Explaining AI-based Decision Support Systems using Concept Localization Maps *

Adriano Lucieri[1,2][0000−0003−1473−4745], Muhammad Naseer Bajwa[1,2][0000−0002−4821−1056], Andreas Dengel[1,2][0000−0002−6100−8255], and Sheraz Ahmed[2][0000−0002−4239−6520]

[1] Technische Universität Kaiserslautern,
Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany
[2] German Research Center for Artificial Intelligence GmbH (DFKI),
Trippstadter Straße 122, 67663 Kaiserslautern, Germany
{adriano.lucieri, naseer.bajwa, andreas.dengel, sheraz.ahmed}@dfki.de

Abstract. Human-centric explainability of AI-based Decision Support Systems (DSS) using visual input modalities is directly related to reliability and practicality of such algorithms. An otherwise accurate and robust DSS might not enjoy trust of domain experts in critical application areas if it is not able to provide reasonable justification of its predictions. This paper introduces Concept Localization Maps (CLMs), which is a novel approach towards explainable image classifiers employed as DSS. CLMs extend Concept Activation Vectors (CAVs) by locating significant regions corresponding to a learned concept in the latent space of a trained image classifier. They provide qualitative and quantitative assurance of a classifier's ability to learn and focus on similar concepts important for human experts during image recognition. To better understand the effectiveness of the proposed method, we generated a new synthetic dataset called Simple Concept DataBase (SCDB) that includes annotations for 10 distinguishable concepts, and made it publicly available. We evaluated our proposed method on SCDB as well as a real-world dataset called CelebA. We achieved localization recall of above 80% for most relevant concepts and average recall above 60% for all concepts using *SE-ResNeXt-50* on SCDB. Our results on both datasets show great promise of CLMs for easing acceptance of DSS in practice.

Keywords: Explainable Artificial Intelligence · Decision Support System · Concept Localization Maps · Concept Activation Vectors

1 Introduction

Inherently inquisitive human nature prompts us to unfold and understand the rationale behind decisions taken by Artificial Intelligence (AI) based algorithms.

* Partially funded by National University of Science and Technology (NUST), Pakistan through Prime Minister's Programme for Development of PhDs in Science and Technology and BMBF projects ExplAINN (01IS19074) and DeFuseNN (01IW17002).

This curiosity has led to the rise of Explainable Artificial Intelligence (XAI), which deals with making AI-based models considerably transparent and building trust on their predictions. Over the past few years, AI researchers are increasingly turning their attention to this rapidly developing area of research not only because it is driven by human nature but also because legislations across the world are mandating explainability of AI-based solutions [14, 32].

The applications of XAI are at least as widespread as AI itself including medical image analysis for disease predictions [22], text analytics [24], industrial manufacturing [25], autonomous driving [10], and insurance sector [19]. Many of these application areas utilize visual inputs in the form of images or videos. Humans recognize images and videos by identifying and localizing various concepts that are associated with objects – for example concepts of shape (bananas are long and apples are round) and colour (bananas are generally yellow and apples are mostly red or green). XAI methods dealing with images also employ a similar approach of identifying and localizing regions in the input space of a given image that correspond strongly with presence or absence of a certain object, or concept associated with the object.

In this work we attempt to provide a new perspective on explainable image classifiers by introducing Concept Localization Maps (CLMs), which are generated to locate human-understandable concepts as learnt and encoded by a classifier in its latent space and map them to input space for a given image. These CLMs validate that the AI-based algorithm learned to focus on pertinent regions in the image while looking for relevant concepts. This work builds on Concept Activation Vectors (CAVs) proposed in [20] and extends concept-based explanation methods by introducing concept localization. The contributions of this work are as follows:

- We propose *Concept Localization Maps* (CLMs) as a means to localize concepts learned by Deep Neural Networks (DNNs) based image classifiers. These CLMs assure that DNNs learn right concept at right location in the image.
- We develop a new synthetically generated dataset, called Simple Concept DataBase (SCDB), of geometric shapes with annotations for 10 concepts and their segmentation maps. This dataset mimics complex relationships between concepts and classes in real-world skin lesion analysis tasks and can assist researchers in classification and localization of concepts.
- We evaluate CLMs qualitatively and quantitatively using three different model architectures (*VGG16* [28], *ResNet50* [15], and *SE-ResNeXt-50* [17]) trained on SCDB dataset to show that the proposed method works across different network architectures. We also demonstrate practicality of this method in real-world applications by applying it on *SE-ResNeXt-50* trained on CelebA dataset.

2 Related Works

2.1 Concept-based Explanation Methods

Concept-based explanation methods try to semantically link meaningful human-understandable concepts to internal states of a trained DNN. Kim et al. [20] developed a line of work based on Concept Activation Vectors (CAVs), which is a concept learning paradigm that maps human-defined concepts to a neural network's latent space. Their work has been used in various medical image analysis applications including histopathology [12], ophthalmology [13], and dermatology [22], and has been extended by an unsupervised variant that does not require pre-definition of concepts [9]. Other works have aimed at decomposing the latent space of networks [34] or mapping concepts to specific units in a network [3, 4]. Recently, Goyal et al. [11] have paired the notion of concepts with counterfactual explanations using conditional Variational AutoEncoders (VAE) to generate counterfactual images for specific concepts.

2.2 Feature Importance Estimators

Saliency or relevance maps are popular means for local interpretation of a trained DNN. Different methodologies exist to generate maps that indicate importance of input features for a given prediction. Gradient-based attribution methods use backpropagation to trace output gradients back to the input feature space. These methods generate heatmaps by backpropagating the gradient either once through the network [2, 26, 27, 30] or by using ensembling techniques [1, 29], which reduces noise considerably. On the other hand, perturbation-based attribution methods generate heatmaps by perturbing an input sample and observing changes in the model's prediction score. A distinction can be drawn between methods that perturb the inputs randomly [23], systematically [33] or based on optimization [7, 6].

Our propsoed CLM method builds on CAV-based explanation method and combines it with heatmap generation methods and highlights salient regions in input space of an image that corresponds to automatically learned class-specific concepts.

3 Datasets

3.1 SCDB: Simple Concept DataBase

Attribution methods proved to work well in simpler detection tasks where entities are spatially easy to separate [18, 8, 26] but often fail to provide meaningful explanations in more complex and convoluted domains like dermatology, where concepts indicative of the predicted classes are spatially overlapping. Therefore,

we developed and released SCDB[3], a new synthetic dataset of complex composition, inspired by the challenges in skin lesion classification using dermatoscopic images. In SCDB, skin lesions are modelled as randomly placed large geometric shapes (base shapes) on black background. These base shapes are randomly rotated and have varying sizes and colours. The *disease biomarkers* indicative of the ground truth labels are given as combinations of smaller geometric shapes within a larger base shape. These *biomarkers* can appear in a variety of colours, shapes, orientations and at different locations. Semi-transparent fill colour allows *biomarkers* to spatially overlap. The dataset has two defined classes, C1 and C2, indicated by different combinations of *biomarkers*. C1 is indicated by joint presence of concepts *hexagon* $\wedge$ *star* or *ellipse* $\wedge$ *star* or *triangle* $\wedge$ *ellipse* $\wedge$ *starmarker*. C2 is indicated by joint presence of concepts *pentagon* $\wedge$ *tripod* or *star* $\wedge$ *tripod* or *rectangle* $\wedge$ *star* $\wedge$ *starmarker*. In addition to the described combinations, further *biomarkers* are randomly generated within the base shape without violating the classification rules. Two more *biomarkers* (i.e. *cross* and *line*) are randomly generated on the base shape without any relation to target classes. Finally, random shapes are generated outside of the base shape as noise.

The dataset consists of 7500 samples for binary image classification and is divided into train, validation and test splits of 4800, 1200, 1500 samples respectively. Another 6000 images are provided separately for concept training. Along with each images, binary segmentation maps are provided for every concept present in the image in order to evaluate concept localization performance. Segmentation maps are provided as the smallest circular area enclosing the *biomarker*. Fig. 1 shows examples of dataset samples.

