iGOS++: integrated gradient optimized saliency by bilateral perturbations

Oregon State University

ACM-CHIL 2021

Saeed Khorram*, Tyler Lawson*, Li Fuxin

CoRIS Institute, Oregon State University

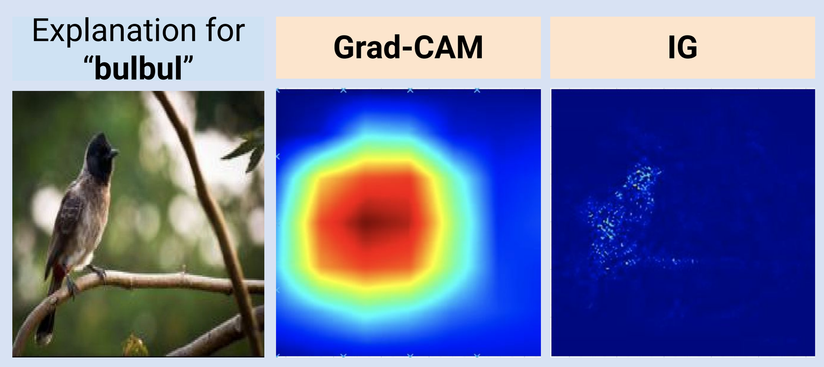
Code available at:
https://github.com/saeed-khorram/IGOS_pp

*** Equal Contributions**

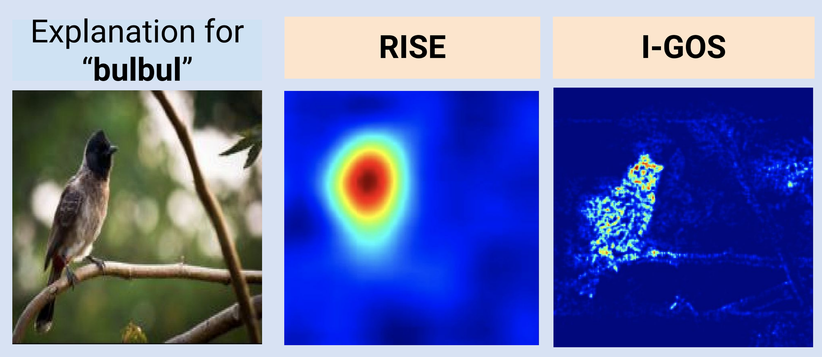
Motivation & Background

Attribution Maps

- **Backpropagation-based**
  - Less class sensitive (GuidedBP)
  - Diffuse (Gradient, IG) or Coarse (Grad-CAM)
  - Relatively fast



- **Perturbation-based**
  - More intuitive explanations
  - Usually flexible resolutions (I-GOS)
  - Relatively slow (RISE)
  - Prone to finding **adversarial masks** (I-GOS, Mask)



Pitfall of Adversarial Masks

- Previous perturbation-based methods (e.g. I-GOS) solely rely on removing evidence
  - Confidence drops quickly when deleting top pixels (i.e. **good deletion** score) but confidence does not go up when retaining top pixels (i.e. **poor insertion** score):

Revealing top 6% pixels from iGOS++, the model is 99.2% confident compared to 1.4% for I-GOS

Model Formulation

Objective Function:

$$\min_{M=(M_x,M_y)} \; F_c(I_0, M) = f_c(\Phi(I_0, \tilde{I}_0, M_x)) - f_c(\Phi(I_0, \tilde{I}_0, 1-M_y)) + f_c(\Phi(I_0, \tilde{I}_0, M_{xy})) - f_c(\Phi(I_0, \tilde{I}_0, 1-M_{xy})) + g(M_{xy})$$

$$\text{subject to} \quad g(M_{xy}) = \lambda_1 ||\mathbf{1} - M_{xy}||_1 + \lambda_2 \mathrm{BTV}(M_{xy});$$
$$M_{xy} = M_x \odot M_y; \quad \mathbf{0} \le M_x, M_y \le \mathbf{1}$$

Optimize for deletion Mx and insertion My masks while tying them together by multiplication Mxy. The regularization term encourages small and smooth output mask Mxy.

