Initialization Noise in Image Gradients and Saliency Maps

Ann-Christin Woerl, Jan Disselhoff, Michael Wand

Johannes Gutenberg University Mainz

This work has been supported by the Carl-Zeiss-Stiftung through grants "Big Data in Atmospheric Physics" and "Emergent Algorithmic Intelligence".

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

MOTIVATION

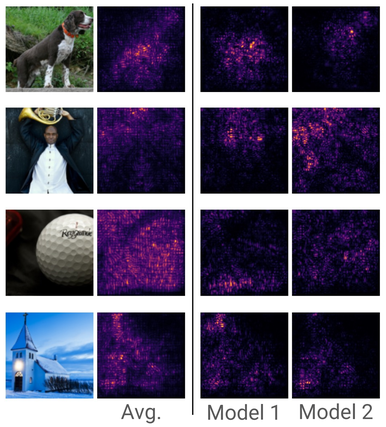
- Visualizing network decisions with different saliency methods
- Understand data by using classifier as a tool

CONTRIBUTIONS

- Initialization and training of a deep network can have strong influence on saliency maps
- Noisy artefacts can be removed by marginalization

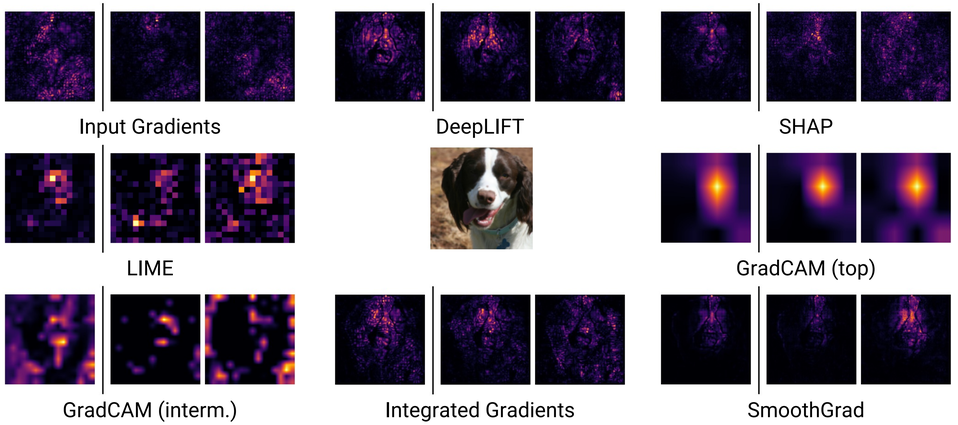
ANALYSIS OF INPUT GRADIENTS

- Logit-by-image gradients
- Mean over 50 models vs. 2 handpicked models



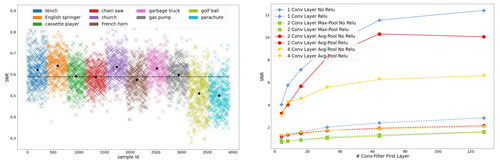
SALIENCY MAP COMPARISON

- First column: mean over 30 differently initialized models
- Column 2-3: single model with random initialization

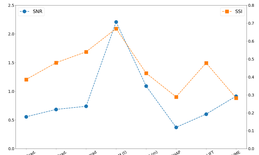


SNR ANALYSIS

- SNR varies for different classes (left)
- Impact of architectural parameter on SNR (right)



- SNR for different architectures, saliency methods and datasets (right)
- SNR vs. SSIM (left)



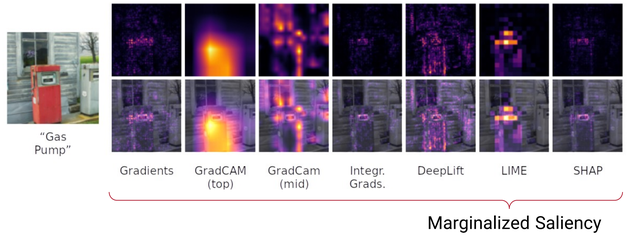
	VGG19	ResNet18	ResNet101
Input Gradient	0.436	0.555	0.435
Integrated Gradient*	0.612	0.742	0.744
Smoothgrad**	0.627	0.789	0.878
GradCAM (top)	0.816	1.593	1.373
GradCAM (interm.)	0.612	1.112	0.680
SHAP	0.252	0.365	0.308
DeepLIFT	0.453	0.628	0.554
LIME**	0.609	0.772	0.826

(a) CIFAR10

	VGG19	ResNet18	ResNet101
Input Gradient	0.585	0.589	0.444
Integrated Gradient*	0.725	0.695	0.577
Smoothgrad**	0.724	0.681	0.470
GradCAM (top)	1.370	2.363	3.171
GradCAM (interm.)	0.748	1.171	1.253
SHAP	0.418	0.386	0.327
DeepLIFT	0.978	0.650	0.587
LIME**	0.919	0.997	0.982

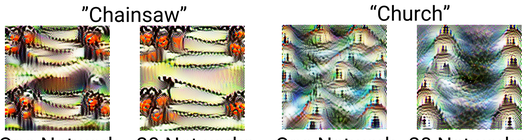
(b) Imagenette

VISUAL COMPARISON



ENSEMBLE DEEP DREAM

- Adapt "inceptionism" approach to use gradients from an ensemble of 20 networks
- Results appear to show more complete features



CONCLUSION

Explaining data by attribution

- Substantial initialization noise
- Marginalization can remove it
- Attribution results are at least incomplete

Contact: awoerl@uni-mainz.de

Project Page: