

- Citation** M. Prabhushankar, and G. AlRegib, "Extracting Causal Visual Features for Limited Label Classification," submitted to *IEEE International Conference on Image Processing (ICIP)*, Sept. 2021.
- Copyright** ©2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
- Contact** mohit.p@gatech.edu OR alregib@gatech.edu
<https://ghassanalregib.info/>

EXTRACTING CAUSAL VISUAL FEATURES FOR LIMITED LABEL CLASSIFICATION

Mohit Prabhushankar and Ghassan AlRegib

School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, 30332-0250
{mohit.p, alregib}@gatech.edu

ABSTRACT

Neural networks trained to classify images do so by identifying features that allow them to distinguish between classes. These sets of features are either causal or context dependent. Grad-CAM is a popular method of visualizing both sets of features. In this paper, we formalize this feature divide and provide a methodology to extract causal features from Grad-CAM. We do so by defining context features as those features that allow *contrast* between predicted class and any contrast class. We then apply a set theoretic approach to separate causal from contrast features for COVID-19 CT scans. We show that on average, the image regions with the proposed causal features require 15% less bits when encoded using Huffman encoding, compared to Grad-CAM, for an average increase of 3% classification accuracy, over Grad-CAM. Moreover, we validate the transfer-ability of causal features between networks and comment on the non-human interpretable causal nature of current networks.

Index Terms— Visual Causality, Contrastive Explanations, COVID-19, Gradients, Causal metrics

1. INTRODUCTION

In the field of image classification, deep learning networks have surpassed the top-5 human error rate [1] on ImageNet dataset [2]. In this dataset, neural networks learn to differentiate between 1000 classes of natural images. The success of deep learning networks on natural images has fostered its usage on computed visual data including biomedical [3] and seismic [4, 5] fields. While the number of learnable classes in these fields is generally limited, neural networks have the additional task of aiding domain specific experts to interpret the explanations behind their decisions to promote trust in the network. For instance, in the field of biomedical imaging, a medical practitioner diagnoses whether a patient is COVID positive or negative based on CT scans [6]. The authors in [7] use transfer learning approaches on CT scans to perform the detection and provide explanatory results using Grad-CAM [8] to justify their network’s efficacy in detecting COVID-19. Grad-CAM highlights features in the image that lead to the network’s decision. In this paper, we analyze a neural network’s causal capability using existing explanatory methods by providing a technique to extract causal features from such explanatory methods.

Probabilistic causation assumes a causal relationship between two events C and P if event C increases the probability of occurrence of P [9]. In image classification networks, P refers to decisions made by neural networks based on features C . A popular method for ascertaining probabilistic causality is through interventions in data [10]. In these models, the set of causal features C is varied by intervening in the generation process of C to ascertain the

change in the observed decision P . Such interventions can however be long, complex, unethical or impossible [11] like in COVID CT scans. Hence, we forego interventionist causality and rely on *observed causality* to derive causation. Observed causality relies on passive observation to determine statistical causality. The authors in [12] propose that non-interventionist observation provides two sets of features - causal C , and context B - that lead to decision P . In other words, a decision P is made based on both causal and context features in an image. Hence, existing explanatory methods including [8, 13, 14] highlight $C \cup B$ features. However, they do not provide a methodology to extract either C or B separately. In this paper, we utilize contrastive features from [15] to approximate B features. We then propose a set-theoretic approach to abstract C out of Grad-CAM’s $C \cup B$ features. The contributions of this paper include:

- Formulating a set theoretic interpretation of causal and context features for observed causality in visual data.
- Expressing context features as contrastive features.
- Providing an evaluation setup that tests for causality in a limited label scenario.

In Section 2, we motivate context via contrast and review Grad-CAM and contrastive explanations. We then motivate the proposed method and detail its procedure in Section 3. We finally present the results in Section 4 before concluding in Section 5.