Smoothness loss, BTV, discourages mask value changes where input is not changing. This helps avoiding finding adversarial masks.

Bilateral Total Variance (BTV):

$$\mathrm{BTV} = \sum_{u \in \Lambda} e^{-\nabla I(u)^2/\sigma^2} \|\nabla M(u)\|_\beta^\beta$$

$$\sum_{s=1}^{S} F_c\left(\frac{s}{S}(M^k - \alpha^k \cdot TG(M^k))\right) - \sum_{s=1}^{S} F_c\left(\frac{s}{S} M^k\right) \le -\alpha^k \cdot \beta \cdot TG(M^k)^T TG(M^k),$$

$$TG(M) = \nabla_{I_0}^{IG} f_c(M_x) + \nabla_{I_0}^{IG} h_c(M_y) + \nabla_{I_0}^{IG} f_c(M_{xy}) + \nabla_{I_0}^{IG} h_c(M_{xy}) + \nabla g(M_{xy}).$$

Solved by using IG as the descent direction. Step size is computed using backtracking line search with revised Armijo condition. TG is the total gradient.

Contributions:

- We developed a novel visualization approach that alleviates finding adversarial masks by incorporating the insertion loss into the conventional mask optimization.
- We proposed a novel smoothness loss, BTV, that weights the variation in the mask space considering the changes in the input space.
- Through extensive qualitative experiments, we show that our method outperforms all the baselines, particularly in terms of insertion score (10-25% improvement).
- We showcase the capabilities of iGOS++ in a real-world application: debugging a COVID-19 classifier on chest x-ray images.

Evaluations and Results

ResNet50	224×224		28×28		7×7	
	Deletion	Insertion	Deletion	Insertion	Deletion	Insertion
GradCam [15]	– –	– –	– –	– –	0.1675	0.6521
Integrated Gradients [20]	0.0907	0.2921	– –	– –	– –	– –
RISE [11]	0.1196	0.5637	– –	– –	– –	– –
Mask [7]	0.0468	0.4962	0.1151	0.5559	0.2259	0.6003
IGOS [12]	0.0420	0.5846	0.1059	0.5986	**0.1607**	0.6632
iGOS++ (ours)	**0.0328**	**0.7261**	**0.0929**	**0.7284**	0.1810	**0.7332**

Table 1. Quantitative comparison in terms of deletion (lower is better) and insertion (higher is better) metrics on ResNet50 model.

Ablation	224×224		28×28	
	Deletion	Insertion	Deletion	Insertion
I-GOS	0.0420	0.5846	0.1059	0.5986
Insertion	0.0760	0.6192	0.1321	0.7231
I-GOS + Insertion (naïve)	0.0322	0.6175	0.2037	0.5103
iGOS++ (no noise)	0.0490	0.5943	0.0904	0.7108
iGOS++ (fix step size)	0.0332	0.5695	0.1052	0.7060
iGOS++ (no BTV)	**0.0245**	0.6742	**0.0813**	0.6825
iGOS++	0.0328	**0.7261**	0.0929	**0.7284**

Table 2. Results from ablation study on ResNet50.

M_x & M_y	224×224		28×28	
	Deletion	Insertion	Deletion	Insertion
M_x	**0.0268**	0.5008	0.1011	0.5536
M_y	0.0594	0.7184	0.1788	0.6912
M_{xy}(iGOS++)	0.0328	**0.7261**	**0.0929**	**0.7332**

Table 3. Comparison of the Insertion/Deletion scores of iGOS++ with Mx and My masks.

Dataset	Accuracy	F1-Score	Precision	Recall
COVIDx	95.19	93.81	**95.75**	91.85
COVIDx++	**95.93**	**95.08**	95.70	**94.49**

Table 4. Classification performance on the validation set of the COVIDx and COVIDx++ (cleaned) datasets.

