# AGOP as Explanation: From Feature Learning to Per-Sample Attribution in Image Classifiers

Raj Kiran Gupta Katakam[1,*]

[1]Credit Karma, San Francisco, CA, USA

**Abstract**

The Average Gradient Outer Product (AGOP) governs feature learning in neural networks: the Neural Feature Ansatz states that weight Gram matrices at each layer align with the corresponding AGOP matrices computed over the training distribution. We ask a complementary question: can this same quantity serve as a *post-hoc attribution method* for explaining individual predictions? We introduce *AGOP-Weighted*: a novel attribution method that multiplies the per-sample gradient by $\sqrt{\mathrm{diag}(M)/\max \mathrm{diag}(M)}$, a training-distribution prior that suppresses gradient noise and amplifies consistently important pixels—a combination not present in any prior attribution method. We formalise two companion variants—AGOP-Local (per-sample gradient, equivalent to VanillaGrad) and AGOP-Global ($\mathrm{diag}(M)$ directly as a zero-cost saliency map)—and implement an efficient training-time accumulation hook; AGOP-Global then requires zero inference cost (disk lookup) while AGOP-Weighted requires only a single gradient pass. We conduct the first rigorous comparison of AGOP attribution against Integrated Gradients (IG), SmoothGrad, GradCAM, and VanillaGrad across two benchmarks with pixel-level ground truth: (i) the synthetic XAI-TRIS benchmark (four classification scenarios, 8×8 images, CNN8by8) and (ii) the photorealistic CLEVR-XAI benchmark (ResNet-18 fine-tuned from ImageNet). AGOP-Weighted achieves 44% higher mIoU than IG on linear tasks; AGOP-Global achieves 7× higher mIoU than IG on multiplicative tasks (where IG falls *below* random) at zero inference cost. Both findings generalise to ResNet-18 on CLEVR-XAI ($+18\%$ and $+37\%$ respectively). We further show that GradCAM fails on small-resolution images due to spatial resolution collapse, and that $\mathrm{diag}(M)$ quality improves monotonically throughout training even after classification accuracy has plateaued.

**Keywords**

explainability, attribution, AGOP, gradient-based saliency, image classification, neural feature ansatz



## 1. Introduction

### 1.1. Motivation

Explaining what a neural network "looks at" when making a prediction has become a central challenge in trustworthy machine learning. In image classification, attribution methods assign a relevance score to each input pixel, producing a *saliency map* that highlights which regions of the image drove the model's decision. The field has produced a rich family of such methods: Integrated Gradients [1] uses a path integral from a baseline to the input; GradCAM [2] projects class-discriminative gradients back onto convolutional feature maps; and SmoothGrad [3] averages gradients over noise-perturbed copies of the input.

These methods share a common limitation: they treat the trained model as a black box, applying geometric constructs independently of how the network learned its representations.

### 1.2. A Different Starting Point: AGOP

Recent theoretical work has revealed that the *Average Gradient Outer Product* (AGOP) is not merely one post-hoc tool among many—it is the mathematical object that governs feature learning in neural networks. Radhakrishnan et al. [4] proved that the weight Gram matrices $W_l^\top W_l$ at each layer align with the corresponding AGOP matrices computed over the training distribution. This is the *Neural Feature*

*Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil*

*Corresponding author.

raj.katakam@creditkarma.com (R. K. G. Katakam)

*Ansatz* (NFA): the geometry of the network's weights is encoded in its gradient statistics. Beaglehole et al. [5] extended this to convolutional networks, showing that patch-based AGOP recovers the filter covariance.

The diagonal of the AGOP matrix, $\mathrm{diag}(M)$, has a natural attribution interpretation: it is the average squared partial derivative of the model output with respect to each input dimension, accumulated over the training set. Input dimensions that the model consistently relies upon receive high attribution under $\mathrm{diag}(M)$—precisely what a good global saliency map should capture.

