SUNY: A Visual Interpretation Framework for Convolutional Neural Networks from a Necessary and Sufficient Perspective

Xiwei Xuan[1], Ziquan Deng[1], Hsuan-Tien Lin[2], Zhaodan Kong[1], Kwan-Liu Ma[1]

1 2

CVPR JUNE 17-21, 2024 SEATTLE, WA

Motivation

- Existing CNN visual explanations often overlook the **causal perspective** that answers the core **"why"** question.
- **Counterfactual thinking** intrinsic to human cognition can better support model interpretation.
- **Necessity** (***N***) and **Sufficiency** (***S***) are two complementary sides of a desirable explanation.
  - ***N***: changing the hypothetical causes and measuring the outcome differences. E.g., RISE, CexCNN.
  - ***S***: keeping the causes unchanged and investigating the outcome stability. E.g., ScoreCAM, Group-CAM.
- We aim at rationalizing explanations toward better human understanding:
  - Consider **input features** as hypothetical causes and **outputs** as outcomes.
  - Quantify the causal ***N*** and ***S*** of input features to construct bi-directional visual explanations of CNNs.
  - Create saliency maps with two-dimensional information.

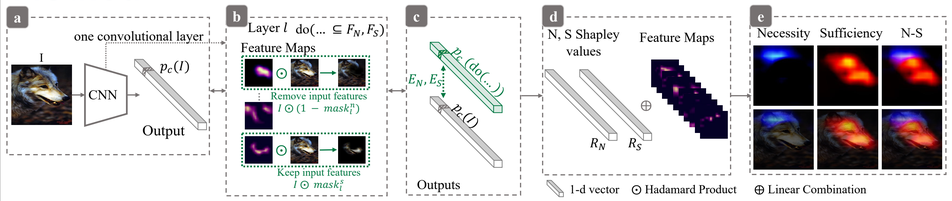
Contribution

- ***N-S* Shapley values**, for a bi-directional importance quantification for CNNs from **Necessity** and **Sufficiency** perspectives.
- A visual explanation framework, ***SUNY***, which can provide CNN interpretation better aligning with human cognition.
- **2D saliency maps** for a more informative visual explanation.

SCAN ME

Please check out our full paper for more details!

Method



N/S Shapley value

$$R_N(f_n) = \sum_{F' \subseteq \{F_N \setminus f_n\}} \frac{|F'|!(|F_N| - |F'| - 1)!}{|F_N|!} \times [E_N(F' \cup f_n) - E_N(F')]$$

$$R_S(f_s) = \sum_{F' \subseteq \{F_S \setminus f_s\}} \frac{|F'|!(|F_S| - |F'| - 1)!}{|F_S|!} \times [E_S(F' \cup f_s) - E_S(F')]$$

N/S value function

$$E_N(F_*) = [p_c(I) - p_c(\text{do}\ (F \setminus F_*))]/p_c(I)$$

$$E_S(F_*) = p_c(\text{do}\ (F_*))/p_c(I)$$

Quantitative Evaluation

Deletion & Insertion Evaluation

Dataset	Methods	VGG 16 Deletion ↓	VGG 16 Insertion ↑	VGG 16 Overall ↑	Inception-v3 Deletion ↓	Inception-v3 Insertion ↑	Inception-v3 Overall ↑	ResNet50 Deletion ↓	ResNet50 Insertion ↑	ResNet50 Overall ↑
ILSVRC	Grad-CAM[20]	0.1098	0.6112	0.5015	0.1276	0.6567	0.5291	0.1796	0.6889	0.5093
	Grad-CAM++[3]	0.1155	0.6033	0.4878	0.1309	0.6476	0.5167	0.1847	0.6799	0.4952
	SmoothGrad [22]	0.1136	0.6023	0.4887	0.1317	0.6465	0.5148	0.1849	0.6800	0.4951
	RISE [16]	0.1185	0.6188	0.5003	0.1404	0.6444	0.5040	0.1303	0.6932	0.5629
	Score-CAM [26]	0.1070	0.6382	0.5312	0.1309	0.6528	0.5219	0.2319	0.6218	0.3898
	CexCNN [5]	0.1161	0.6025	0.4864	0.1355	0.6543	0.5188	0.1886	0.6443	0.4557
	Group-CAM [35]	0.1138	0.6218	0.5080	0.1292	0.6545	0.5253	0.1794	0.6904	0.5110
	SUNY	0.1005	0.6468	0.5462	0.1215	0.6603	0.5388	0.1323	0.6988	0.5665
	SUNY-N	0.1057	0.6038	0.4981	0.1257	0.6453	0.5196	0.1374	0.6552	0.5178
	SUNY-S	0.1144	0.6389	0.5245	0.1309	0.6530	0.5221	0.2220	0.6922	0.4702
CUB	Grad-CAM[20]	0.0558	0.7617	0.7059	0.0963	0.7323	0.6360	0.0930	0.6452	0.5522
	Grad-CAM++[3]	0.0589	0.7541	0.6951	0.0950	0.7281	0.6331	0.0972	0.6407	0.5434
	SmoothGrad [22]	0.0594	0.7489	0.6895	0.0977	0.7244	0.6266	0.0974	0.6405	0.5431
	RISE [16]	0.0560	0.7583	0.7023	0.0855	0.7168	0.6314	0.0570	0.6567	0.5996
	Score-CAM[26]	0.0542	0.7575	0.7033	0.0901	0.7326	0.6424	0.0995	0.6351	0.5355
	CexCNN [5]	0.0630	0.7389	0.6760	0.1017	0.7283	0.6267	0.1014	0.6173	0.5159
	Group-CAM [35]	0.0606	0.7521	0.6915	0.0971	0.7290	0.6318	0.0926	0.6458	0.5532
	SUNY	0.0518	0.7591	0.7073	0.0842	0.7361	0.6519	0.0562	0.6645	0.6083
	SUNY-N	0.0537	0.7497	0.6960	0.0854	0.7165	0.6311	0.0667	0.6443	0.5776
	SUNY-S	0.0555	0.7577	0.7022	0.0894	0.7328	0.6434	0.0939	0.6577	0.5638