2. BACKGROUND AND RELATED WORKS

Causal and Context features: The authors in [12] define causal features as visual features that exist within the physical body of the object in an image and context features as visual features that surround the object in the image. In this paper, we forego the definitions based on physical locations in favor of the feature’s membership towards predicting a class P . We define causal features C as those features whose presence increases the likelihood of occurrence of decision P in any CT scan x . Conversely, the absence of causal features C decreases the probability of decision P . The above two definitions of causal features are derived from the Common Cause Principles [16] and are used in [17] to evaluate causality. We follow a slightly altered methodology to showcase the causal effectiveness of our method. We define context features B contrastively, as features that allow differentiating predicted class P and a contrast class Q , without necessarily causing P .

Context and Contrast features: In the field of human visual saliency, the authors in [18] provide an argument for the existence of contextual features of a class that are represented by their relationship with features of other classes. In [19], the implicit saliency of a neural network is extracted as an expectancy-mismatch between

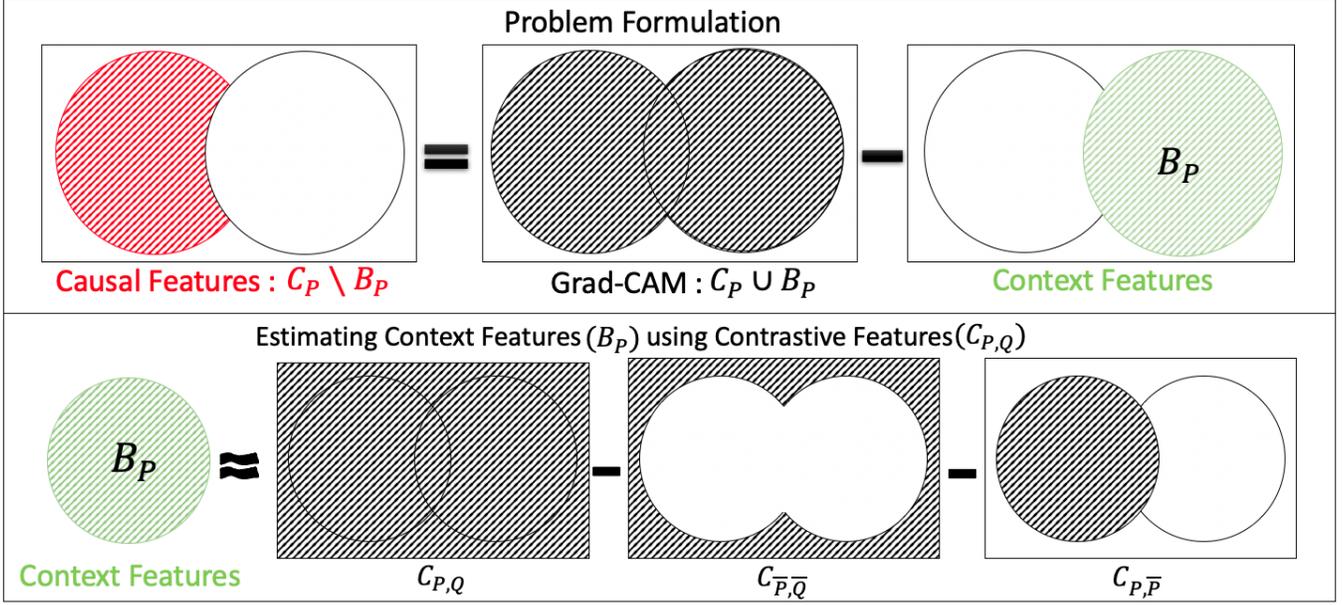


Fig. 1. Top : Venn diagram for problem formulation based on Eq. 1. Bottom : Estimating context features from Eq. 2. Note that because the network does not classify with 100% confidence, we cannot resolve $C_P \cap B_P$.

the predicted class against all learned classes thereby empirically validating the existence of *contrastive* information within neural networks. The authors in [15] extract this information and visualize them as explanations. In this paper, we represent the context features B as contrast features.

