SEA-NN: Submodular Ensembled Attribution for Neural Networks

Piyushi Manupriya [1] J. Saketha Nath [1] Vineeth N Balasubramanian [1]

Abstract

We propose a framework of ensembling attribution maps to learn a Submodular attribution function for neural networks. Most of the existing attribution algorithms assign attribution scores independently to each (group of) feature. The use of Submodularity in our method brings in context-awareness and also results in attribution maps being sparse, thus reducing false positives. We demonstrate this through our experiments on Brain Tumor Detection dataset and a subset of Imagenet classification dataset. To the best of our knowledge, our work is the first Submodular attribution algorithm for neural networks as well as the first method that explores the power of non-linear ensembling of attribution maps. Code for the paper: https://github.com/Piyushi-0/SEA-NN.

1. Introduction

Deep neural networks(DNNs) have excelled in some of the most complex tasks in diverse domains like image recognition, video synthesis, speech-to-text conversion and autonomous navigation to name a few (Pouyanfar et al.). One such task in which deep learning has become the go-to choice is image classification.

While the advancements in deep learning led to continuous surge in accuracy scores, this was at the cost of making the neural network‘s decisions incomprehensible. It is hard to visualize the decision boundary learnt by these state-of-the-art models, mainly because of their non-linear structure, the high dimensionality of data, a large number of classes in datasets like Imagenet (Deng et al.) and addition of noise being a popular choice for regularization. Interpretability of DNNs is extremely important not only to foster trust in DNN's prediction but also to debug these complex networks. Over the last few years, there has been a diverse set of approaches for interpretability, some of which include visualizing filters of a trained CNN (Zeiler & Fergus, 2013), approximating the original model with an interpretable surrogate model (Ribeiro et al., 2016; Guidotti et al., 2018), doing a case-based reasoning (Li et al.; Chen et al.), modifying the training of CNNs (Zhang et al., 2018; Hwa Yoo et al., 2019) and quantifying feature importances using attribution algorithms. Our work focuses on attribution algorithms for neural networks. These algorithms output an attribution map that represents feature-wise contribution of input features towards the prediction (Zeiler & Fergus, 2013; Montavon et al.; Shrikumar et al., 2017; Sundararajan et al.; Chattopadhyay et al.).

Motivated by the efficacy of ensembling in Machine Learning, we propose a new algorithm, **Submodular Ensembled Attribution(SEA)**, that learns a Submodular attribution function based on attribution maps of different attribution algorithms. (Rieger & Hansen, 2019) performed pixel-wise averaging of attribution maps and tried to show that aggregating explanation methods stabilizes explanations. Our approach explores the power of non-linear aggregation, with benefits of Submodularity. To the best of our knowledge, the only other work that uses Submodularity for interpretation of DNNs is (Elenberg et al.) where the authors used cardinality-constrained Submodular optimization to select the top-k most important pixels but their objective function was neither monotone nor Submodular, in general. Moreover, their algorithm was only for selection of a subset of features and not for assigning attribution scores.

We assume access to reliable attribution maps, which we refer to as component attribution maps and use it to learn a non-negative monotonically non-decreasing Submodular value function. Using this value function that scores a subset of input features, we quantify attribution score for a feature as the marginal gain of adding that feature to an already existing optimal subset of features. Greedy algorithms provide a constant factor approximation guarantee for maximization of a non-negative monotone Submodular function (Nemhauser et al.). The use of marginal gain and diminishing returns brings in context-awareness in our attribution algorithm. State of the art DNNs are designed to leverage feature inter-dependencies which should be taken into account while computing attribution scores. Context Aware Second-Order Interpretation(CASO) proposed in (Singla

[1] Department of Computer Science And Engineering, Indian Institute of Technology, Hyderabad, Telangana, India. Correspondence to: Piyushi Manupriya <cs18mtech11019@iith.ac.in>, J. Saketha Nath <saketha@cse.iith.ac.in>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

et al.) also tries to address the issue of context-awareness. Although, CASO computes attribution scores of group features, it still ignores the inter-dependency between different groups of features. Sparsity in CASO is tuned by a hyperparameter but the use of Submodularity makes our attribution maps inherently sparse. Moreover, we also explore the power of ensembling attribution maps, which CASO doesn't. We compare our results against (i) the component attribution maps, (ii) Pixel-wise averaged attribution map and (iii) CASO as baselines.