Localization Evaluation

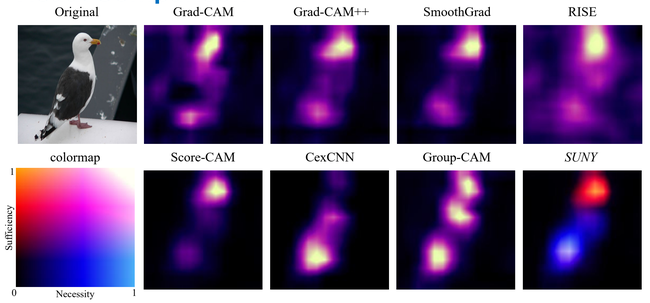
Dataset	Methods	Proportion (%) ↑ VGG16	Inception-v3	ResNet50
ILSVRC	Grad-CAM[20]	57.68	66.35	59.84
	Grad-CAM++[3]	61.31	65.93	61.74
	SmoothGrad[22]	62.18	65.78	61.75
	RISE[16]	58.93	59.26	59.48
	Score-CAM[26]	64.25	65.94	66.72
	CexCNN[5]	65.24	66.33	57.39
	Group-CAM[35]	62.70	66.17	60.68
	SUNY	65.61	66.71	68.02
CUB	Grad-CAM[20]	43.06	40.05	39.02
	Grad-CAM++[3]	45.45	40.45	41.25
	SmoothGrad[22]	47.12	40.34	41.28
	RISE[16]	37.28	34.74	36.32
	Score-CAM[26]	49.68	40.67	47.42
	CexCNN[5]	37.13	41.38	41.22
	Group-CAM[35]	43.53	41.08	40.36
	SUNY	49.97	41.96	43.21

Saliency Attack

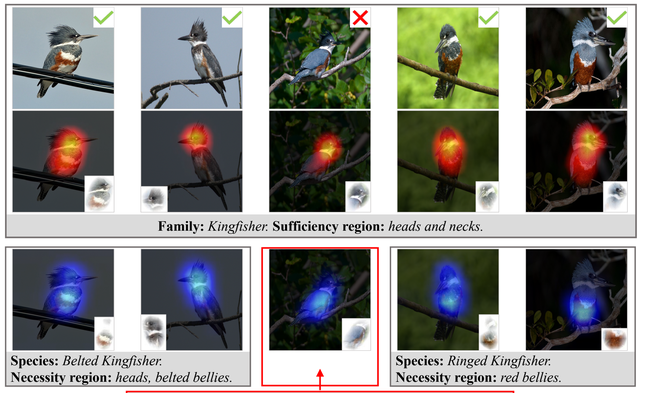
Dataset	Methods	Saliency $Attack_{score}$ ↑ VGG16	Inception-v3	ResNet50
ILSVRC	Grad-CAM[20]	0.9615	1.0435	0.7674
	Grad-CAM++[3]	0.9991	0.9821	0.8751
	SmoothGrad[22]	1.0449	0.9675	0.8776
	RISE[16]	0.9928	0.7353	1.0259
	Score-CAM[26]	0.5326	0.9673	0.3378
	CexCNN[5]	1.6341	1.0653	0.6393
	Group-CAM[35]	1.1556	1.0200	0.8020
	SUNY	2.0452	1.9874	1.5619
	SUNY-N	1.6344	1.0885	1.0726
	SUNY-S	0.5434	0.9685	0.5564
CUB	GradCam[20]	0.5969	0.5985	0.4694
	GradCam++[3]	0.6670	0.5950	0.5257
	SmoothGrad[22]	0.7783	0.5929	0.5260
	RISE [16]	0.5063	0.3860	1.1286
	Score-CAM[26]	1.2215	0.5989	0.8027
	CexCNN[5]	1.2673	0.5898	0.4171
	Group-CAM[35]	0.6742	0.5951	0.4991
	SUNY	2.8111	1.0475	1.7747
	SUNY-N	1.5863	0.7658	1.2083
	SUNY-S	1.2317	0.5884	0.8238

Qualitative Evaluation

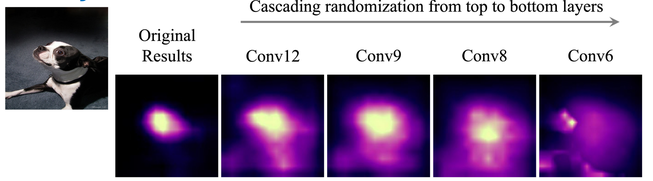
Visual Comparison



Explanations for Failure Cases



Sanity Check



Acknowledgement: This work is supported in part by NIBIB under grant no. P41 EB032840. H.-T. Lin is partially supported by the Ministry of Science and Technology in Taiwan via MOST 111-2918-I-002-006 and 112-2628-E-002-030.