### 1.3. Contributions

1. *AGOP-Weighted: a novel attribution method* that multiplies the per-sample gradient by a training-distribution prior (normalised RMS gradient $w$), bridging per-sample and global attribution for the first time. While AGOP is an established theoretical construct [4], its application as a pixel-level attribution method for image classifiers—and specifically the AGOP-Weighted formulation—is introduced here. Companion variants: AGOP-Local (equivalent to VanillaGrad, situating it within the AGOP framework) and AGOP-Global ($\mathrm{diag}(M)$ as a zero-cost dataset-level saliency prior, not a per-instance explanation).
2. *AGOPTrainingHook*: an implementation that accumulates $\mathrm{diag}(M)$ during standard training using per-sample `autograd.grad` backward passes; AGOP-Global then operates at zero inference cost (disk lookup of $\mathrm{diag}(M)$) while AGOP-Weighted requires a single gradient pass per input.
3. *First rigorous evaluation on XAI-TRIS* [6] across four classification scenarios and two background conditions, using five localization and faithfulness metrics.
4. *Empirical findings*: AGOP-Weighted achieves the best mIoU on linear tasks ($+44\%$ over IG); AGOP-Global achieves $7\times$ IG's mIoU on multiplicative tasks; GradCAM fails on small-resolution images.
5. *Generalisation to CLEVR-XAI* [7] with ResNet-18 fine-tuned from ImageNet, confirming all core findings.

## 2. Related Work

**Gradient-based attribution.** Throughout, $f_c(x)$ denotes the logit for class $c$. VanillaGrad [8] computes $|\partial f_c/\partial x|$ with a single backward pass. Integrated Gradients [1] addresses gradient saturation by integrating along a path from a reference baseline to the input, satisfying *completeness* and *sensitivity* axioms. SmoothGrad [3] reduces gradient noise by averaging over $n$ noise-perturbed copies of the input. GradCAM [2] and GradCAM++ [9] weight convolutional feature maps by gradient-based channel importance and upsample to input resolution.

**AGOP, NFA, and recursive feature machines.** Radhakrishnan et al. [4] proved the NFA for fully-connected networks trained with gradient descent. Beaglehole et al. [5] extended the result to convolutional networks via patch-based AGOP. The xRFM framework [10] uses the diagonal of the AGOP matrix for coordinate-level feature importance in tabular settings but does not compare against pixel-level XAI baselines on image benchmarks. Our work fills this gap.

**Perturbation-based and global methods.** LIME [11] and SHAP [12] treat the model as a black box, fitting surrogate models or computing Shapley values on perturbed samples; both are complementary to gradient-based methods but more expensive at inference. Most attribution methods are *local*—producing a separate map per input; AGOP-Global is *global*, acting as a dataset-level diagnostic of learned features, not a per-instance explanation.

**XAI benchmarks.** XAI-TRIS [6] provides 8×8 synthetic images with deterministic pixel masks across four task types of increasing non-linearity. CLEVR-XAI [7] provides photorealistic 3D scenes with per-question ground-truth object masks. OpenXAI [13] provides a tabular XAI evaluation framework; Samek et al. [14] establish the deletion/insertion faithfulness protocol we adopt.

# 3. Attribution Methods

We evaluate five standard baselines and three AGOP variants, plus a random attribution baseline for reference. All methods produce a non-negative saliency map $A \in \mathbb{R}^{H\times W}$ for a given input image $x \in \mathbb{R}^{C\times H\times W}$ and predicted class $c$.

## 3.1. Standard Baselines

**VanillaGrad [8].** $\text{attr}_{\text{VG}}(h,w) = \sum_{ch} |\partial f_{\hat{c}}(x)/\partial x[ch,h,w]|$, where $ch$ indexes colour channels and $\hat{c}$ is the predicted class. Computed with a single backward pass (1× cost).

**Integrated Gradients (IG) [1].** $\text{IG}_j(x) = (x_j - x'_j)\cdot \frac{1}{T}\sum_{k=1}^{T} \frac{\partial f_c(x' + \frac{k}{T}(x-x'))}{\partial x_j}$, where $j$ indexes input dimensions, $x' = \mathbf{0}$ (zero baseline, standard practice; baseline sensitivity left to future work), $T = 50$ steps (50× cost). IG can fail on multiplicative interactions where individual pixel gradients are small even at signal pixels.

**SmoothGrad [3].** Averages $\nabla_x f_c(x+\varepsilon_k)$ over $K = 50$ noise-perturbed inputs with $\varepsilon_k \sim \mathcal{N}(0, \sigma^2 I)$, $\sigma = 0.15$ (50× cost).

**GradCAM [2] / GradCAM++ [9].** Weight convolutional feature maps by gradient-based channel importance and bilinearly upsample to input resolution (1× cost). Spatial resolution is hard-capped at the target layer's feature map size.

## 3.2. AGOP Variants

Let $\{(x^{(s)}, y^{(s)})\}_{s=1}^{n}$ be the training set with predicted classes $\hat{y}^{(s)}$. The AGOP matrix and its diagonal (where $j$ indexes input dimensions) are:

$$M = \frac{1}{n}\sum_{s=1}^{n} \nabla_x f_{\hat{y}^{(s)}}(x^{(s)})\, \nabla_x f_{\hat{y}^{(s)}}(x^{(s)})^{\top}, \quad \text{diag}(M)[j] = \frac{1}{n}\sum_{s=1}^{n}\left(\frac{\partial f_{\hat{y}^{(s)}}(x^{(s)})}{\partial x_j}\right)^2$$

$\text{diag}(M)$ is tractable (64 floats for $8\times 8$ images; ≈600 KB for $224\times 224\times 3$).

**AGOP-Local.** Equivalent to VanillaGrad: $\text{attr}_{\text{L}}(h,w) = \sum_{ch} |\partial f_{\hat{c}}(x)/\partial x[ch,h,w]|$. Its role is conceptual: it anchors the Local→Weighted→Global progression by showing that VanillaGrad is simply AGOP evaluated on a single sample, with AGOP-Weighted and AGOP-Global incorporating progressively more training-set statistics (1× cost). The exact single-sample AGOP diagonal is $g_j^2$; we use $|g_j|$ to match standard saliency implementations—both produce the same pixel ordering so PG, Del, and Ins are identical, though mIoU and Energy-GT may differ in principle.

**AGOP-Weighted.** Uses the training-accumulated $\text{diag}(M)$ as a prior to modulate the per-sample gradient: With weight $v[i] = \sqrt{\text{diag}(M)[i]/\max_j \text{diag}(M)[j]}$ (defined above):

$$\text{attr}_{\text{W}}(h,w) = \sum_{ch}\left|\frac{\partial f_{\hat{c}}(x)}{\partial x[ch,h,w]}\right| \cdot v[ch,h,w].$$

The weight $v[i]$ suppresses noisy gradient spikes in favour of consistently important pixels. *Inference cost*: 1× (diag pre-loaded from disk).

**Algorithm 1** AGOP Diagonal Accumulation During Training

*Input:* model $f$, training dataloader $\mathcal{D}$, flag `only_correct`
$\mathrm{diag}(M) \leftarrow \mathbf{0} \in \mathbb{R}^d$; $n_{\text{acc}} \leftarrow 0$
**for** each minibatch $(X, Y)$ in $\mathcal{D}$ **do**
  $\hat{Y} \leftarrow \arg\max f(X)$ ▷ forward pass
  **for** each sample $s$ in batch **do**
    **if** `only_correct` and $\hat{y}^{(s)} \neq y^{(s)}$ **then** *continue*
    **end if**
    $g^{(s)} \leftarrow \nabla_x f_{\hat{y}^{(s)}}(x^{(s)})$ ▷ per-sample backward via `autograd.grad`
    $\mathrm{diag}(M)$ += $(g^{(s)})^{\odot 2}$; $n_{\text{acc}}$ += 1
  **end for**
  `optimizer.step()` ▷ AGOP hook does NOT alter gradients
**end for**
$\mathrm{diag}(M) \leftarrow \mathrm{diag}(M)/n_{\text{acc}}$ ▷ normalise
*Save* $\mathrm{diag}(M)$ to disk

**AGOP-Global.** $\text{attr}_G = \mathrm{diag}(M)$ (same for every input). Powerful when signal regions are positionally consistent across inputs; fails for spatially invariant tasks where $\mathrm{diag}(M)$ averages to a diffuse map. *Inference cost*: $0\times$ (disk lookup).

## 4. The AGOPTrainingHook

The hook performs one additional per-sample backward pass (separate from the weight-update backward), and does not alter training. Overhead scales with batch size and can be reduced via `torch.vmap` for architectures without normalization layers. $\mathrm{diag}(M)$ is computed once: AGOP-Global then costs $0\times$ at inference (disk lookup); AGOP-Weighted costs $1\times$ (one gradient pass).

Key design choices: (1) `only_correct=True` accumulates $\mathrm{diag}(M)$ only over correctly-classified samples, preventing early-training errors from contaminating gradient statistics; (2) mean normalisation produces interpretable units (expected squared gradient per dimension); (3) snapshots every 100 gradient steps enable convergence analysis.

## 5. Experiments

### 5.1. XAI-TRIS Benchmark

XAI-TRIS [6] provides 8×8 grayscale images where a 4-pixel tetromino shape (T or L) serves as the class-discriminative signal, with a deterministic GT mask marking exactly those pixels. We evaluate four scenarios of increasing attribution difficulty: *Linear* — signal pixels have higher mean intensity (additive mixing, $\alpha = 0.18$); gradients align with the signal, making this the easiest case. *Multiplicative* — signal pixels *modulate* background values (pixel-wise multiplication); per-sample gradients at signal pixels are proportional to random background noise, making first-order methods unreliable. *Translations+Rotations* — the tetromino appears at a random position and orientation per sample; global statistics are diffuse, but per-sample methods must track each image's specific shape. *XOR* — class label is the XOR of binarised tetromino values; gradients cancel in expectation ($\mathbb{E}[\partial f/\partial a] \approx 0$), defeating all first-order methods. Each scenario uses two background conditions: *uncorrelated* (clean signal) and *correlated* (spurious class-specific background, producing ≈50% model accuracy as a sanity check).

*Model (CNN8by8)*: Four Conv2d+ReLU+MaxPool blocks, 234 parameters, trained with Adam (lr $= 10^{-3}$, WD$= 10^{-4}$), cosine annealing, 100 epochs; accuracy 80–94% (uncorrelated), ≈50% (correlated). *Metrics*: *PG* — does the peak pixel fall inside GT mask?; *mIoU* — IoU of binarised map vs GT; *Energy-GT* — attribution mass inside GT (random ≈6.25%); *Del*↓/*Ins*↑ — faithfulness via progressive masking/revealing.

**Table 1**
Scenario 1—Linear / uncorrelated. Model acc. ≈80%. Underline denotes best value per column.

| Method | PG | mIoU | Energy-GT | Del↓ | Ins↑ | ms/s. |
|---|---|---|---|---|---|---|
| Random | 0.093 | 0.059 | 0.062 | 0.560 | 0.563 | <1 |
| VanillaGrad | 0.667 | 0.201 | 0.300 | 0.466 | 0.635 | 0.8 |
| IG | 0.697 | 0.218 | 0.365 | 0.463 | 0.636 | 2.4 |
| SmoothGrad | 0.579 | 0.181 | 0.226 | 0.466 | 0.635 | 1.2 |
| GradCAM | 0.000 | 0.113 | 0.097 | 0.442 | 0.638 | 1.0 |
| GradCAM++ | 0.000 | 0.109 | 0.096 | 0.471 | 0.614 | 1.1 |
| AGOP-Local | 0.667 | 0.201 | 0.300 | 0.466 | 0.635 | 1.0 |
| AGOP-Weighted | 0.642 | 0.314 | 0.456 | 0.457 | 0.637 | 1.0 |
| AGOP-Global | 0.491 | 0.235 | 0.308 | 0.456 | 0.634 | 0 |

**Table 2**
Scenario 2—Multiplicative / uncorrelated.

| Method | PG | mIoU | Energy-GT | Del↓ | Ins↑ | ms/s. |
|---|---|---|---|---|---|---|
| Random | 0.067 | 0.060 | 0.063 | 0.547 | 0.585 | <1 |
| VanillaGrad | 0.304 | 0.090 | 0.167 | 0.462 | 0.634 | 1.3 |
| IG | 0.100 | 0.048 | 0.062 | 0.460 | 0.636 | 3.3 |
| SmoothGrad | 0.446 | 0.165 | 0.222 | 0.454 | 0.630 | 1.2 |
| GradCAM | 0.000 | 0.014 | 0.023 | 0.504 | 0.594 | 1.0 |
| GradCAM++ | 0.000 | 0.059 | 0.051 | 0.534 | 0.576 | 1.0 |
| AGOP-Local | 0.304 | 0.090 | 0.167 | 0.462 | 0.634 | 1.0 |
| AGOP-Weighted | 0.377 | 0.204 | 0.286 | 0.454 | 0.631 | 1.0 |
| AGOP-Global | 0.506 | 0.337 | 0.331 | 0.453 | 0.621 | 0 |