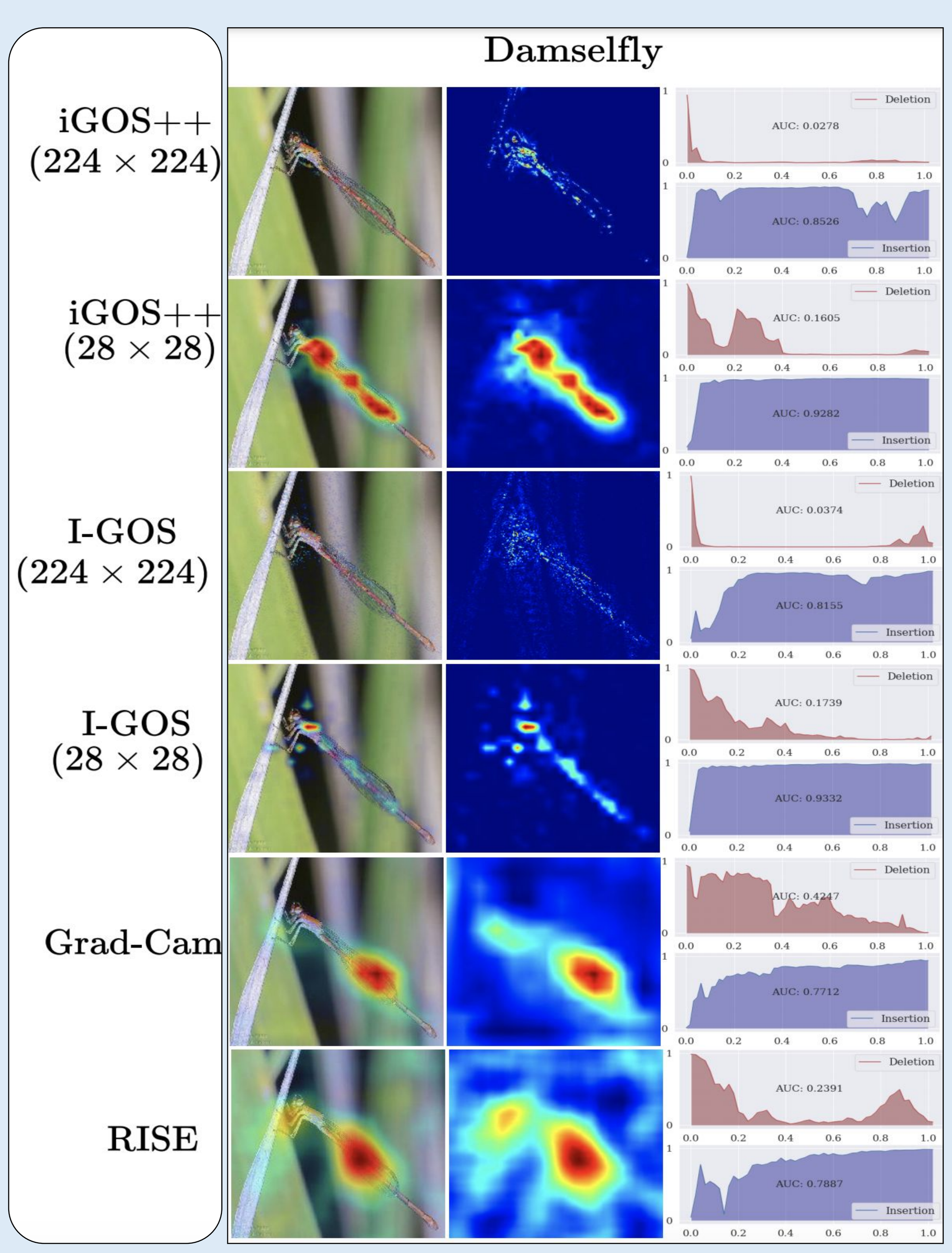


Fig 1. Visual comparison of iGOS++ where it has better insertion/deletion curves than baselines.