Contributions:

- We propose Submodular Ensembled Attribution(SEA), a novel Submodular attribution algorithm for neural networks, by ensembling attribution maps of different attribution algorithms.
- Attribution maps of SEA are sparser and have better visual coherence.
- We propose Minimal Discriminative Region Size(MDRS) metric to measure the discriminative power and sparsity of an attribution method. SEA outperforms baseline attribution algorithms based on MDRS.
- We evaluated attribution algorithms from a Human Interpretability perspective and found that SEA performed reasonably well.

2. Proposed Method

2.1. Background

Our work builds upon properties of Submodular functions. Submodular functions are a special kind of discrete functions, characterized by diminishing returns property. Submodular functions appear naturally in many discrete maximization problems like clustering, sensor placement and document summarization (Krause & Golovin, 2014).

For a set function $f : 2^V \to R$ defined on a ground set V, the marginal gain on adding an element e in the context of set A can be defined as $f(e|A) = f(A \cup e) - f(A)$. f is said to be Submodular if for any $e \notin B$, for all $A \subseteq V$ and $B \subseteq V$ such that $A \subseteq B$, $f(e|A) \geq f(e|B)$ ie. the smaller set has a larger gain on addition of a new element. On the other hand, if both the sets have equal marginal gain ie. $f(e|A) = f(e|B)$, then f is said to be Modular. In most of the applications where f acts as a valuation function, f is desired to be non-negative ie. $f(A) \geq 0$ for all $A \subseteq V$. Additionally, if $f(A) \leq f(B)$ for all $A \subseteq B$, then f is said to be Monotonic.

2.2. Learning Submodular Ensembled Attribution map

2.2.1. Formulation:

We draw inspiration from Deep Submodular Function(DSF) proposed in (Dolhansky & Bilmes) and learn DSF for attribution. A DNN whose weights are restricted to be non-negative and the activation functions used are monotone non-decreasing concave for non-negative reals, constitutes a non-negative monotone non-decreasing Submodular function when given Boolean input vectors. This is referred to as Deep Submodular Function(DSF) that can be trained in the similar way as DNNs. For more details, we urge interested readers to refer to section(3) in (Dolhansky & Bilmes). Here, we describe our generic procedure to train a DSF on component attribution maps and to assign attribution scores using it. With some modifications in the pre-processing steps, our procedure to generate attribution maps for high-dimensional images is presented in section(2.2.4).

For a given image, we take the component attribution maps and obtain a set of representative sets of pixels from it by hard-thresholding the maps. We use $\mathcal{S}$ to denote this set. After hard-thresholding an attribution map at a threshold k, we obtain a set of pixels(a boolean vector) with only those pixels present whose attribution scores were in the top-k percentile.

The set of all pixels present in an image is denoted by V. We define $\mathcal{K} = \{|S| \mid S \in \mathcal{S}\}$ and $\mathcal{S}_k = \{S \mid S \in \mathcal{S} \text{ and } |S| = k\}\ \forall k \in \mathcal{K}$. Now, our goal is to learn a Deep Submodular Function(DSF) that induces high values for the sets $S \in \mathcal{S}_k\ \forall k \in \mathcal{K}$. We use $f_{\mathbf{w}}$ to denote the DSF parameterized by $\mathbf{w}$, weights of the DNN. With hyperparameter λ controlling the amount of regularization and $\delta > 0$ representing a small margin, the optimization problem for learning the parameters of DSF becomes:

$$\min_{\mathbf{w}\geq 0} \sum_{k\in\mathcal{K}} \sum_{S\in\mathcal{S}_k} \left(\delta + \max_{A\subseteq V, |A|\leq k} f_{\mathbf{w}}(A) - f_{\mathbf{w}}(S)\right)^+ + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{1}$$