**Table 3**
Scenario 3—Translations+Rotations / uncorrelated.

| Method | PG | mIoU | Energy-GT | Del↓ | Ins↑ | ms/s. |
|---|---|---|---|---|---|---|
| Random | 0.052 | 0.061 | 0.063 | 0.502 | 0.519 | <1 |
| VanillaGrad | 0.774 | 0.225 | 0.348 | 0.455 | 0.644 | 1.3 |
| IG | 0.973 | 0.540 | 0.784 | 0.455 | 0.647 | 3.1 |
| SmoothGrad | 0.764 | 0.242 | 0.336 | 0.455 | 0.647 | 1.2 |
| GradCAM | 0.140 | 0.098 | 0.101 | 0.465 | 0.608 | 1.0 |
| GradCAM++ | 0.167 | 0.111 | 0.116 | 0.468 | 0.614 | 1.1 |
| AGOP-Local | 0.774 | 0.225 | 0.348 | 0.455 | 0.644 | 1.0 |
| AGOP-Weighted | 0.767 | 0.233 | 0.357 | 0.455 | 0.644 | 1.0 |
| AGOP-Global | 0.098 | 0.080 | 0.074 | 0.488 | 0.556 | 0 |

**Linear (Table 1).** AGOP-Weighted achieves the best mIoU (0.314) and Energy-GT (0.456), outperforming IG by $+44\%$ in mIoU. GradCAM completely fails (PG $= 0$) due to spatial resolution collapse ($3 \times 3$ feature maps). AGOP-Local $\equiv$ VanillaGrad, confirming theoretical equivalence.

**Multiplicative (Table 2).** AGOP-Global dominates: mIoU $= 0.337$ vs. IG's 0.048—a $7\times$ margin. IG falls *below* the random baseline (0.048 vs. 0.060): its path-integral traverses a noisy gradient landscape from pixel-wise multiplication and actively anti-localises the signal. AGOP-Global's training-accumulated $\mathrm{diag}(M)$ averages over all samples, cancelling per-sample noise and recovering the stable signal.

**Translations+Rotations (Table 3).** IG is the clear winner (PG $= 0.973$, mIoU$= 0.540$): its path-integral correctly tracks each image's specific shape location. AGOP-Global completely fails (mIoU $=$ 0.080) because $\mathrm{diag}(M)$ averages over all positions, producing a diffuse map.

**Table 4**
Scenario 4—XOR / uncorrelated. Negative mIoU: background receives more attribution than signal.

| Method | PG | mIoU | Energy-GT | Del↓ | Ins↑ | ms/s. |
|---|---|---|---|---|---|---|
| Random | 0.048 | −0.008 | 0.001 | 0.550 | 0.569 | <1 |
| VanillaGrad | 0.208 | −0.038 | −0.019 | 0.474 | 0.623 | 0.8 |
| IG | 0.202 | −0.142 | −0.158 | 0.459 | 0.638 | 2.4 |
| SmoothGrad | 0.326 | −0.025 | −0.004 | 0.474 | 0.630 | 1.2 |
| GradCAM | 0.000 | −0.017 | 0.002 | 0.496 | 0.523 | 1.0 |
| GradCAM++ | 0.000 | 0.007 | 0.013 | 0.498 | 0.542 | 1.0 |
| AGOP-Local | 0.208 | −0.038 | −0.019 | 0.474 | 0.623 | 0.9 |
| AGOP-Weighted | 0.349 | −0.080 | −0.026 | 0.458 | 0.638 | 1.1 |
| AGOP-Global | 0.497 | −0.062 | 0.002 | 0.455 | 0.650 | 0 |

**XOR (Table 4).** All methods struggle (no mIoU above $0.01$; IG reaches $-0.142$): XOR requires second-order methods. *Correlated background*: all methods are near-random at ≈50% accuracy—a model that cannot classify cannot explain.

**Cross-scenario summary.** AGOP-Weighted dominates when signal is positionally consistent; IG dominates with high local spatial contrast; all methods fail on XOR. *No single method wins across all scenarios.* AGOP-Weighted achieves 0.314 mIoU at 1.0 ms/sample vs. IG's 0.218 at 2.4 ms/sample—44% better at half the cost. AGOP-Global is a free disk lookup, ideal for batch pipelines. Attribution quality improves monotonically throughout training: $\mathrm{diag}(M)$ continues to improve through epoch 100 even after accuracy saturates at epochs 70–85.

