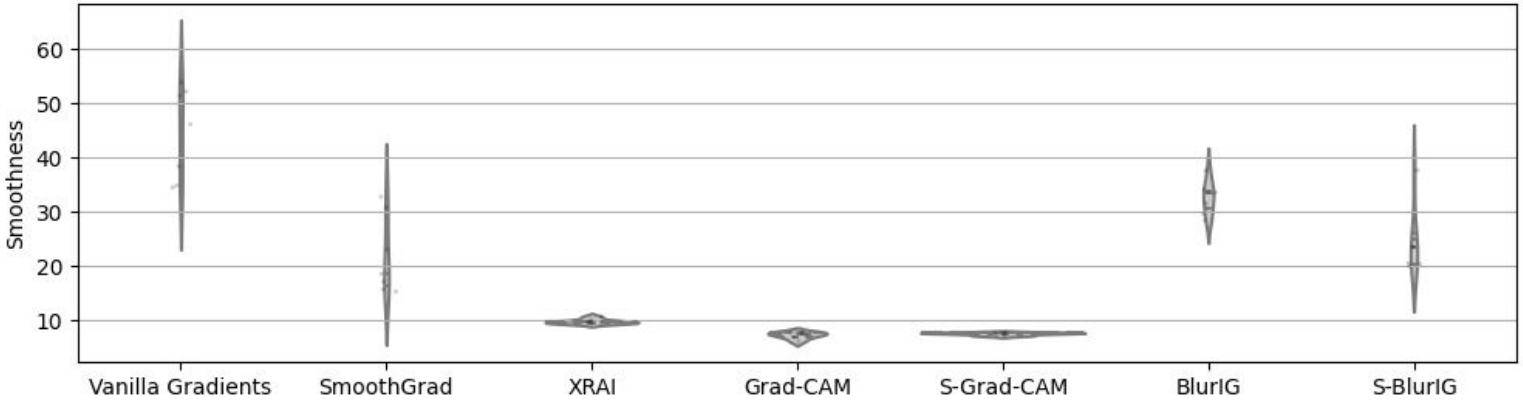
Quantitative Results

MoRF (Deletion)



