Quantifying Explainability with Multi-Scale Gaussian Mixture Models

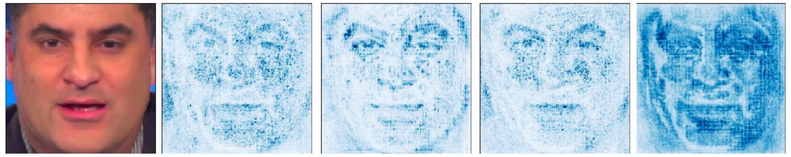
Anthony Rhodes, Intel Labs
Yali Bian, Intel Labs
Ilke Demir, Intel Labs

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivation

❖ *Which represents the best saliency?*

Saliency maps significantly vary between different XAI methods, different evaluated detectors, and different data distributions.



❖ *How to systematically compare and evaluate saliency map quality?*

Traditional metrics cannot capture the multi-level information embedded in a saliency map; whereas model-dependent approaches limit comparisons across models, datasets, and domains with costly inference per input manipulation.

Image Space	Distribution Space	Input Space
SSIM, IoU, RMSE	KL, EMD	IC, AD, ADD, I/D AuC
Sensitive to noise, pixel, and spatial changes	Not hierarchical, spatial structure is lost	Model-dependent, heavy-inference time

XMGD

We introduce a novel explainability comparison metric, eXplainable Multi-Scale GMM Distance (**XMGD**). XMGD provides a principled probabilistic algorithm for analyzing and quantifying any model or dataset similarity through the lens of explainability. Through experimental results, we demonstrate several critical advantages of XMGD over alternative saliency comparison metrics:

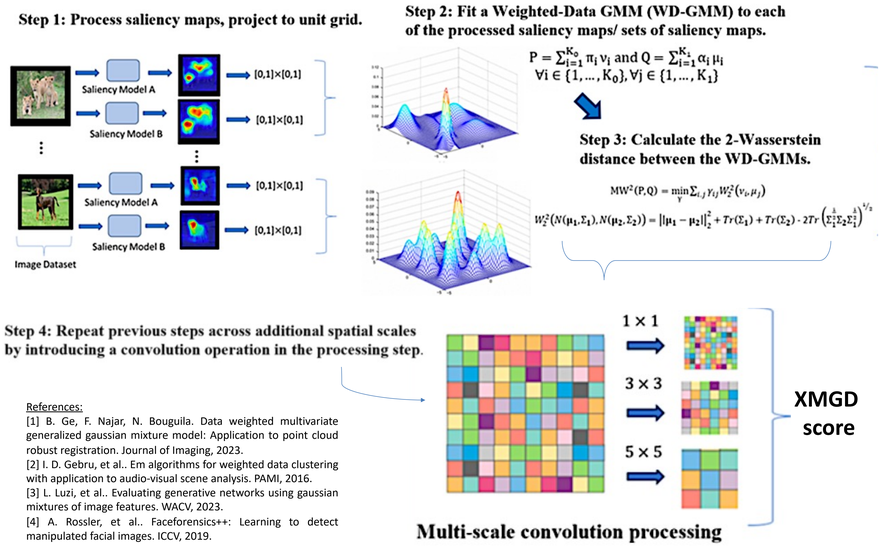
❖ XMGD is more robust against individual input/pixel saliency.
❖ XMGD is not sensitive to dataset size so it does not suffer from poor convergence properties due to small datasets.
❖ XMGD operates directly on the saliency images, without the need for input or model manipulation.
❖ XMGD can enhance explainability across a large number of diverse use cases, including saliency comparisons for individual images, entire datasets, or outputs of different models.

Method

Given two input saliency maps (or two aggregations over saliency maps) representing different models or datasets, XMGD proceeds as:

1. Saliency maps are extracted and projected onto the 2D unit grid.
2. A Weighted-Data GMM (WD-GMM) [1] is fit to each of the saliency maps.
3. Mixed 2-Wasserstein distance is calculated between the WD- GMMs.
4. Steps (1-3) are repeated across spatial scales by a convolution operation.
5. Final *XMGD score* is computed as a weighted sum of multi-scale distances.

WD-GMM is solved using expectation maximization over a duplicated point GMM [2], by a tunable binning parameter *b* over saliency maps on the unit grid. The distance between GMMs is computed using discrete optimal transport as in [3]. Then we repeat it across *l* scales using *l-1* 2D convolutions of sizes *w*. *{l, w, b, g}* can be adjusted per domain, dataset, model, map size.

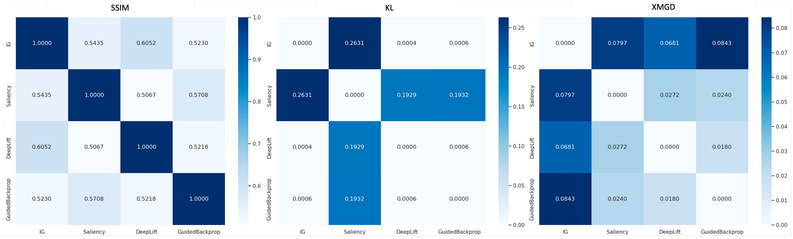


References:
[1] B. Ge, F. Najar, N. Bouguila. Data weighted multivariate generalized gaussian mixture model: Application to point cloud robust registration. Journal of Imaging, 2023.
[2] I. D. Gebru, et al.. Em algorithms for weighted data clustering with application to audio-visual scene analysis. PAMI, 2016.
[3] L. Luzi, et al.. Evaluating generative networks using gaussian mixtures of image features. WACV, 2023.
[4] A. Rossler, et al.. Faceforensics++: Learning to detect manipulated facial images. ICCV, 2019.

Experiments

❖ Diversity of distances

Experimental results for generated saliencies are aggregated over seven detectors, then their pairwise similarity is compared across four XAI methods.



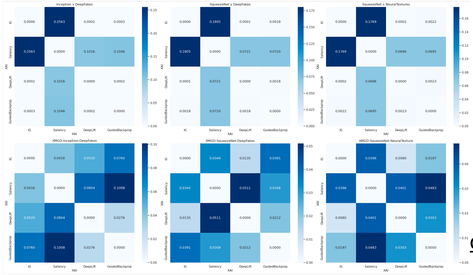
When compared to KL Divergence and SSIM, XMGD demonstrates the most differentiated levels of distances between distributions.

❖ Relevance to model-based metrics

On FaceForensics[4] with five generators, using the same seven detectors, we compute model-dependent metrics IIC, AD, and ADD as for four XAI methods. IG performs the best on all. In Tab. 1 and Fig. 3, IG is preferred by both XMGD and model-based metrics, demonstrating that XMGD can capture model-based insights without input manipulation and additional inference.

Metric	IG	Saliency	DeepLift	GuidedBP
IIC↑	**0.612**	0.358	0.561	0.518
AD↓	**-0.082**	0.344	0.289	0.229
ADD ↑	**1.064**	-0.070	0.201	-0.120

❖ Preserved correlations between detectors and generators



Assuming that same detector/ generator comparisons to surface as high correlations, DF+SQ (mid) vs. DF+Inc (left) yields 0.813, DF+SQ vs. NT+SQ (right) gives 0.810, but the pair without common components yields only 0.742. XMGD captures underlying similarities.