Ranking Explainable AI Methods Using Pixel-Level Evidence in Classification Tasks

Anh Khoa Pham – vibe_coding_scientist

LUH|AI

Leibniz Universität Hannover

1 Motivation

- Introduce XAI methods for Deep Learning
- Quantitatively compare Grad-CAM, Saliency Maps, and Integrated Gradients using a Top-K overlap metric
- Identify the most suitable and faithful explanation method for image classification tasks with pixel-level ground truth

2 XAI Methods

- Saliency Maps: $SM_c^{(i,j)}(x) = \max_{k\in\{1,\ldots,C\}} \left|\frac{\partial f_c(x)}{\partial x_{j,j,k}}\right|$
- Integrated Gradients:

$$IG_C(x) = (x - x') \odot \int_{\alpha}^{1} \frac{\partial f_C(x' + \alpha(x - x'))}{\partial x} d\alpha$$

- Grad-CAM:

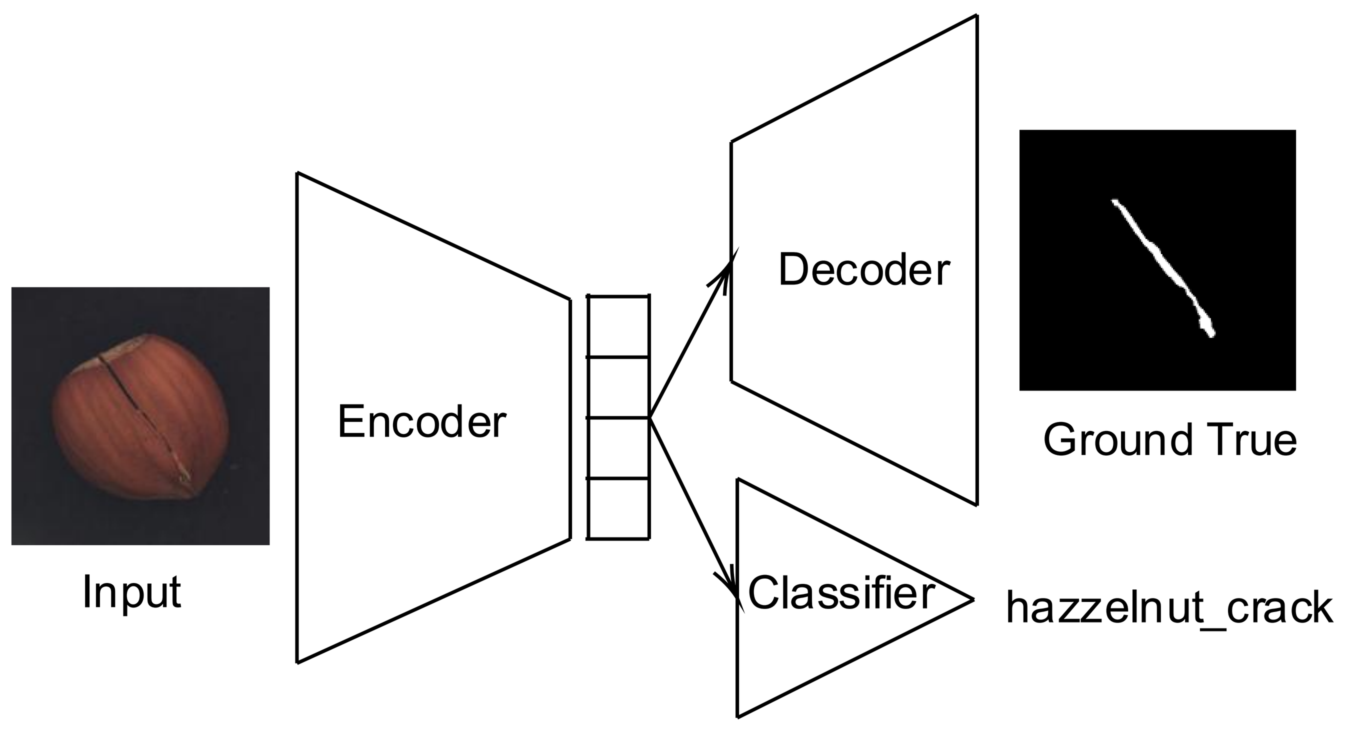
$$\alpha_k^C = \frac{1}{z}\sum_{\mathrm{i}}\sum_{\mathrm{j}} \frac{\partial f_C}{\partial A_{ij}^k}$$

$$GC_C(x) = Re\,LU\left(\sum \alpha_k^C A^k\right)$$

3 Dataset

- Subset of the MVTec AD dataset, focusing exclusively on hazelnut images
- Each sample includes an input image, a pixel-level ground-truth annotation, and an image-level defect label
- Five classes are considered: crack, cut, good, hole, and print
- Primary task: image classification

4 Training Pipeline



- Multi-task learning loss:

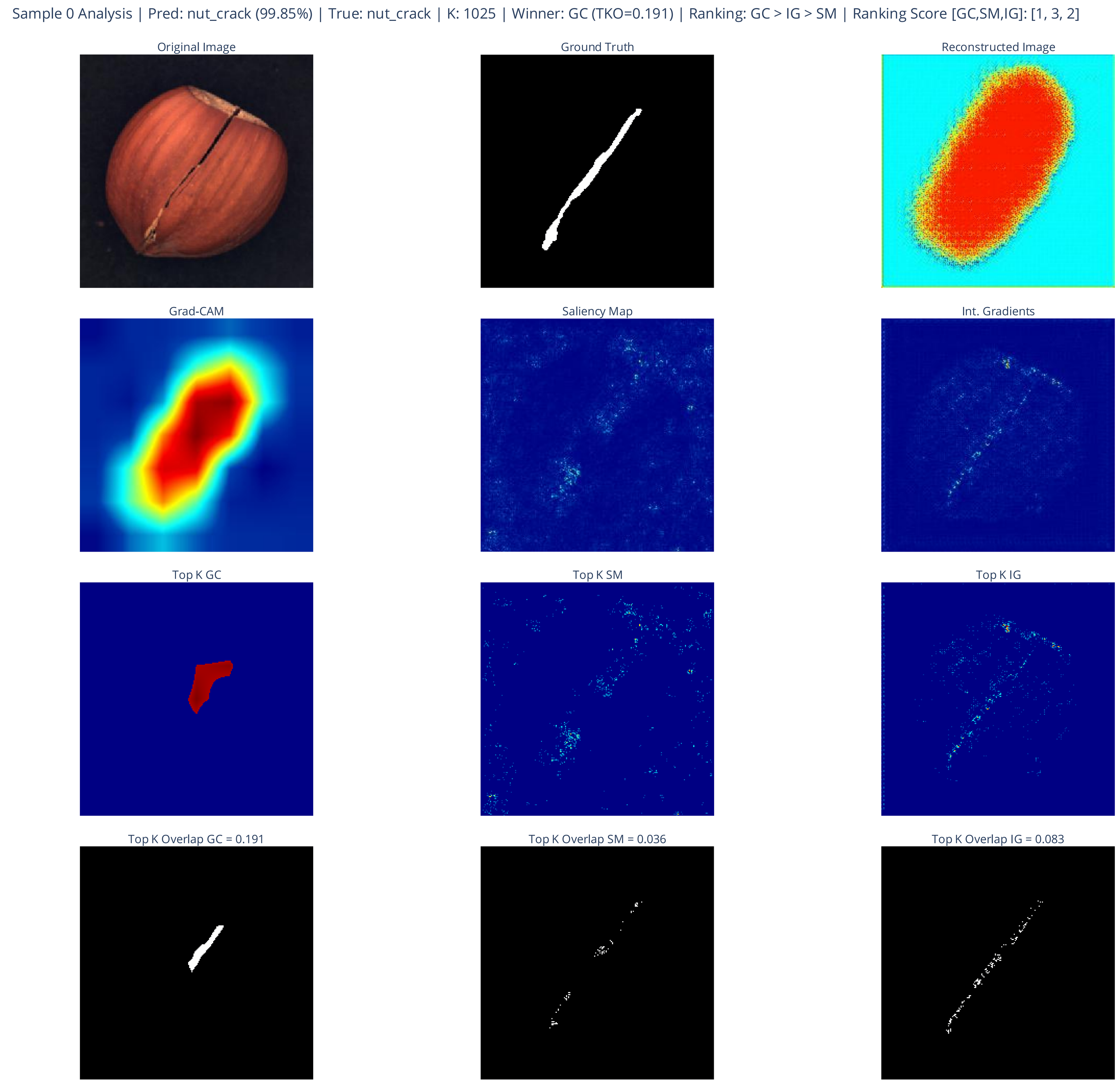
$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(\hat{y}, y) + \lambda\mathcal{L}_{BCE}(\hat{y}_{res}, y_{gt})$$

- Performance of the model averaged 10 seeds

Split	Loss	Loss CE	Loss BCE	Accuracy
Train	0.0631±0.0530	0.0581±0.0527	0.0050±0.0007	0.9837±0.0144
Test	0.2087±0.1745	0.2006±0.1736	0.0081±0.0014	0.9443±0.0467