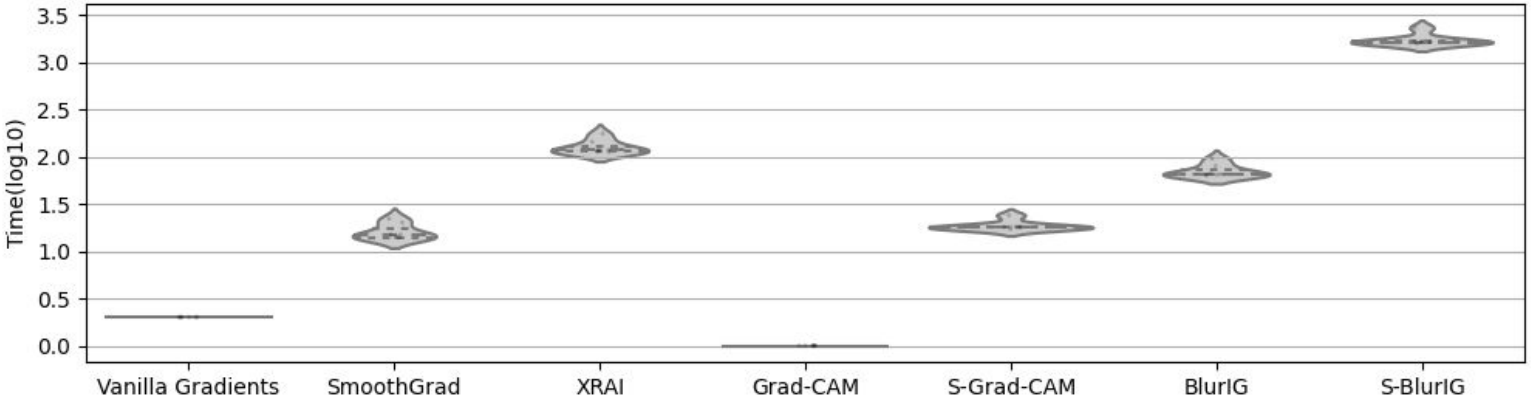
Quantitative Results

Smoothness



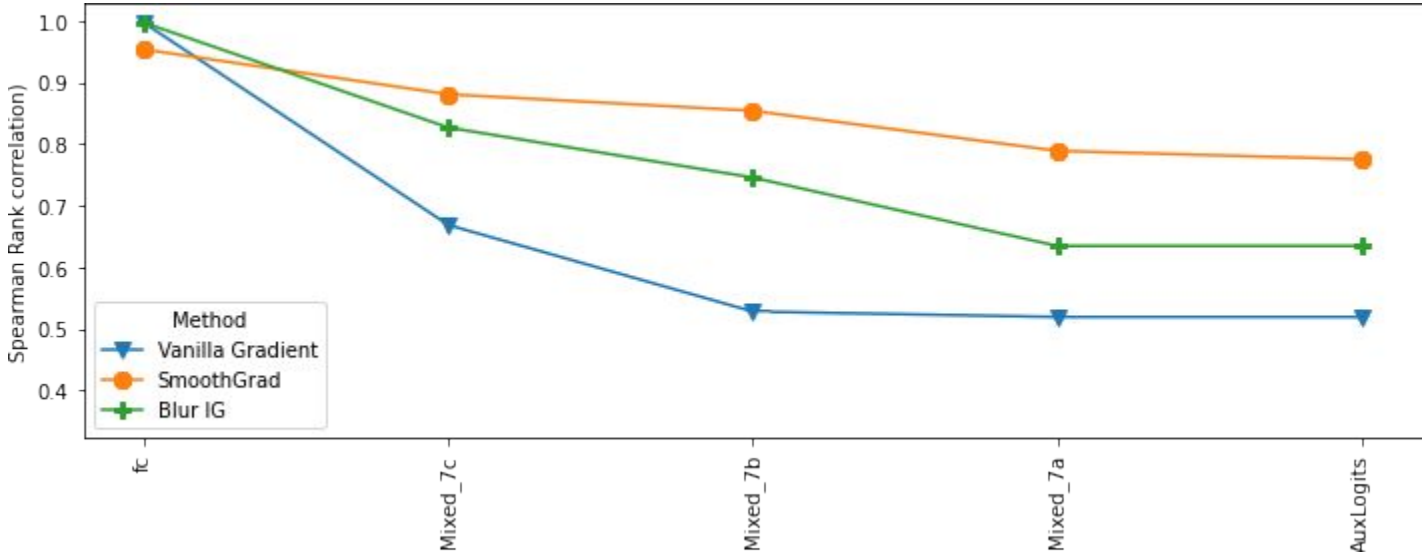
Quantitative Results

Time (log10 scale)