The subgradient of an element w_i belonging to the weight vector $\mathbf{w}$ is $\sum_{k\in\mathcal{K}} \sum_{S\in\mathcal{S}_k} \left(\frac{\partial f_{\mathbf{w}}(A^*)}{\partial w_i} - \frac{\partial f_{\mathbf{w}}(S)}{\partial w_i}\right) + \lambda w_i$, where A^* is a solution to the inner maximization that maximizes DSF subject to a cardinality constraint. We solve it using the constant factor greedy approximation algorithm for maximization of a non-negative monotone Submodular function, proposed in (Nemhauser et al.). We can efficiently compute this subgradient by backpropagating through the DNN. We update the weight vector $\mathbf{w}$ using Projected Gradient Descent.

The algorithm for assigning attribution scores is described in algorithm(1). Our attribution algorithm scores the features based on the marginal gain that the feature causes on being

Algorithm 1 Attribution Algorithm

```
**Input:** Trained DSF $f$, Set of candidate elements $V$
Initialize feature subset $A = \{\}$
Initialize $n = |V|$
Initialize attribution map $G[i] = 0$ for $i = 0, 1, \dots, n$
$M = \{argmax_{v \in V \setminus A} f(v|A)\}$
Pick $e \in M$
**while** $|A| < n$ and $f(e|A) > 0$ **do**
  **for** $p \in M$ **do**
    $G[p] = \frac{f(e|A)}{|M|}$
  **end for**
  $A = A \cup \{e\}$
  $V = V \setminus M$
  $M = \{argmax_{v \in V} f(v|A)\}$
  Pick $e \in M$
**end while**
```

added to an already existing (approximately)optimal feature subset.

2.2.2. Properties of Submodular Ensembled Attribution map

Axiom 1. For a given subset of pixels from the input, the output of a well-trained DSF can accurately reflect the relevance of that subset towards the class predicted by the classification network. Thus, a DSF well-trained on reliable component attribution maps of a given class can be looked as a discrete surrogate of the classification network's function corresponding to that class.

Proposition 1. *If for all subsets A that do not contain features v_i or v_j, DSF$(A \cup \{v_i\})$ = DSF$(A \cup \{v_j\})$ then*

SEA$_i(x)$ = *SEA*$_j(x)$

Based on axiom(1), if the marginal gain caused by two features, computed using DSF, always comes out to be the same, we can expect the two features to be equally important to the classification network's function. Thus, it is desirable that both the features get equal attribution scores.

Proposition 2. *If for all subsets A that do not contain feature v_i, DSF$(A \cup \{v_i\})$ = DSF(A) then*

SEA$_i(x) = 0$

Based on axiom(1), if the marginal gain caused by a feature computed using DSF, is always zero, we can expect that feature to be unimportant to the classification network's function and hence should get a zero attribution score.

Proposition 3. *With a trained DSF: $2^V \rightarrow \mathcal{R}$ and A as the feature subset that we get at the end of our attribution algorithm(1)*

$\sum_{i=1}^{n}$ *SEA*$_i(x)$ = *DSF*(A) − *DSF*$(\{\})$

Our method can be made to satisfy an axiom called Completeness by scaling the values of the trained DSF. Completeness axiom says that the attributions should add up to the difference between the output of classification network F at the input x and the baseline x'. We can scale the values of trained DSF such that DSF$(\{\}) = F(x')$ and DSF$(A) = F(x)$, This makes SEA satisfy Completeness ie.
$\sum_{i=1}^{n} \text{SEA}_i(x) = F(x) - F(x')$.