Grad-CAM and Contrastive Explanations: Consider a trained binary classifier $f()$. Given an input image x , $y = f(x)$ are the logit outputs of dimensions 2×1 . The predicted class P of image x is the index of the maximum element in y i.e. $P = \arg \max_i y_i, \forall i \in [1, 2]$. Grad-CAM localizes all features in x that leads to a decision P by backpropagating the logit y_P to the last convolutional layer l . The per-channel gradients in layer l are summed up to obtain an importance score α_k for a channel $k, k \in [1, K]$ and multiplied with the activations in their respective channels A_k . The importance score weighted activation maps are averaged to obtain the Grad-CAM mask $\mathcal{G}_P = ReLU(\sum_{k=1}^K \alpha_k A^k)$ for class P . The authors in [15] modified the Grad-CAM framework to backpropagate a loss function $\mathcal{L}_{P,Q}$ between predicted class P and a contrast class Q . With the other steps remaining the same, a contrast-importance score α_k^c weighted contrast mask is given by $\mathcal{C} = ReLU(\sum_{k=1}^K \alpha_k^c A^k)$ for predicted and contrast classes P and Q . Note that gradients are used as features in multiple works including [20, 21, 22].

3. PROPOSED METHOD

We first motivate our method based on set theory before describing the process of extraction of causal features.

3.1. Theory

Consider the setting as described in Section 2 where a binary classification network $f()$ is trained on COVID-19 CT scans [6]. Once trained, for any given scan from the dataset, Grad-CAM provides visual features that combine both causal and context features. Hence, Grad-CAM provides a mask, $\mathcal{G}_P = C_P \cup B_P$ for the prediction P on a given scan x . If the network classifies x correctly with 100% confidence, then $f()$ has resolved causal and context features in-

dependently such that $\mathcal{G}_P = C_P + B_P$. However, this rarely occurs in practice and we assume $\mathcal{G}_P = C_P + B_P - (C_P \cap B_P)$. Hence, our goal is to extract the relative complement $C_P \setminus B_P$ given $\mathcal{G}_P = C_P \cup B_P$. This is illustrated in Fig. 1. Based on a visual inspection of the venn diagram, we can rewrite $C_P \setminus B_P$ as,

$$C_P \setminus B_P = \mathcal{G}_P - B_P. \quad (1)$$

Note that we do not have access to either C_P or B_P . We are only provided with \mathcal{G}_P . In this paper, we estimate the context features B_P using contrastive features from [15]. Specifically, continuing the notations from Section 2, we represent B_P as,

$$B_P = \mathcal{C}_{P,Q} - \mathcal{C}_{\bar{P},\bar{Q}} - \mathcal{C}_{P,\bar{P}}. \quad (2)$$

Substituting Eq. 2 back in Eq. 1, we obtain our final formulation,

$$C_P \setminus B_P = \mathcal{G}_P - [\mathcal{C}_{P,Q} - \mathcal{C}_{\bar{P},\bar{Q}} - \mathcal{C}_{P,\bar{P}}]. \quad (3)$$

A venn diagram visualization is presented in Fig. 1. We qualitatively explain all the contrastive terms.

3.2. Contrastive features

$\mathcal{C}_{P,Q}$: Highlights features that answer ‘Why P or Q ?’. This term contrastively leads to either decisions of P or Q . In the binary setting, we approximate this to be all possible features \mathcal{U} . Borrowing notations from Section 2, $\mathcal{C}_{P,Q}$ is obtained by backpropagating a loss $\mathcal{L}(y, [1, 1])$ to obtain a contrast-importance score $\alpha_k^{P,Q}$.

$\mathcal{C}_{\bar{P},\bar{Q}}$: Highlights features that answer ‘Why neither P nor Q ?’. The features in this term do not increase the probability of either P or Q . $\mathcal{C}_{\bar{P},\bar{Q}}$ is obtained by backpropagating a loss $\mathcal{L}(y, [0, 0])$ to obtain a contrast-importance score $\alpha_k^{\bar{P},\bar{Q}}$.

$\mathcal{C}_{P,\bar{P}}$: Highlights features that answer ‘Why not P with 100% confidence?’. Hence, it highlights all unresolved causal features. $\mathcal{C}_{P,\bar{P}}$ is obtained by backpropagating a loss $\mathcal{L}(y, [1, 0])$ to obtain a contrast-importance score $\alpha_k^{P,\bar{P}}$.