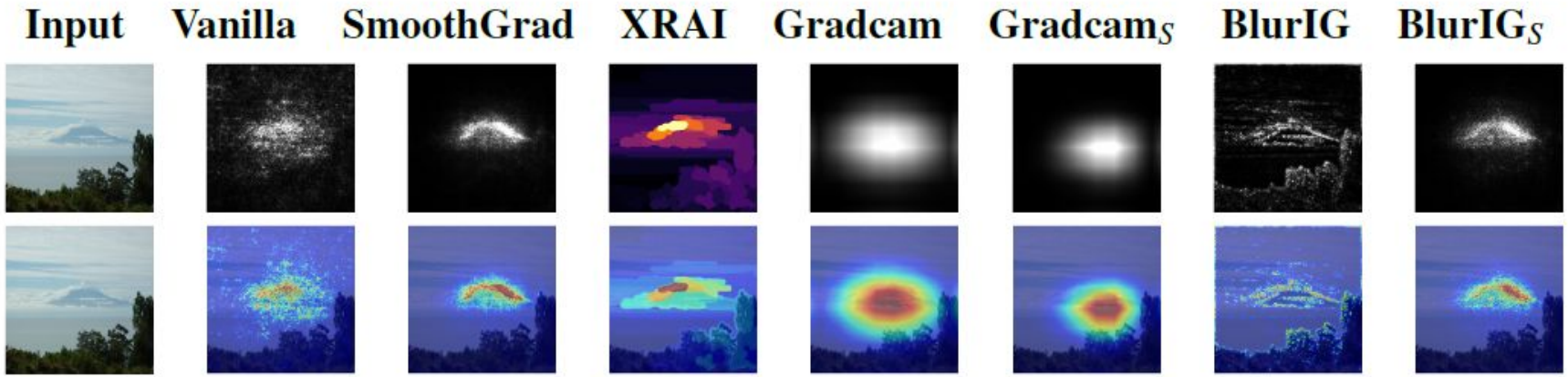
Quantitative Results

Invariance to Model Weights - Independent Randomization



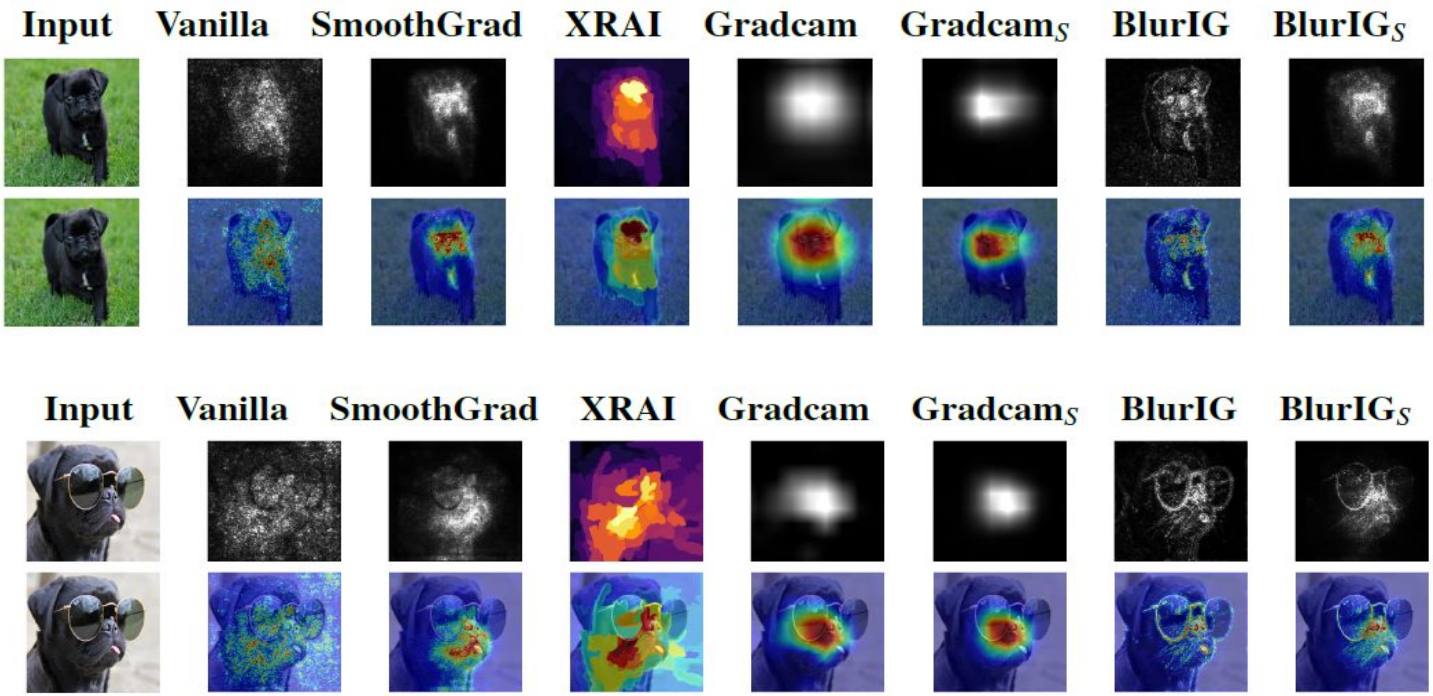
Qualitative Results

Saliency Map Examples - Volcano



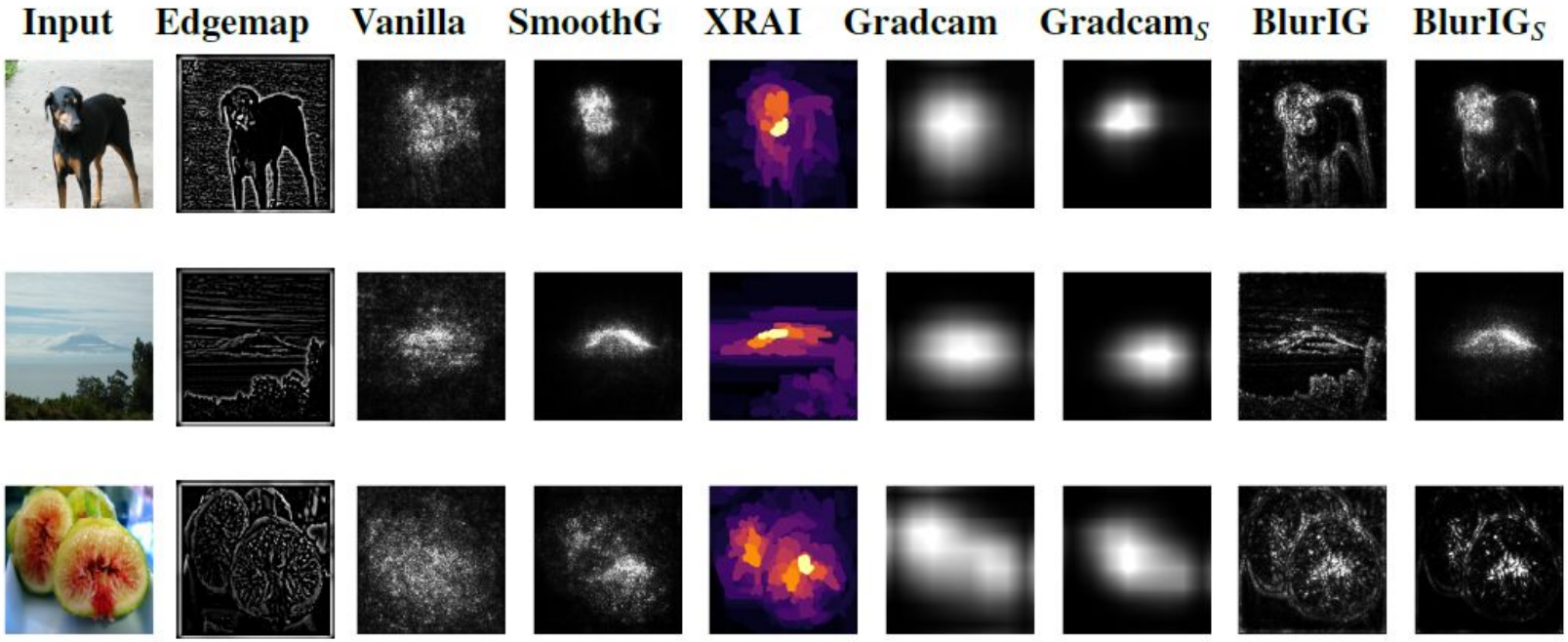
Qualitative Results

Saliency Map Examples - Pug & Pug with Glasses



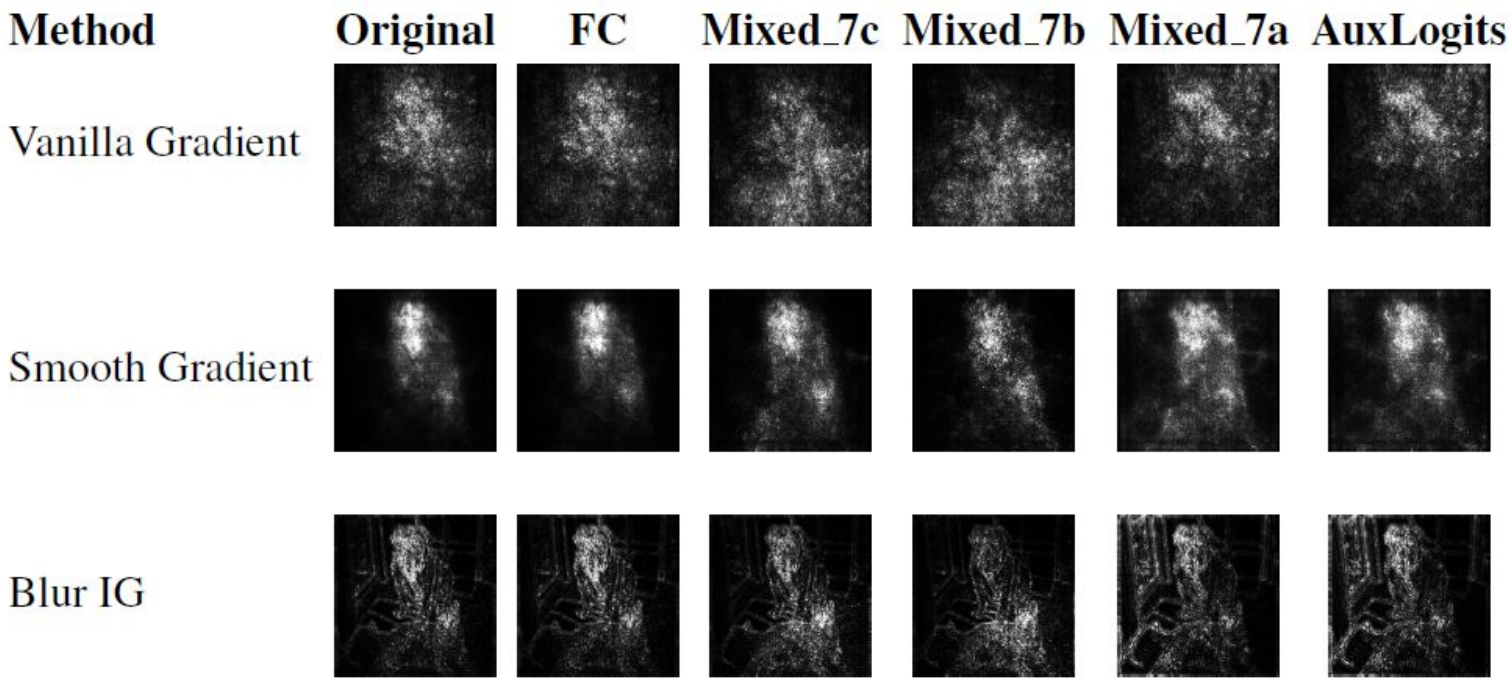