2.2.3. Computational Effort:

At every epoch, we call the cardinality constraint maximization function only once with k_{max} as the cardinality constraint, where k_{max} is the maximum element in the set $\mathcal{K}$. Use of the greedy algorithm ensures that while computing A^* corresponding to k_{max}, the A^*'s corresponding to $k < k_{max}$ are also computed. The greedy approximation algorithm for cardinality-constrained non-negative monotone Submodular maximization with cardinality k has time complexity O($k|V|$). Our DSF is a neural network, and so we want as less number of function calls as possible. We can reduce the number of calls to DSF from $k_{max}|V|$ to k_{max} by passing all candidate elements of size $|V|$ as a single batch.

2.2.4. Scaling to high-dimensional images

For high-dimensional images, learning DSF as a function of pixels is computationally heavy. Here, we describe the approach of learning DSF as a function of sub-pixels.

First, we segment the image for which we want the attribution map and obtain segmented component attribution maps with attribution score for a segment being the normalised sum of attribution scores of pixels present in that segment. A similar approach was earlier used in (Kapishnikov et al., 2019). We then hard-threshold the segmented attribution maps and obtain a set of segments in the top-k percentile. Segmentation algorithms do not guarantee equal-sized segments. This restricts us from learning DSF as a function of segments. DSF is a monotonically non-decreasing function so the greedy algorithm for cardinality-constrained maximization will be biased to pick a larger segment because of it's larger marginal gain. Hence, we sub-sample the thresholded segmented component attribution maps and learn the DSF as a function of sub-pixels following the same procedure described in section (2.2.1). For sub-sampling, we stride a window of size 8x8, in a non-overlapping manner, across the segmented thresholded component attribution map and pick the mode of the pixel values present in the window. After training the DSF, we get our attribution map using algorithm (1) with V as the set of sub-pixels. Finally, we upsample the attribution map to match the original resolution of input image.

We also tried sub-sampling just the thresholded attribution maps but thresholding after segmentation gave less noisy thresholded maps.

2.2.5. Minimal Discriminative Region Size(MDRS) metric

Most of the popular metrics proposed in (Bach et al., 2015), (Fong & Vedaldi, 2017), IROF (Rieger & Hansen, 2020) and Causal metric (Petsiuk et al., 2018) perturb the input pixels/super-pixels according to it's attribution scores and measure the change in the output logit or output probability of the classification network. While all the attribution algorithms that we compare against access the classification network directly, SEA is learnt only on component attribution maps without having explicit access to the classification network. So, to make the comparison fair, we propose, Minimal Discriminative Region Size(MDRS) metric, that only depends on the class label predicted by the classification network after perturbing the input.

We define MDRS as the minimum number of pixels/super-pixels that we need to perturb in order to change the class predicted by the classification network. In our experiments, we obtain MDRS greedily from an attribution map by perturbing the input pixels starting from the most relevant one and continuing until the predicted class changes. For high-dimensional inputs, we obtain segmented attribution maps in the same way as described in section(2.2.4) and then start perturbing the segments starting from the most relevant one and continuing until the predicted class changes. An attribution algorithm with lower MDRS across inputs can be expected to capture the discriminative regions in the input image with less number of false positives.

3. Experiments

We conducted all our experiments on NVIDIA GTX 1080 Ti and used PyTorch for implementation. For implementation of some of the baseline attribution algorithms, we used PyTorch's Captum (Kokhlikyan et al., 2019) library. We experimented on Brain Tumor Detection dataset(3.1) and on 1000 images randomly selected from Imagenet classification dataset(3.2). Following the procedure described in section(2.2.4) for scaling to high-dimensional images, we learn a DSF as a function of sub-pixels and obtain an attribution map of resolution 28x28 which we later upsample to 224x224. Please note that although our proposed algorithm and CASO involve solving an optimization per-image, we tuned the hyper-parameters for optimization only on one randomly picked image and used that for all the images.