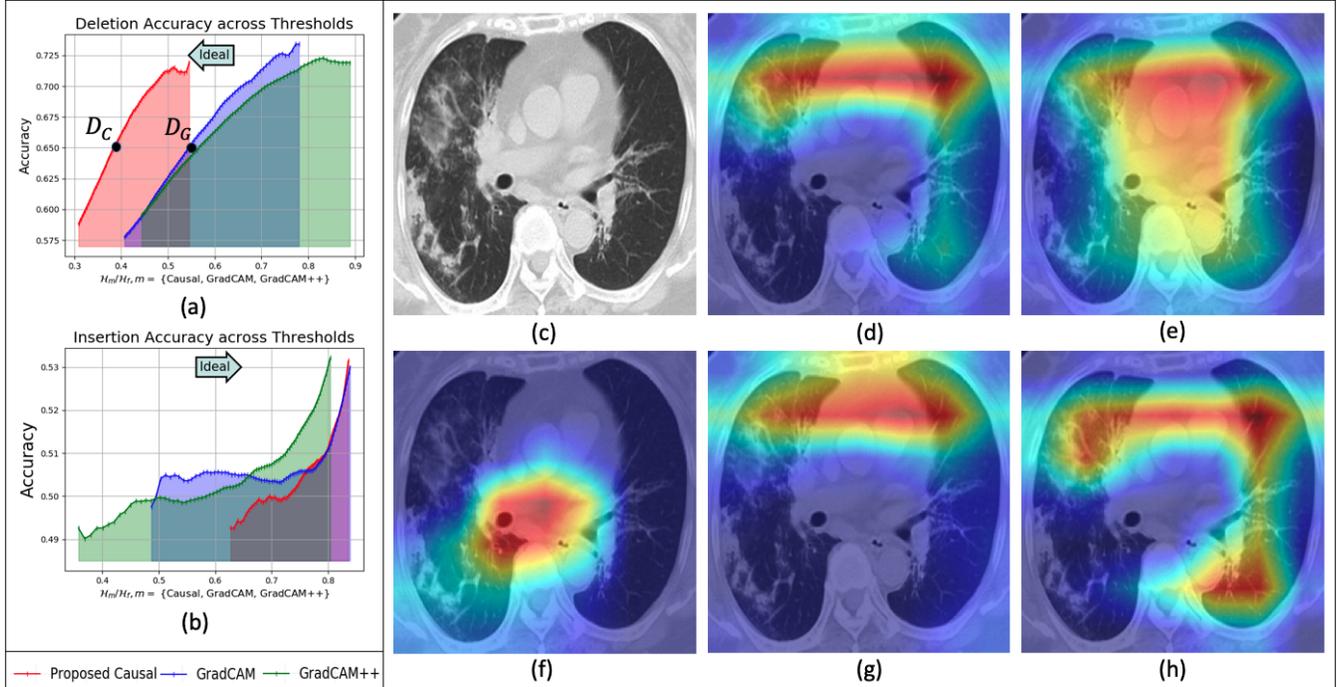


Fig. 2. (a) Deletion - Curves to the left are ideal. (b) Insertion - Curves to the right are ideal. (c) Original scan. (d) Grad-CAM. (e) Grad-CAM++. (f) Proposed causal explanation. (g) $\mathcal{C}_{\mathcal{P},\mathcal{Q}}$. (h) $\mathcal{C}_{\overline{\mathcal{P}},\overline{\mathcal{P}}}$

3.3. Implementation

Continuing the notations from Section 2, the implementation equivalent of Eq. 3 is given by,

$$\mathcal{C}_{\mathcal{P}} \setminus \mathcal{B}_{\mathcal{P}} = \text{ReLU} \left(\sum_{k=1}^K - \left[\alpha_k - \alpha_k^{P,\mathcal{Q}} + \alpha_k^{\overline{\mathcal{P}},\overline{\mathcal{Q}}} + \alpha_k^{P,\overline{\mathcal{P}}} \right] A^k \right), \quad (4)$$

where α_k represents the importance score from Grad-CAM and α_k^c represents the normalized importance score from contrast maps. The overall negative sign occurs because α are gradients whose directions are opposite to the feature minima. The final map is normalized and is visualized. A representative COVID negative scan and its Grad-CAM [8] and Grad-CAM++ [13] explanations are shown in Figs. 2c, 2d, and 2e respectively. The causal map from Eq. 4 and contrastive maps $\mathcal{C}_{\mathcal{P},\mathcal{Q}}$ and $\mathcal{C}_{\overline{\mathcal{P}},\overline{\mathcal{P}}}$ are visualized in Figs. 2f, 2g and 2h respectively. Note that while $\mathcal{C}_{\mathcal{P},\mathcal{Q}}$ appears similar to \mathcal{G} , $\mathcal{C}_{\mathcal{P},\mathcal{Q}}$ is biased by normalization and its α_k values are lesser.