Qualitative Results

Similarity to Edge Map



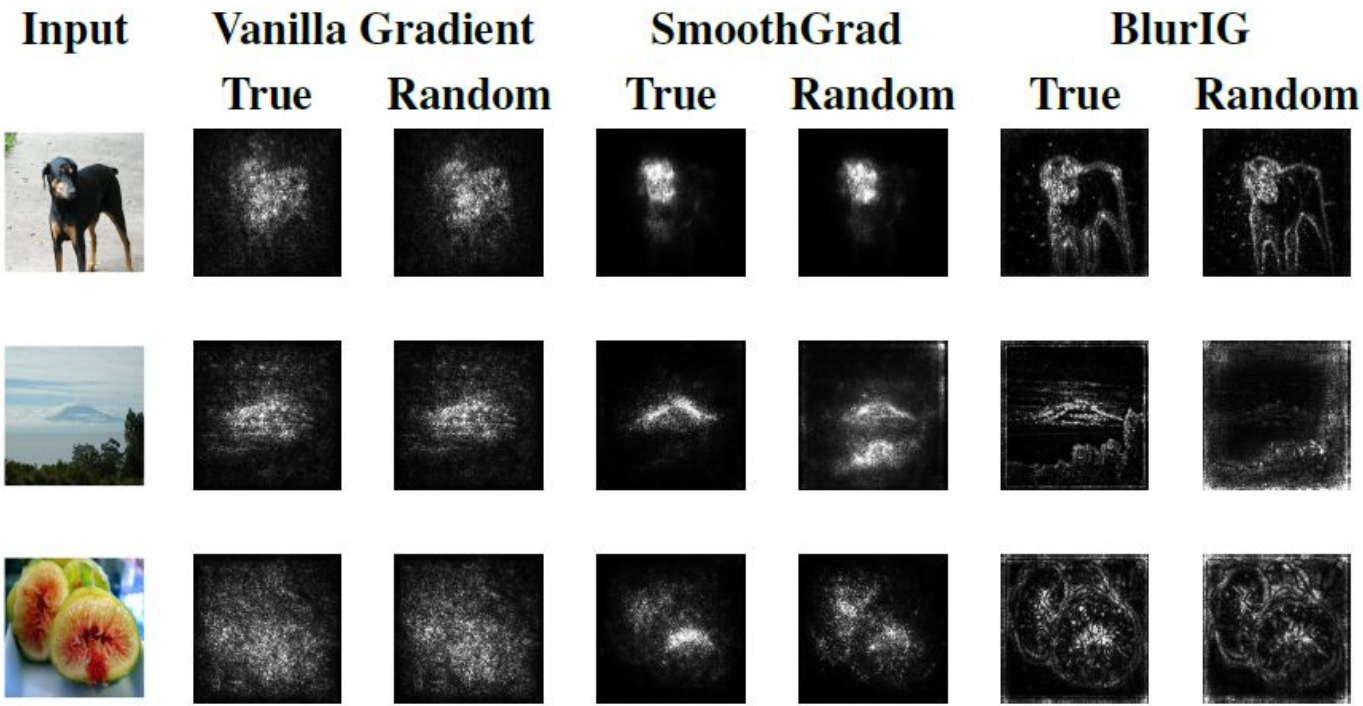
Qualitative Results

Invariance to Model Weights - Independent Randomization



Qualitative Results

Invariance to Data Labeling



Quantitative Results

Method	MoRF/Deletion	Smoothness	Similarity to Edge Map	Invariance to data	Runtime
Vanilla Gradient	0.111 ± 0.074	41.510 ± 7.814	0.10	0.86	2
SmoothGrad	0.083 ± 0.072	20.559 ± 6.698	0.15	0.79	15.286
XRAI	0.158 ± 0.140	9.776 ± 0.396	-	-	118
Grad-CAM	0.133 ± 0.127	7.370 ± 0.558	0.13	-	1
Grad-CAM$_S$	0.142 ± 0.127	7.599 ± 0.216	0.13	-	18.714
BlurIG	0.064 ± 0.067	31.687 ± 2.836	0.25	0.43	66.429
BlurIG$_S$	-	24.553 ± 5.727	0.17	-	1728.857

Conclusion

According to

- MoRF -> BlurIG
- Smoothness -> Grad-CAM
- Invariance to data labeling -> Vanilla Gradient
- Similarity to edge map -> Vanilla Gradient
- Runtime -> Vanilla Gradient & Grad-CAM
- Invariance to model weights (Independent Randomization) -> Vanilla Gradient

provides **better** explanation.

Future Work

- Running experiments for more images.
- Using additional metrics to evaluate methods.

Thank You!