## 5.2. CLEVR-XAI Benchmark

CLEVR-XAI [7] contains rendered 3D scenes with 3–10 geometric objects of varying colours, sizes, and materials (metal or rubber). Each scene is paired with a question about a specific object's material, so the *same image* can carry different labels depending on which object is queried—a naive classifier trained on the raw dataset achieves only ≈58% accuracy (barely above chance). We resolve this by restricting to scenes where one material *dominates*: ≥60% of objects share the queried material. The task becomes "does metal or rubber dominate this scene?"—a *modified proxy task* (not the original VQA grounding problem), but a signal learnable from the image alone. This yields 4,952 samples (3,962 train / 990 val), accuracy ceiling ≈91.5%. We use `all_objects` GT masks (≈16.2% of pixels) rather than `single_object` masks, since the model must attend to *all* same-material objects; evaluating against a single-object mask (≈2.3%) structurally undercounts correct attributions by up to $N\times$ when $N$ objects of the dominant material are present.

*Model*: ResNet-18 [15] (ImageNet pretrained, final FC replaced with Linear(512, 2)), trained with Adam (lr $= 10^{-4}$), CosineAnnealingLR, 50 epochs; best val_acc $= 91.52\%$ at epoch 31. $\mathrm{diag}(M)$ computed via post-hoc full pass over training samples at best checkpoint.

Both AGOP-Global (mIoU = 0.149) and AGOP-Weighted (0.129) exceed IG (0.109) by $+37\%$ and $+18\%$ respectively. IG leads on Pointing Game (0.290) but under-covers the multi-object mask. AGOP-Global achieves competitive mIoU at 0.02 ms/sample—a $6{,}600\times$ speedup over IG (131.98 ms/sample). GradCAM's spatial collapse is confirmed at ResNet-18 scale (`layer4`: $2 \times 2$ map, PG = 0.002 despite 16.2% mask coverage). AGOP-Local ≡ VanillaGrad confirmed across architectures.

Notably, the ranking between AGOP-Global and AGOP-Weighted *inverts* compared to XAI-TRIS Linear (where Weighted 0.314 > Global 0.235). The reason is the same Global–Local trade-off: the CLEVR-XAI task (material dominance) has a positionally consistent signal—dominant-material objects tend to occupy similar scene regions across training images—so $\mathrm{diag}(M)$ alone captures the distribution-level pattern directly. On XAI-TRIS Linear, the signal is at a *fixed pixel position*, making the per-sample gradient always informative; on CLEVR-XAI, slight positional variability within the dominant material

**Table 5**
CLEVR-XAI, ResNet-18, val_acc = 91.52%, `all_objects` masks, $n$=500 val samples.

| Method | PG | mIoU | Energy-GT | Del↓ | Ins↑ | ms/s. |
|---|---|---|---|---|---|---|
| Random | 0.158 | 0.130 | 0.154 | 0.479 | 0.503 | 0.07 |
| VanillaGrad | 0.282 | 0.124 | 0.162 | 0.479 | 0.503 | 11.04 |
| IG | <u>0.290</u> | 0.109 | 0.151 | <u>0.479</u> | 0.503 | 131.98 |
| SmoothGrad | 0.108 | <u>0.179</u> | <u>0.169</u> | 0.482 | <u>0.505</u> | 103.50 |
| GradCAM | 0.002 | 0.129 | 0.151 | 0.548 | 0.497 | 15.72 |
| GradCAM++ | 0.004 | 0.128 | 0.149 | 0.539 | 0.492 | 19.92 |
| AGOP-Local | 0.282 | 0.124 | 0.162 | 0.479 | 0.503 | 20.33 |
| AGOP-Weighted | 0.204 | 0.129 | 0.166 | 0.480 | 0.504 | 18.65 |
| AGOP-Global | 0.000 | 0.149 | 0.165 | 0.479 | 0.504 | <u>0.02</u> |

region means the global prior alone is more reliable than the Weighted combination.