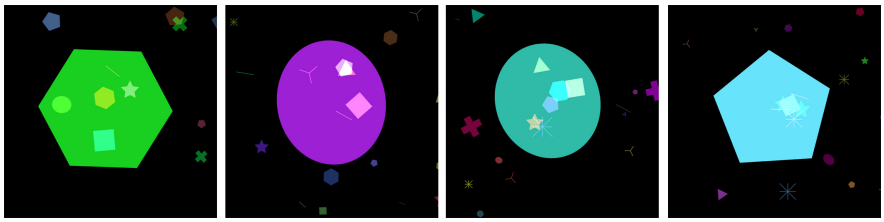


Fig. 1: Training samples from the proposed dataset.

3.2 CelebA

The proposed method can be evaluated on those real-world datasets that provide concept annotations. *CelebA* [21] is such a dataset of faces with rich face attribute annotations. The dataset contains 202599 images of 10177 identities each annotated with regards to 40 binary attributes. The dataset has been split

[3] https://github.com/adriano-lucieri/SCDB

in train, validation, and test splits of 129 664, 32 415 and 40 520 samples, evenly divided with respect to gender labels.

An important aspect to consider while selecting datasets was to find a dataset that not only contains annotations of fine-grained concepts but also high-level concepts that can be indicated by solving an interim task of fine-grained concept detection. We chose CelebA because the gender annotation allows for solving a non-trivial classification task that relies on some of the remaining annotated concepts like *baldness*, *mustache*, *lipstick* and *makeup*, which statistically indicate the gender in the given data distribution.

4 Concept Localization Maps: The Proposed Approach

The CLM method obtains a localization map m_{Cl} for a concept C learnt on DNN's layer l, that locates the relevant region essential for the prediction of a concept classifier $g_C(f_l(x;\theta))$ given an input image $x \in X$. The linear concept classifier g_C generates a concept score for concept C given a latent vector of trained DNN $f_l(x;\theta)$ with optimal weights θ at layer l. The resulting map m_{Cl} corresponds to the region in the latent space of DNN that encodes the concept C.

4.1 g-CLM

To apply gradient-based attribution methods for concept localization we must find a binary mask m_{binC} that filters out latent dimensions that contribute least to the classification of concept C. For each concept we determine those dimensions by thresholding the concept classifier's weight vector v_C, also known as CAV. High absolute weight values imply a strong influence of the latent feature dimension to the concept prediction and shall thus be retained. Therefore, a threshold value T_C is computed automatically based on 90% percentile of weight values in v_C.

Gradient-based attribution methods are applied once the latent feature dimension is masked and the concept-relevant latent subset $f_{lC}(x,\theta)$ is obtained. The methods evaluated in our work apply SmoothGrad^2 (SG-SQ) and VarGrad as ensembling approaches using plain input gradients as base-estimator defined in equations 1 and 2. The noise vector $g_j \sim \mathcal{N}(0,\, \sigma^2)$ is drawn from a normal distribution and sampling is repeated $N = 15$ times. SG-SQ and VarGrad were proven to be superior to classical SmoothGrad (SG) in [16] in terms of trustworthiness and spatial density of attribution maps. Henceforth, all experiments referring to gradient-based CLM will be denoted by g-CLM.

$$m_{Cl} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{d\, f_l C(x_{i,j}+g_j)}{d\, x_{i,j}+g_j}\right)^2 \tag{1}$$

$$m_{Cl} = Var\left(\frac{d\, f_l C(x_{i,j}+g_j)}{d\, x_{i,j}+g_j}\right) \tag{2}$$

4.2 p-CLM

The application of perturbation-based attribution methods requires local manipulation of the input image to observe changes in prediction output. In the case of CLM, the output is the predicted score of the concept classifier instead of image classifier. The systematic occlusion method from [33] is used in the experimentation section. For all reported experiments, a patch-size of 30 and stride of 10 is used, as it provides a good trade-off between smoothness of obtained maps and localization performance. Occluded areas are replaced by black patches. All experiments referring to the perturbation-based CLM method are denoted as p-CLM.

5 Experiments

This section provides a quantitative and qualitative evaluation of CLM on SCDB and CelebA to prove its feasibility in practice.