For both the datasets, we used $\{97, 97.5, 98, 98.5, 99, 99.5\}$ as the set of percentiles for thresholding component attribution maps. The DSF architecture was chosen to be a 4-layer feed-forward neural network with square root activation function. Weights of the DNN to learn a DSF have to be non-negative. We sampled DSF's initial weights from a uniform distribution between 2 and 2.25, for both the datasets. We found that using weights close to zero resulted in faster convergence but the choice of weights had negligible effect on the results.

3.1. Results on Brain Tumor Detection dataset

We fine-tuned a pre-trained VGG-11 network on a publicly avaialble dataset for Brain Tumor Detection (https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection). The validation accuracy at the end of training was 82%. We used Fenzenswalb's algorithm (Felzenszwalb & Huttenlocher) for segmentation of image with scale as 50 and selected the attribution maps of Integrated Gradients(IG) (Sundararajan et al.), Deep Lift(DL) (Shrikumar et al., 2017), DeepLift-Shap(DL-Shap) (Lundberg & Lee, 2017), Input.Gradient(INP-GR) and Guided GradCam(GGC) (Selvaraju et al., 2016) as the component attribution maps for training DSF. We used Adam optimizer with a weight decay of $(1e-6)$ and the value of λ as 10 and trained the DSF for 10 epochs, updating the weights using Projected Gradient Descent. For CASO, λ_1 was chosen to be 0.002 and the smoothness argument was set.

In the Brain Tumor Detection dataset, white mass inside the skull is known to be indicative of Brain Tumor. Figures (1) and (2) show that SEA correctly points out the infected regions. Sparsity is highly beneficial in medical diagnosis where false positives could be fatal. We quantitatively evaluated the attribution maps using the MDRS metric (2.2.5), perturbing at the level of segments. The sum of MDRS scores across the dataset of 253 images is shown in Table(1). For all the 3 perturbation values, our proposed algorithm performs better than all the baseline algorithms.

Table(3) shows the time taken by CASO, component attribution algorithms and our proposed algorithm, averaged across inputs. The time taken by our algorithm is the time for learning the DSF. We would also like to mention that although we did not select Smooth Integrated Gradients as a component attribution map, but the average time taken by Smooth Integrated Gradients(with 10 samples) was 17.43 seconds.

3.2. Results on Imagenet

We used a pre-trained AlexNet model for Imagenet classification and experimented on a randomly picked subset of 1000 images. We first segment the images using Fenzenswalb's algorithm (Felzenszwalb & Huttenlocher) with a scale of 500. As component attribution maps, we selected the attribution maps of Integrated Gradients(IG) (Sundararajan et al.), Deep Lift(DL) (Shrikumar et al., 2017),

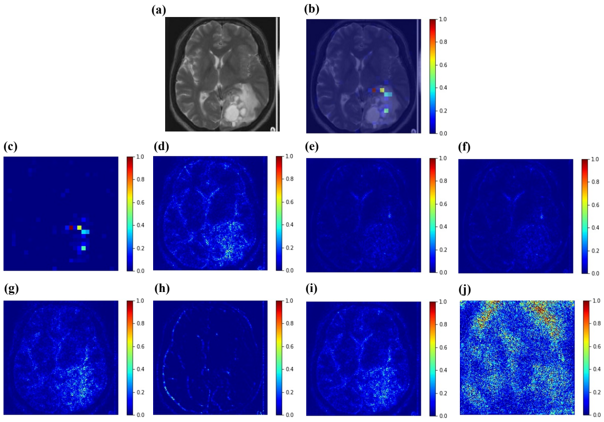

Figure 1. (a) Image of class Tumor (b) Proposed attribution map overlayed on image (c) Proposed attribution map (d) Integrated gradients (e) Deep Lift (f) Deep Shap (g) Input.Gradient (h) Guided Grad Cam (g) Pixel-wise average of component attribution maps (h) CASO