Effect of number of classes: In a binary classification setting, we need four feature maps - one Grad-CAM and three contrastive maps to extract causal features. These are obtained by backpropagating $\{(0, 0), (1, 0), (1, 1)\}$ and the logit for Grad-CAM. Hence, we back-propagate the power set of all possible class combinations. This translates to 2^N backpropagations for N classes. Therefore, this technique is suitable for a limited class scenario.

4. EXPERIMENTS

In this section, we detail the experiments to validate the causal nature of our proposed features. We perform two sets of experiments to validate within-network and inter-network causality. The COVID-19 dataset [6] consists of 349 COVID positive CT scans and 463 COVID negative CT scans. We train ResNets-18,34,50 [1] and DenseNets-121,169 [23] as described in [7].

4.1. Within-network causality : Deletion and Insertion

The authors in [17] propose two causal metrics - deletion and insertion. In deletion, the identified causes are deleted pixel by pixel and the probability of predicted class, as a function of the fraction of the removed pixels, is monitored. In insertion, the non-causal pixels are added and the increase in probability as a function of added fraction of pixels is noted. However, in a binary setting, the probability for a class rarely decreases to a large extent even after removing a majority of the pixels. Hence, we modify the deletion and insertion setup to measure accuracy instead of probability on masked images.

A threshold is applied on the Grad-CAM [8], Grad-CAM++ [13] and proposed causal maps. For deletion, the pixels greater than the given threshold are made equal to 1 with the rest being 0. And vice-versa for insertion. The binary mask is then multiplied with the original input image and the masked image is passed through the model. The model's prediction is noted. This is conducted on all images in the testing set and the average test accuracy is calculated. The experiment is conducted with 81 thresholds ranging from 0.1 to 0.8 with an increment of 0.01. The average accuracy in each case is noted. From Figs. 2f and 3, we see that the area highlighted by the proposed causal features is lesser than the compared methods. To objectively measure this area, we encode the original and masked images using Huffman coding [24] as \mathcal{H}_f and \mathcal{H}_m respectively. The ratio of the bits is taken as $\mathcal{H} = \mathcal{H}_m / \mathcal{H}_f$. Each average accuracy for a threshold is now associated with an \mathcal{H} . All 81 accuracies are plotted as a function of their \mathcal{H} and depicted in Fig. 2a. Consider two points D_C and D_G on the proposed causal and Grad-CAM curves respectively. These points depict roughly 65% averaged accuracy. From the corresponding bit rates in the x-axis, the causal features achieve this accuracy at a lower bit rate compared to Grad-CAM. Hence, dense causal features are encoded by lesser bits in the proposed method. This is validated in the insertion plot in Fig. 2b as well.

Table 1. Causal Feature Transference from ResNet-18 to other architectures.

Threshold	Huffman (\downarrow)		Accuracies (\uparrow)							
			ResNet-34		ResNet-50		DenseNet-121		DenseNet-169	
	GradCAM	Causal	GradCAM	Causal	GradCAM	Causal	GradCAM	Causal	GradCAM	Causal
0.1	0.7802	0.5456	0.6158	0.6502	0.7586	0.7537	0.6404	0.6453	0.7044	0.7291
0.2	0.6442	0.4549	0.5911	0.6355	0.7734	0.7783	0.6158	0.6256	0.7143	0.7685
0.3	0.5329	0.3879	0.5665	0.5764	0.7241	0.7980	0.6108	0.6207	0.6946	0.7389
0.4	0.4434	0.3329	0.5074	0.5419	0.67	0.7882	0.5911	0.5961	0.6305	0.7192
0.5	0.3715	0.2886	0.5025	0.5222	0.601	0.7586	0.5911	0.6108	0.6059	0.6847

Table 2. Causal Feature Transference from ResNet-34 to other architectures.