5.1 Experimental Setup

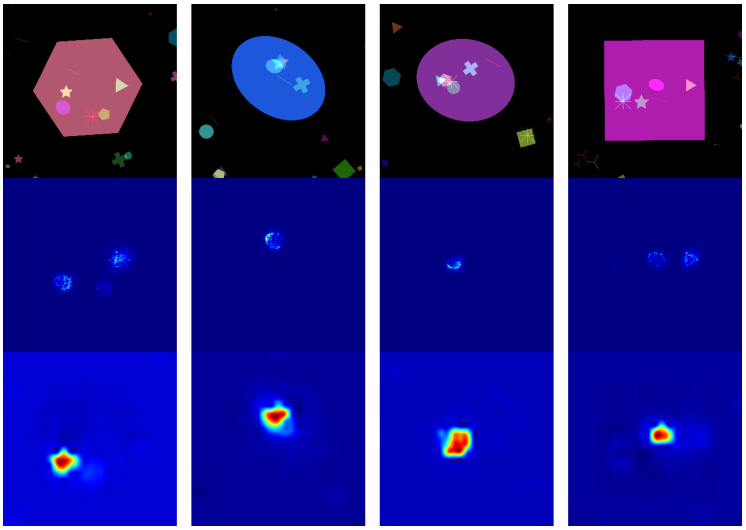
Three DNN types, namely VGG16, ResNet50 and SE-ResNeXt-50 are examined using CLM in order to study the influence of architectural complexity on concept representation and localization. All models were initialized with weights pre-trained on ImageNet [5]. Hyperparameter tuning on optimizer and Initial Learning Rate (ILR) provided best results for optimization using RMSprop [31] with ILR of 10^{-4}. Experiments were conducted for maximum of 100 epochs using learning rate decay with factor 0.5 and tolerance of 5 epochs, and early stopping after 10 epochs without improvement in validation loss.

5.2 Evaluation on Synthetic SCDB Dataset

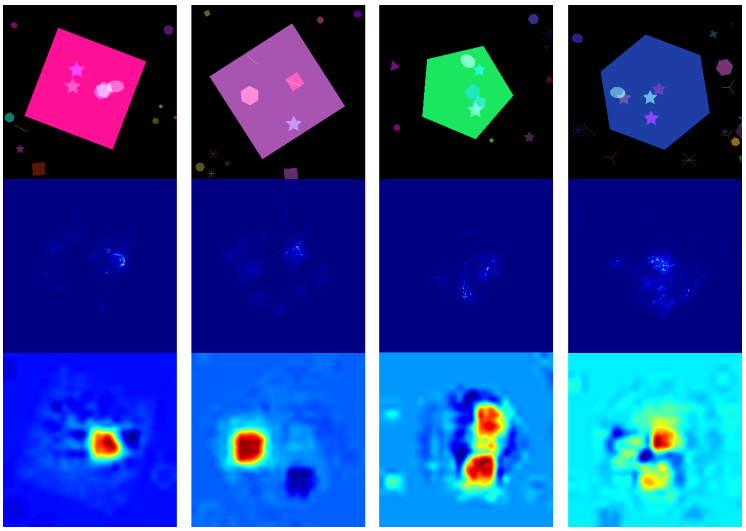
The resulting accuracies of models trained on SCDB are shown in Table 1. Surprisingly, the simplest and shallowest architecture achieved the highest test accuracy. However, the average concept classification accuracies on the architectures' last pooling layers (*pool5*) indicate that complex architectures posses more informed representations of concepts. Fig. 2 shows some examples of SCDB along with generated CLMs. Rows two and three correspond to g-CLM (SG-SQ) and p-CLM, respectively. The examples presented in this figure reveal that g-CLMs can be used to localize concepts in many cases. However, it appears that

Table 1: Test accuracies on SCDB dataset and average concept accuracies on *pool5* layer of each architecture.

	VGG16	ResNet50	SE-ResNeXt-50
Image Classification Accuracy [%]	97.5	93.5	95.6
Concept Classification Accuracy [%]	85.7	81.1	72.8



(a) Random images along with their generated CLMs for concept *ellipse*. Middle row shows g-CLM (SG-SQ) and lower row shows p-CLM.