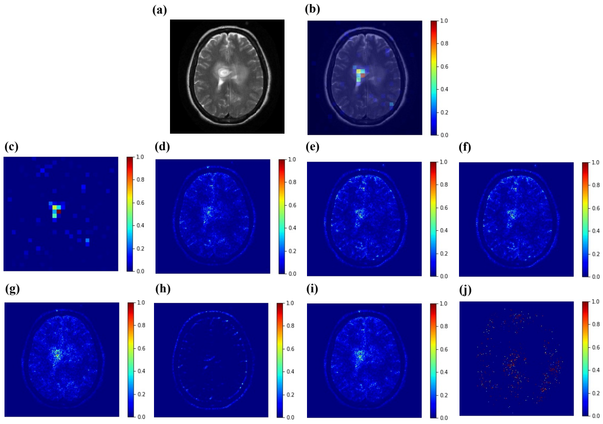

Figure 2. (a) Image of class Tumor (b) Proposed attribution map overlayed on image (c) Proposed attribution map (d) Integrated gradients (e) Deep Lift (f) Deep Shap (g) Input.Gradient (h) Guided Grad Cam (g) Pixel-wise average of component attribution maps (h) CASO

Table 1. MDRS on Brain Tumor Detection dataset(lower is better)

ALGORITHM	ZERO	MEAN	0.5
IG	136	217	156
DL	140	220	147
DL-SHAP	140	220	147
INP-GR	130	221	200
GGC	177	232	232
PIXEL-WISE AVG	125	215	206
CASO	29158	37476	28872
PROPOSED	**99**	**97**	**105**

Table 2. MDRS on 1000 randomly selected images from Imagenet(lower is better)

ALGORITHM	ZERO	MEAN	0.5
IG	2879	3728	3283
SG	2983	3781	3610
DL	3126	3798	3556
DL-SHAP	3126	3798	3556
GGC	2632	3527	3478
PIXEL-WISE AVG	2811	3641	3408
CASO	10187	14708	14071
PROPOSED	**2471**	**3055**	**2902**

DeepLift-Shap(DL-Shap) (Lundberg & Lee, 2017), Smooth Integrated Grad(SG) (Smilkov et al., 2017) and Guided GradCam(GGC) (Selvaraju et al., 2016). We used Adam optimizer with a weight decay of $(1e-3)$ and the value of λ as 10 and trained the DSF for 15 epochs, updating the weights using Projected Gradient Descent. For CASO, λ_1 was chosen to be 0.02 and the smoothness argument was set.

Table(2) shows the quantitative evaluation based on MDRS that we get by perturbing at the level of segments. MDRS summed across the 1000 images shows that our proposed method achieves the best score across all perturbation values. Qualitatively, we can see that we are able to highlight a sparser portion in the images that is the most relevant to the predicted class. Figure(3) is an image of the class 'wishing cap'. Our proposed attribution algorithm correctly highlights the cap but the other attribution algorithms highlight the entire face of the person. Similarly, Figure(4) shows that in the image belonging to the class 'Respirator', our attribution map highlights the respirator in the person's mouth but the other algorithms highlight a lot of unrelated regions too.

Table(3) shows the time taken by CASO, component attribution maps and our proposed attribution map, averaged across inputs. The time shown for our proposed attribution map is the time for training DSF.

3.2.1. HUMAN INTERPRETABILITY ON IMAGENET

We conducted an experiment to see if ensembling improves human interpretability. To evaluate the attribution maps on the basis of human interpretability, we use the human annotations provided by (Mohseni & Ragan, 2018) for ninety eight images of the Imagenet dataset. We compute the Jaccard score between an attribution map and the annotated heat map after hard-thresholding them to keep only the top-k pixels. We compute this for multiple thresholds k and show the plot in Figure(5). For lower threshold values, our proposed attribution map consistently has the highest