## 6. Discussion

**Why AGOP-Weighted outperforms IG on area-overlap metrics.** IG makes a strong assumption: attribution computed relative to a fixed black baseline. The training-accumulated $\mathrm{diag}(M)$ in AGOP-Weighted makes no baseline assumption—it accumulates gradient statistics over the *actual* training distribution—producing better area-coverage (mIoU, Energy-GT) even when IG wins on Pointing Game. On multiplicative tasks, the *average* squared gradient concentrates on tetromino pixels even though individual evaluations are noisy; the global average "sees through" the noise, explaining AGOP-Global's dominance at zero cost.

**GradCAM resolution collapse.** For CNN8by8 ($3 \times 3$ feature map, 9 regions, PG= $0.000$) and ResNet-18 at $64 \times 64$ ($2 \times 2$ map, 4 regions, PG= $0.002$), aggressive downsampling destroys spatial precision. This is a limitation at small resolutions, not a blanket failure; standard 224×224 networks have larger feature maps. *GradCAM should not be used when the target layer's feature map is coarser than the ground-truth mask.* The Global–Local trade-off follows directly: AGOP-Global excels when signal regions are positionally consistent; IG excels with high per-sample spatial contrast; AGOP-Weighted occupies the middle ground.

**Limitations.** We use only $\mathrm{diag}(M)$, discarding off-diagonal terms that may capture non-linear interactions. The `only_correct` flag and normalisation choices depart from the original Radhakrishnan et al. formulation. Both experiments use smaller models than production scale (ResNet-50, ViT-L at 224×224).

## 7. Conclusion

We have formalised AGOP as a post-hoc attribution method and provided the first systematic comparison against established baselines on two pixel-level benchmarks. Key findings: (1) AGOP-Weighted outperforms IG on area-overlap metrics ($+44\%$ mIoU on XAI-TRIS linear, $+18\%$ on CLEVR-XAI) at the same $1\times$ inference cost; (2) AGOP-Global outperforms all methods on multiplicative tasks ($7\times$ IG's mIoU where IG falls below random; $+37\%$ on CLEVR-XAI) at zero inference cost; (3) GradCAM fails entirely on small-resolution images due to spatial collapse; (4) IG excels on spatially coherent tasks; (5) $\mathrm{diag}(M)$ quality improves with training even after accuracy saturates; (6) all core findings generalise from XAI-TRIS to CLEVR-XAI. Future work should extend to larger networks, class-conditional AGOP variants, and off-diagonal interaction-based attribution.

## Acknowledgments

The author thanks the XAI-TRIS and CLEVR-XAI benchmark authors for making their datasets and code publicly available.

## Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: writing assistance and code review. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

## References

[1] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of ICML 2017, 2017, pp. 3319–3328.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of ICCV 2017, 2017, pp. 618–626.

[3] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, SmoothGrad: Removing noise by adding noise, in: ICML 2017 Workshop on Visualization for Deep Learning, 2017. ArXiv:1706.03825.

[4] A. Radhakrishnan, D. Beaglehole, P. Pandit, M. Belkin, Mechanism for feature learning in neural networks and backpropagation-free machine learning models, Science 383 (2024) 1461–1467.

[5] D. Beaglehole, A. Radhakrishnan, P. Pandit, M. Belkin, Mechanism of feature learning in convolutional neural networks, arXiv preprint arXiv:2309.00570 (2024).

[6] B. Clark, R. Wilming, S. Haufe, XAI-TRIS: Non-linear image benchmarks to quantify false positive post-hoc attribution of feature importance, Machine Learning 113 (2024) 6871–6910.

[7] L. Arras, A. Osman, W. Samek, CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations, Information Fusion 81 (2022) 14–40.

[8] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: ICLR 2014 Workshop on Learning Representations, 2014. ArXiv:1312.6034.

[9] A. Chattopadhyay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: Proceedings of WACV 2018, 2018, pp. 839–847.

[10] A. Radhakrishnan, et al., xRFM: Accurate, scalable, and interpretable feature learning models for tabular data, in: Workshop on AI for Time Series and Dynamic Data (AITD) at NeurIPS 2025, 2025. ArXiv:2508.10053.

[11] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of ACM KDD 2016, 2016, pp. 1135–1144.

[12] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017, pp. 4765–4774.

[13] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, OpenXAI: Towards a transparent evaluation of post hoc model explanations, in: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Curran Associates, Inc., 2022.

[14] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Transactions on Neural Networks and Learning Systems 28 (2017) 2660–2673.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of CVPR 2016, 2016, pp. 770–778.