(b) Random images along with their generated CLMs for concepts *hexagon* (columns 1 and 2) and *star* (columns 3 and 4). Middle row shows g-CLM (SG-SQ) and lower row shows p-CLM.

Fig. 2: Images from proposed SCDB dataset shown in first rows along with corresponding concept localization maps from *SE-ResNeXt-50* on layer *pool5*.

the method often highlights additional *biomarkers* that do not correspond to the investigated concept. For some concepts, localization failed for almost all ex-

amples. Furthermore, the generated maps appear to be sparse and distributed, which is typical for methods based on input gradients. The heatmaps obtained from p-CLM are extremely meaningful and descriptive, as shown in lower rows of Figures 2a and 2b. The granularity of these heatmaps is restricted by the computational cost (through chosen patch-size and stride) as well as the average concept size on the image. It is evident that the method is able to separate the contributions of specific image regions to the prediction of a certain concept. This even holds true if shapes are overlapping.

Quantitative Evaluation: To quantify CLMs performance, we compute average IoU, precision and recall between predicted CLMs and their respective ground truth masks for all images in the validation set of SCDB dataset. Therefore, the predicted CLMs are binarized using a per-map threshold from the 98% percentile. The metrics are computed for all images with a positive concept ground truth which means that images with incorrect concept prediction are included as well. Average results over all 10 concepts for all networks and variants are presented in Fig. 3. Concept localization performance of all methods increased with model complexity. This suggests that concept representations are most accurate in *SE-ResNeXt-50*. Results also clearly show that both variants of g-CLM are outperformed by p-CLM over all networks. p-CLM achieved best average localization recall of 68% over all 10 concepts, followed by g-CLM (SG-SQ) with 38% and g-CLM (VarGrad) with 36%. Most concepts relevant to the classification achieved recalls over 80% with p-CLM. The best IoU of 26% is also scored by p-CLM. It needs to be noted that IoU is an imperfect measure considering the sparsity and gradient-based CLMs and the granularity of p-CLM.

Both qualitative and quantitative analyses suggest that the performance of CLM and thus the representation of concepts is improved with the complexity of the model architecture.

5.3 Evaluation on CelebA Dataset

Learning from our experiments on SCDB, we trained only *SE-ResNeXt-50* model on the binary gender classification task in CelebA. The resulting network achieved 98.6% accuracy on the test split. Concepts that achieved highest accuracies are often strongly related to single classes like facial hair (e.g. *goatee*, *mustache*, *beard* and *sideburns*) or makeup (e.g. *heavy makeup*, *rosy cheeks* and *lipstick*). Fig. 4 shows images with their corresponding CLMs generated with our method. Due to the absence of ground truth segmentation masks in this dataset, no quantitative evaluation can be made. However, qualitative evaluation of some particularly interesting concepts is discussed below.

Lipstick: Fig. 4a shows examples of CLMs for the *lipstick* concept. Although it is quite likely that the network learnt a more abstract notion of female and male lips for the classification, the robust localization indicates that the network