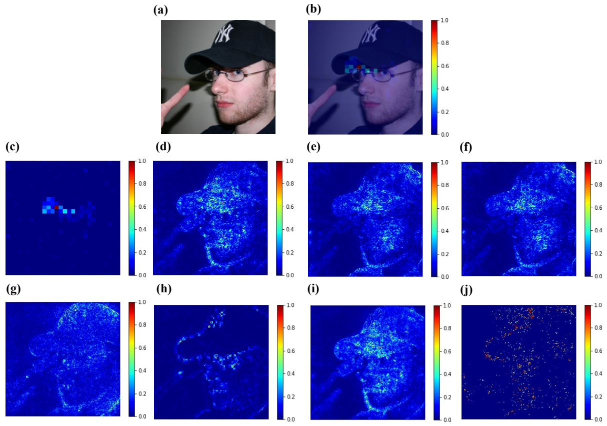

Figure 3. (a) Image of class Wishing Cap (b) Proposed attribution map overlayed on image (c) Proposed attribution map (d) Integrated gradients (e) Deep Lift (f) Deep Shap (g) Smooth Integrated Grad (h) Guided Grad Cam (g) Pixel-wise average of component attribution maps (h) CASO

Jaccard score. This shows that the most important regions captured by our attribution map aligns well with human judgement. However, for higher values of thresholds we do not get good results. We believe that this is because our DSF has been trained using subsets that belong to the top-k percentile based on their attribution scores and so it could not identify the subset of pixels that belonged to the bottom-k percentiles well.

Table 3. Time(in sec) comparison on Brain Tumor Detection Dataset and a subset of Imagenet dataset

ATTRIBUTION ALG.	BRAIN	IMAGENET
IG	2.55	0.57
DL	0.33	0.11
DL-SHAP	2.17	0.45
INP-GR	0.19	NOT USED
SG	NOT USED	3.90
GGC	0.29	0.13
CASO	9.29	0.83
PROPOSED	3.35	3.38

4. Conclusion and Future Work

In this work, we presented an interesting application of Deep Submodular Functions (Dolhansky & Bilmes) and proposed a novel Submodular attribution algorithm(SEA) for neural networks by ensembling different attribution maps. The experiments showed that the sparse attribution maps of SEA rightly captured the discriminative regions in the input.

In our future work, we would like to incorporate efficient minimization of DSF on the subsets belonging to the bottom-k percentile in the component attribution maps. Time taken by our method can be further reduced by replacing the greedy approximation algorithm with faster algorithms like Lazier than Lazy Greedy (Mirzasoleiman et al., 2014).

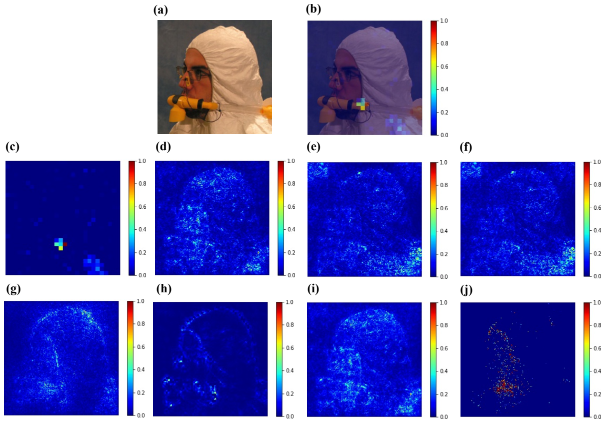

Figure 4. (a) Image of class Respirator (b) Proposed attribution map overlayed on image (c) Proposed attribution map (d) Smooth Integrated Grad (e) Deep Lift (f) Deep Shap (g) Input.Gradient (h) Guided Grad Cam (g) Pixel-wise average of component attribution maps (h) CASO