Threshold	Huffman (\downarrow)		Accuracies (\uparrow)							
			ResNet-18		ResNet-50		DenseNet-121		DenseNet-169	
	GradCAM	Causal	GradCAM	Causal	GradCAM	Causal	GradCAM	Causal	GradCAM	Causal
0.1	0.8352	0.6531	0.7094	0.7044	0.7783	0.7241	0.6108	0.6552	0.7389	0.7586
0.2	0.7493	0.5646	0.7044	0.6995	0.7931	0.7586	0.6059	0.6256	0.7537	0.7586
0.3	0.6584	0.4781	0.6749	0.6749	0.8177	0.7537	0.6059	0.6059	0.7389	0.7340
0.4	0.5672	0.3983	0.6502	0.6650	0.7635	0.7685	0.6059	0.5911	0.7192	0.7044
0.5	0.4749	0.3292	0.6010	0.6059	0.7783	0.7537	0.5764	0.5616	0.6897	0.6552

4.2. Inter-network causality : Transference of features

In this section, we mask input images based on features obtained from the proposed causal and Grad-CAM methods using ResNet-18 [1]. We then pass these masked images through other trained networks including ResNets-34,50 [1] and DenseNets-121,169 [23]. This experiment is designed to validate the transfer-ability of causal features identified by ResNet-18 to other networks. The accuracy and Huffman ratio results for 5 different thresholds are shown in Table 1. It can be seen that the Huffman ratio for the proposed method is lesser than Grad-CAM for all thresholds. Hence, it is able to identify dense causal features from Grad-CAM. The averaged accuracy of masked images is also shown for 4 other networks. In 19 of the 20 categories, the proposed causal feature masked images outperform Grad-CAM feature masked images with a lesser Huffman ratio. In Table 2, we extract masks using ResNet-34, perform deletion based on shown thresholds and obtain Huffman ratios for all test images. These masked images are then passed into the corresponding networks and the accuracy results are shown. In 10 of the 20 categories, the proposed causal features outperform Grad-CAM features.

4.3. Qualitative Analysis

The authors in [17] argue that humans must be kept out of the loop when evaluating causality. However, by definition, explanations are rationales used by networks to justify their decisions [25]. These justifications are made for the benefit of humans. Such justifications are required in fields like biomedical imaging where deep learning tools are used as aids by medical practitioners. We visualize Grad-CAM and their underlying causal features from the proposed technique in Fig. 3. Both original scans are from COVID positive patients. In Fig. 3a, Grad-CAM fails to highlight the circled red region that depicts COVID. More importantly, the extracted causal features are at the bottom right. Feeding the masked image into ResNet-18, the network classifies both correctly but with a higher confidence in the causal features. In Fig. 3b, we pick a scan whose Grad-CAM and causal features were classified with the same confidence but from different regions within the scans.

These results suggest that it is the context features that add human interpretability and causal features that aid classification. In real-world biomedical applications like in the considered COVID-19 detection, it is imperative to identify and make decisions based

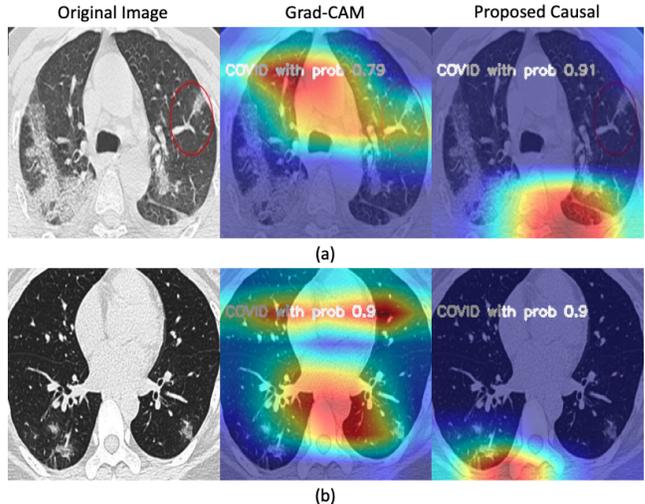


Fig. 3. (a) Non human-interpretable causal feature has higher prediction confidence. (b) The prediction confidences from both explanations are equal.

on causal features. It merits further study into designing better networks whose causal features are more human interpretable, similar to Grad-CAM’s causal and context feature set.