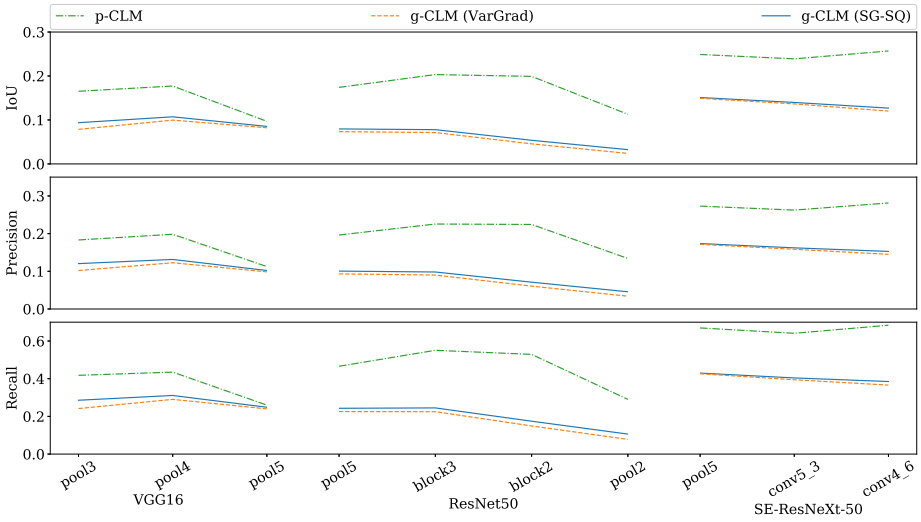


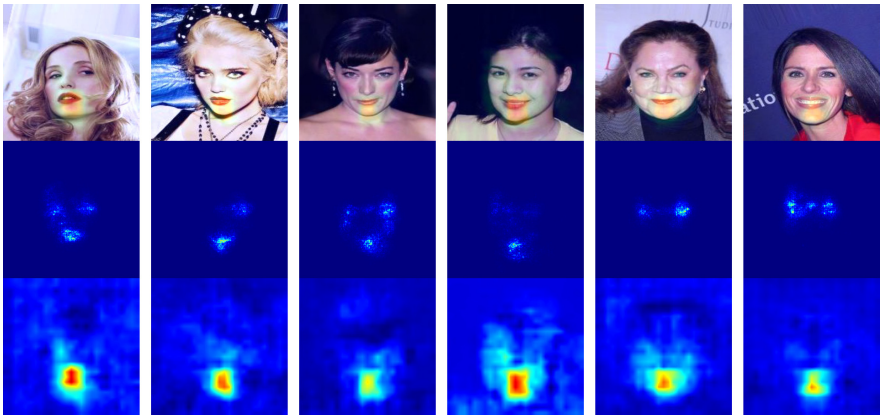
Fig. 3: Average IoU, precision and recall over all 10 concepts for predicted CLMs applied to three network architectures.

indeed encodes a lip related concept in the learnt CAV direction. It is striking that g-CLM often fails, highlighting the cheeks as well.

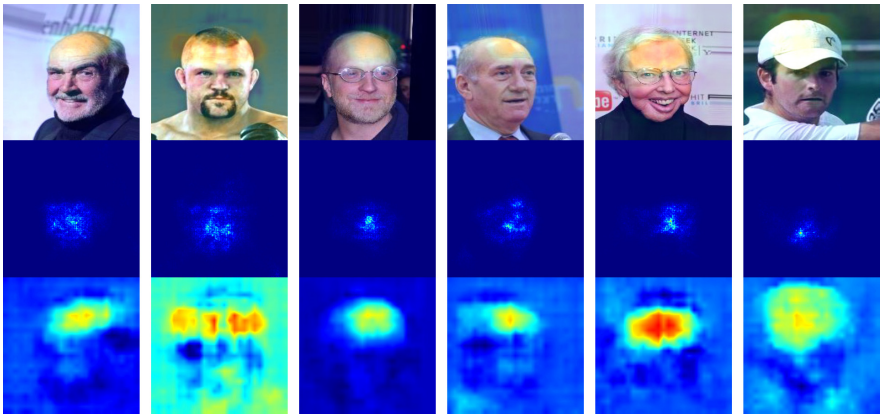
Facial Hair: All concepts related to facial hair achieved concept accuracies exceeding 80%. However, inspecting the generated concept localization maps reveals that the CAVs do not properly correspond with the nuances in concept definitions. The localization maps reveal that the concept *sideburns* never actually locates sideburns but beards in general. For the *goatee* and *mustache* it can be observed that a distinction between both is rarely made. It is thus very likely that the network learned a general representation of facial hair instead of different styles, as it would not aid solving the target task of classifying males versus females.

Bald: The *bald* concept produces almost perfect p-CLMs focusing on the forehead and bald areas. It perfectly demonstrates how the network learnt an intermediate-level feature from raw input that is strongly correlated to a target class. An intriguing finding is that often times, hats are confused with baldness as seen in the last column of figure 4b. However, g-CLM consistently failed to locate this concept.