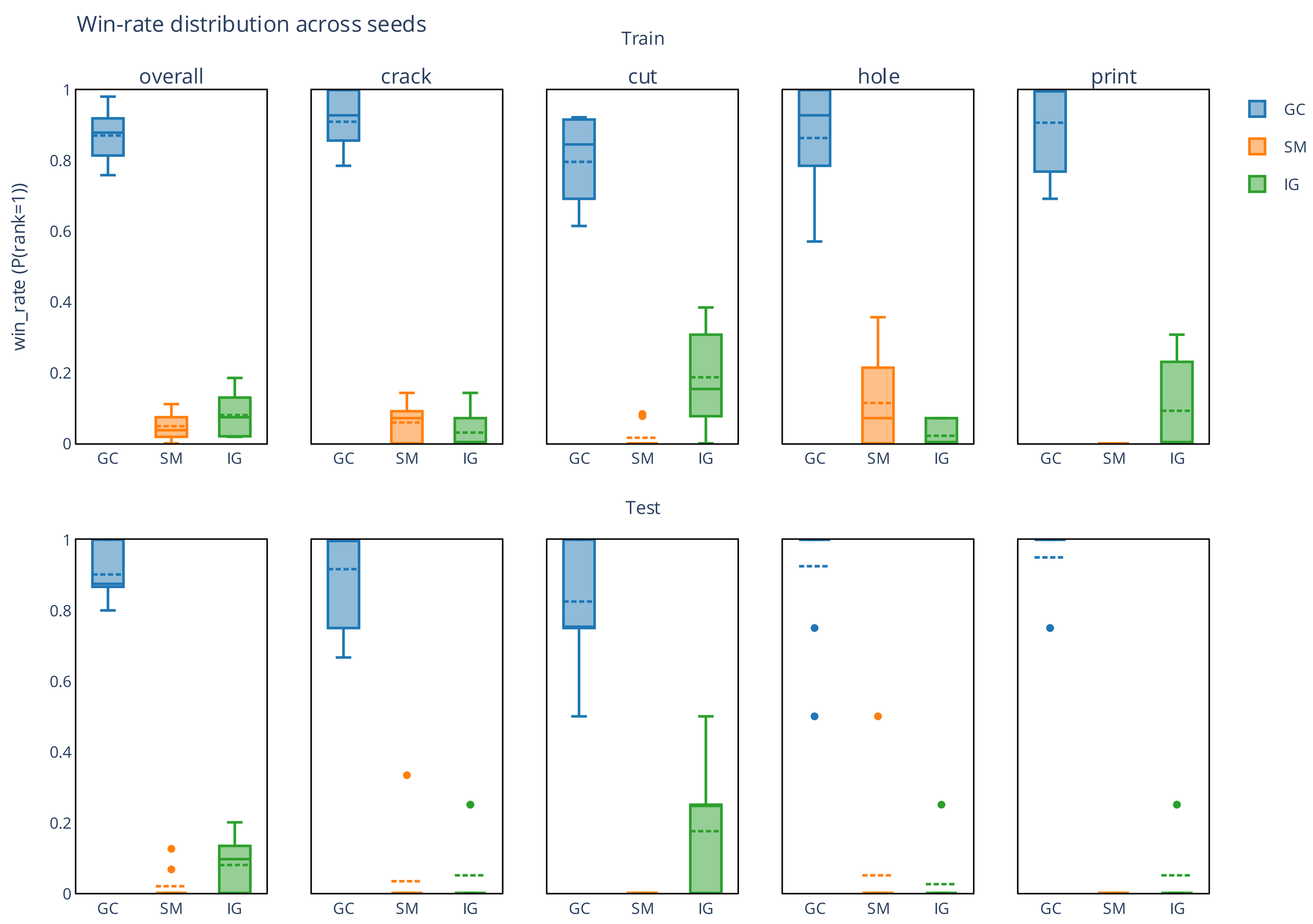
5 Top-K Overlap Evaluation

- Set K as the number of positive pixels in the ground-truth mask
- Select the Top-K highest-scoring pixels from each explanation map
- Compute the overlap between selected pixels and the ground-truth region
- Higher overlap indicates more faithful explanations



6 Win-Rate

- Win-rate of each XAI method on the train and test splits, computed over the entire dataset and per class.



- Win-rate of XAI methods averaged 10 seeds

Split	Class	GC	SM	IG
Train	Overall	**0.8713**±0.0678	0.0486±0.0382	0.0801±0.0564
	Crack	**0.9104**±0.0740	0.0591±0.0543	0.0305±0.0495
	Cut	**0.7968**±0.1119	0.0160±0.0321	0.1872±0.1181
	Hole	**0.8643**±0.1445	0.1143±0.1204	0.0214± 0.0327
	Print	**0.9077**±0.1278	0.0000±0.0000	0.0923±0.1278
Test	Overall	**0.9017**±0.0739	0.0192±0.0405	0.0792± 0.0712
	Crack	**0.9167**±0.1291	0.0333±0.1000	0.0500± 0.1000
	Cut	**0.8250**±0.1601	0.0000±0.0000	0.1750±0.1601
	Hole	**0.9250**±0.1601	0.0500±0.1500	0.0250±0.0750
	Print	**0.9500**±0.1000	0.0000±0.0000	0.0500±0.1000

7 Discussion

- **Conclusion:**
  - GC consistently performs best across all classes and the full dataset.
  - GC produces smoother, less noisy explanations for CNN-based image classification by focusing on high-level semantic information.
  - Selecting the appropriate XAI method is task- and data-dependent.
- **Limitations**
  - GC is limited to CNN-based image models.
- **Future Work**
  - Apply GC to video-based tasks, such as object detection.