5. CONCLUSION

In this paper we formalize the causal and context features that a neural network bases its decision on. We express context features in terms of contrastive features between classes that the neural network has implicitly learned. This allows separation between causal and context features. Grad-CAM is used as the explanatory mechanism from which causal features are extracted. We validate and establish the transfer-ability of these causal features across networks. The visualizations suggest that the causal regions that a neural network bases its decision on is not always human interpretable. This calls for more work in designing human-interpretable causal features especially in fields like biomedical imaging.

6. REFERENCES

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Sathesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [3] Dogancan Temel, Melvin J Mathew, Ghassan AlRegib, and Yousuf M Khalifa, “Relative afferent pupillary defect screening through transfer learning,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 788–795, 2019.
- [4] Yazeed Alaudah, Patrycja Michałowicz, Motaz Alfarraj, and Ghassan AlRegib, “A machine-learning benchmark for facies classification,” *Interpretation*, vol. 7, no. 3, pp. SE175–SE187, 2019.
- [5] Muhammad A Shafiq, Mohit Prabhushankar, Haibin Di, and Ghassan AlRegib, “Towards understanding common features between natural and seismic images,” in *SEG Technical Program Expanded Abstracts 2018*, pp. 2076–2080. Society of Exploration Geophysicists, 2018.
- [6] Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie, “Covid-ct-dataset: a ct scan dataset about covid-19,” *arXiv preprint arXiv:2003.13865*, 2020.
- [7] Xuehai He, Xingyi Yang, Shanghang Zhang, Jinyu Zhao, Yichen Zhang, Eric Xing, and Pengtao Xie, “Sample-efficient deep learning for covid-19 diagnosis based on ct scans,” *medRxiv*, 2020.
- [8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [9] Christopher Hitchcock, “Probabilistic causation,” 1997.
- [10] Judea Pearl et al., “Models, reasoning and inference,” *Cambridge, UK: CambridgeUniversityPress*, 2000.
- [11] Mark Steyvers, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum, “Inferring causal networks from observations and interventions,” *Cognitive science*, vol. 27, no. 3, pp. 453–489, 2003.
- [12] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou, “Discovering causal signals in images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6979–6987.
- [13] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.
- [14] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.
- [15] Mohit Prabhushankar, Gukyeong Kwon, Dogancan Temel, and Ghassan AlRegib, “Contrastive explanations in neural networks,” in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3289–3293.
- [16] Gábor Hofer-Szabó, Miklós Rédei, and László E Szabó, “On reichenbach’s common cause principle and reichenbach’s notion of common cause,” *The British Journal for the Philosophy of Science*, vol. 50, no. 3, pp. 377–399, 1999.
- [17] Vitali Petsiuk, Abir Das, and Kate Saenko, “Rise: Randomized input sampling for explanation of black-box models,” *arXiv preprint arXiv:1806.07421*, 2018.
- [18] Aude Oliva and Antonio Torralba, “The role of context in object recognition,” *Trends in cognitive sciences*, vol. 11, no. 12, pp. 520–527, 2007.
- [19] Yutong Sun, Mohit Prabhushankar, and Ghassan AlRegib, “Implicit saliency in deep neural networks,” in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2915–2919.
- [20] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib, “Distorted representation space characterization through backpropagated gradients,” in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2651–2655.
- [21] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib, “Backpropagated gradient representations for anomaly detection,” in *European Conference on Computer Vision*. Springer, 2020, pp. 206–226.
- [22] Jinsol Lee and Ghassan AlRegib, “Gradients as a measure of uncertainty in neural networks,” in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2416–2420.
- [23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [24] David A Huffman, “A method for the construction of minimum-redundancy codes,” *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.
- [25] Philip Kitcher and Wesley C Salmon, *Scientific explanation*, vol. 13, U of Minnesota Press, 1962.