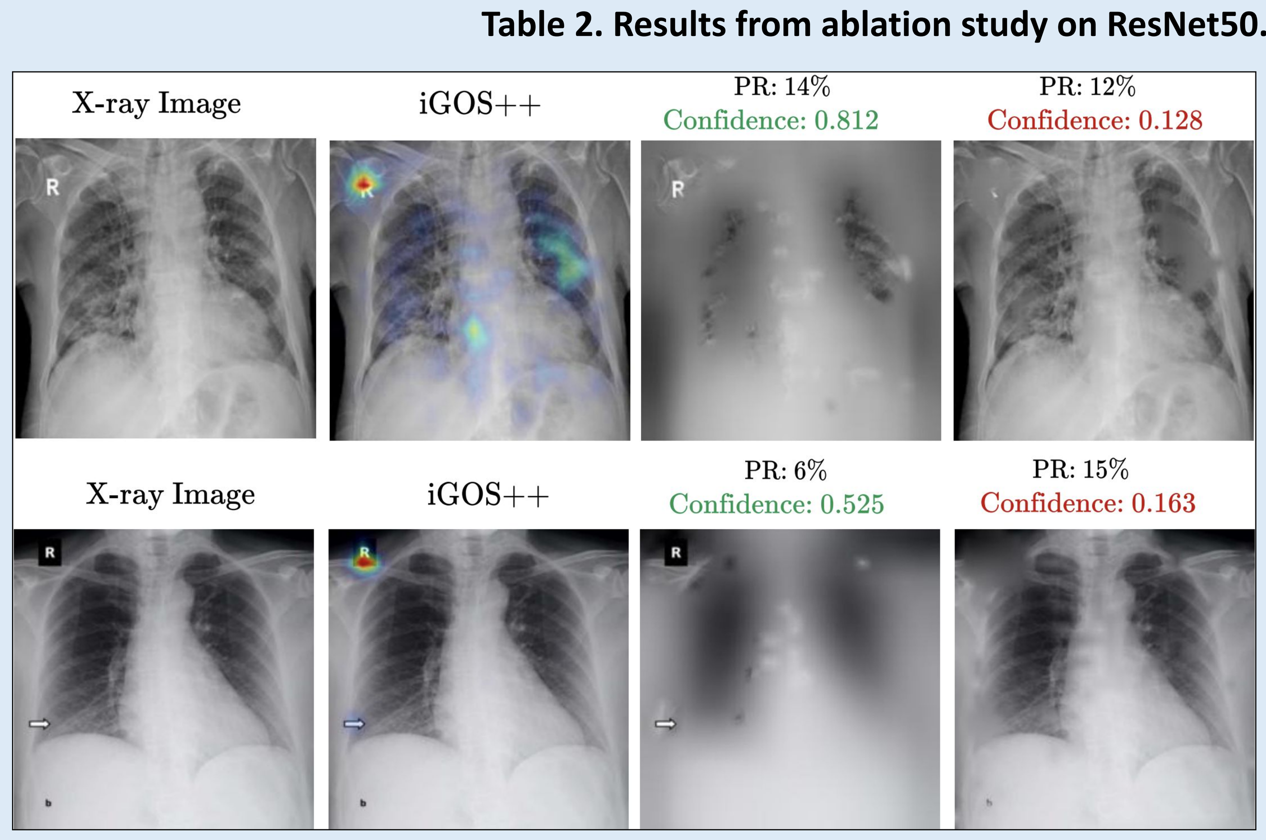


Fig 2. Examples when revealing or removing the text regions causes COVID-19 prediction or misclassification.