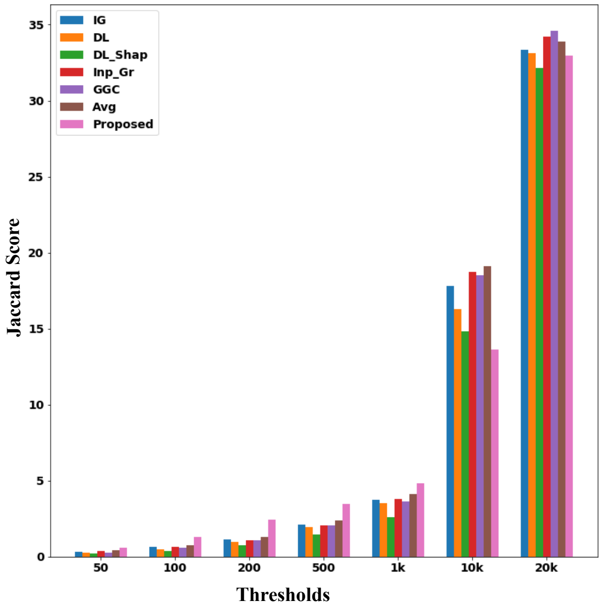

Figure 5. Jaccard score using human annotated attribution maps

5. Acknowledgement

We are grateful to the Ministry of Human Resource Development, India; Department of Science and Technology, India for the financial support of this project. We also thank the Japan International Cooperation Agency and IIT-Hyderabad for the provision of GPU servers used for this work. We thank the anonymous reviewers for their valuable feedback.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 07 2015. doi: 10.1371/journal.pone.0130140. URL http://dx.doi.org/10.1371%2Fjournal.pone.0130140.

Chattopadhyay, A., Manupriya, P., Sarkar, A., and Balasubramanian, V. N. Neural network attributions: A causal perspective. In *Proceedings of the 36th International Conference on Machine Learning*.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems 32*.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*.

Dolhansky, B. W. and Bilmes, J. A. Deep submodular functions: Definitions and learning. In *Advances in Neural Information Processing Systems 29*.

Elenberg, E., Dimakis, A. G., Feldman, M., and Karbasi, A. Streaming weak submodularity: Interpreting neural networks on the fly. In *Advances in Neural Information Processing Systems 30*.

Felzenszwalb, P. F. and Huttenlocher, D. P. *International Journal of Computer Vision*.

Fong, R. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. *CoRR*, abs/1704.03296, 2017. URL http://arxiv.org/abs/1704.03296.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018.

Hwa Yoo, C., Kim, N., and Kang, J.-W. Relevance regularization of convolutional neural network for interpretable classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Kapishnikov, A., Bolukbasi, T., Viegas, F., and Terry, M. Xrai: Better attributions through regions. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Reynolds, J., Melnikov, A., Lunova, N., and Reblitz-Richardson, O. Pytorch captum. https://github.com/pytorch/captum, 2019.

Krause, A. and Golovin, D. Submodular function maximization., 2014.

Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-pre pdf.

Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J., and Krause, A. Lazier than lazy greedy. *CoRR*, abs/1409.7938, 2014.

Mohseni, S. and Ragan, E. D. A human-grounded evaluation benchmark for local explanations of machine learning. *CoRR*, abs/1801.05075, 2018.

Montavon, G., Bach, S., Binder, A., Samek, W., and Müller, K. Explaining nonlinear classification decisions with deep taylor decomposition. *CoRR*, abs/1512.02479.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions - I. *Math. Program.*, 14(1).

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. E. P., Shyu, M., Chen, S., and Iyengar, S. S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5).

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

Rieger, L. and Hansen, L. K. Aggregating explainability methods for neural networks stabilizes explanations. *CoRR*, abs/1903.00519, 2019. URL http://arxiv.org/abs/1903.00519.

Rieger, L. and Hansen, L. K. Irof: a low resource evaluation metric for explanation methods. *ArXiv*, abs/2003.08747, 2020.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

Singla, S., Wallace, E., Feng, S., and Feizi, S. Understanding impacts of high-order loss approximations and features in deep learning interpretation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. SmoothGrad: removing noise by adding noise. *ICML workshop on visualization for deep learning*, June 2017.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. URL http://arxiv.org/abs/1311.2901.

Zhang, Q., Nian Wu, Y., and Zhu, S.-C. Interpretable convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.