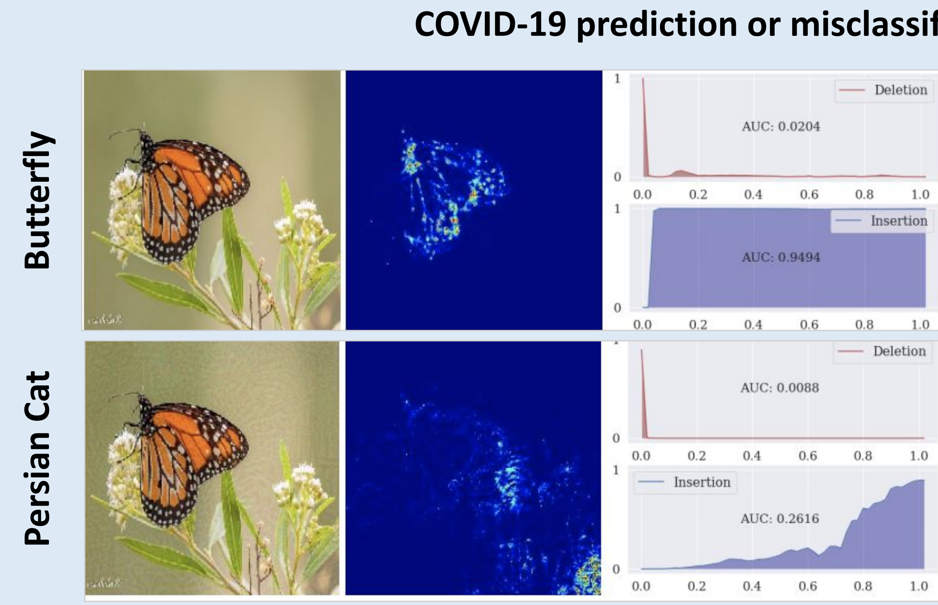


Fig 3. iGOS++ explanations for natural image of butterfly (top) and adversarial image of persian cat (bottom). The explanations are quite different. Also, for adversarial images the insertion curve goes up at the end. This shows only relying on deletion curve can be misleading.

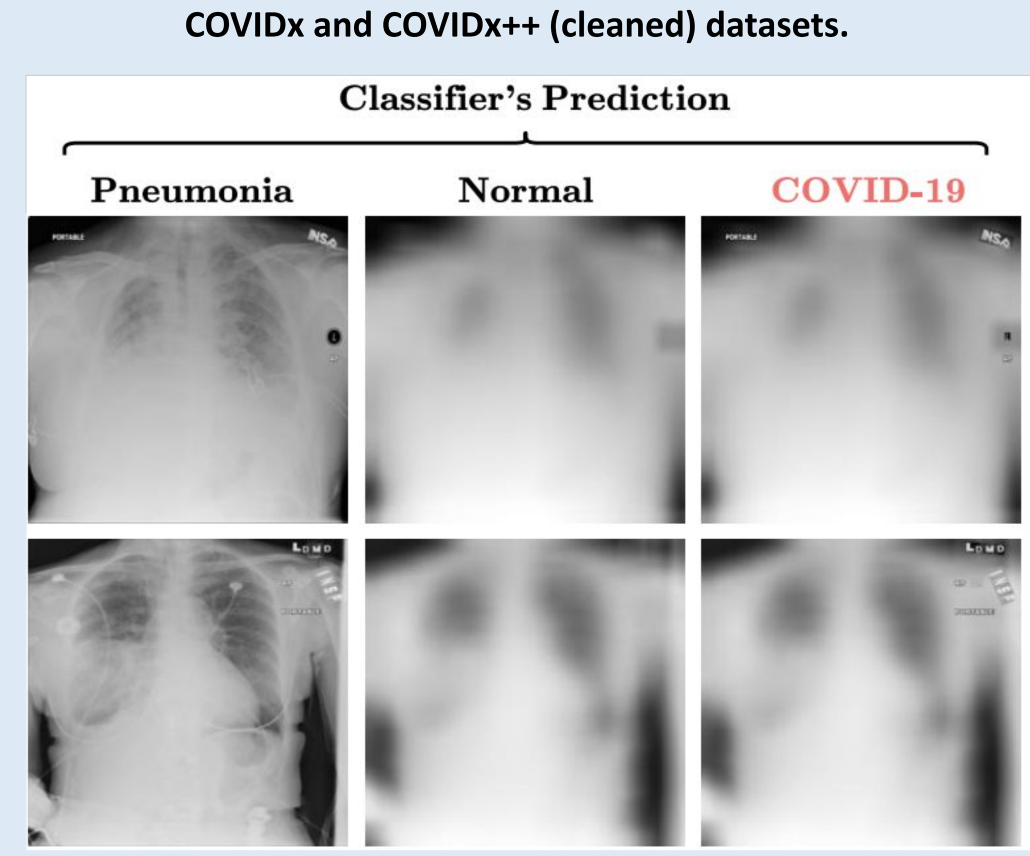


Fig 4. Chest x-ray images of "Pneumonia" patients (left). Highly blurred images are predicted as "normal" (middle). Only revealing the text regions mistakenly causes the classifier to make COVID-19 prediction (right).

This work was partially supported by DARPA contract N66001-17-2-4030.