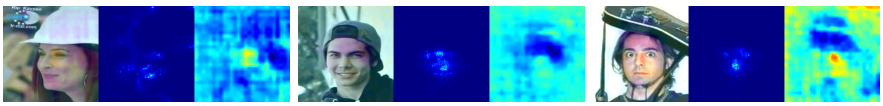
Hats: In addition to being sometimes mistaken for baldness, the localization maps for *hats* in figure 4c prove that the network struggled to learn correct representation of a hat.



(a) Images with correct prediction of concept *lipstick*. First row shows the original image with the heatmap overlay of p-CLM, second row shows g-CLM (SG-SQ) and the last row shows p-CLM.



(b) Images with positive predictions of concept *bald*. First row shows the original image with the heatmap overlay of p-CLM, second row shows g-CLM (VarGrad) and the last row shows p-CLM.



(c) Images with positive predictions of concept *hat* along with with their CLMs. Left column shows the original image with overlay of p-CLM, middle column shows g-CLM (SG-SQ) and right column shows p-CLM.

Fig. 4: Examples for CLMs generated from *SE-ResNeXt-50* trained on binary classification of gender with CelebA dataset.

6 Conclusion

We introduced the novel direction of concept localization for explanation of AI-based DSS and proposed a robust perturbation-based concept localization method (p-CLM) that has been evaluated on a synthetically generated dataset as well as a publicly available dataset of natural face images. p-CLM considerably outperformed two gradient-based variants (g-CLM) in qualitative and quantitative evaluation. Our initial results are promising and encourage further refinement of this approach. The computational efficiency and quality of heatmaps can be greatly improved by utilizing optimization-based perturbation methods like [6] and [7]. Not only will they reduce the number of network propagations by optimizing the prediction score, but also the flexible shape of masks would be beneficial for the quality of CLMs. Perturbation-based methods always introduce some distribution shift which might distort predicted outcomes. However, more sophisticated methods like image impainting could minimize distribution shifts through perturbation. The method of CLM is another step towards explainable AI that could help break the barriers in the way of practical utilization of AI-based solutions. Our SCDB dataset is also publicly available for research community to advance XAI using concept based interpretation.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems. pp. 9505–9515 (2018)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7) (2015)
3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
4. Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: Gan dissection: Visualizing and understanding generative adversarial networks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
6. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2950–2958 (2019)
7. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3429–3437 (2017)
8. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Deep learning interpretation of echocardiograms. NPJ digital medicine **3**(1), 1–10 (2020)

9. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Advances in Neural Information Processing Systems 32, pp. 9277–9286. Curran Associates, Inc. (2019)
10. Glomsrud, J.A., Ødegårdstuen, A., Clair, A.L.S., Smogeli, Ø.: Trustworthy versus explainable ai in autonomous vessels. In: Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC) 2019. pp. 37–47. Sciendo (2020)
11. Goyal, Y., Shalit, U., Kim, B.: Explaining classifiers with causal concept effect (cace). arXiv preprint arXiv:1907.07165 (2019)
12. Graziani, M., Andrearczyk, V., Müller, H.: Regression concept vectors for bidirectional explanations in histopathology. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 124–132. Springer (2018)
13. Graziani, M., Brown, J.M., Andrearczyk, V., Yildiz, V., Campbell, J.P., Erdogmus, D., Ioannidis, S., Chiang, M.F., Kalpathy-Cramer, J., Müller, H.: Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. In: Medical Imaging 2019: Computer-Aided Diagnosis. vol. 10950, p. 109501R. International Society for Optics and Photonics (2019)
14. Guo, W.: Explainable artificial intelligence (xai) for 6g: Improving trust between human and machine. arXiv preprint arXiv:1911.04542 (2019)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
16. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: Advances in Neural Information Processing Systems. pp. 9734–9745 (2019)
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
18. Jolly, S., Iwana, B.K., Kuroki, R., Uchida, S.: How do convolutional neural networks learn design? In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 1085–1090. IEEE (2018)
19. Kahn, J.: Artificial intelligence has some explaining to do (2018), https://www.bloomberg.com/news/articles/2018-12-12/artificial-intelligence-has-some-explaining-to-do
20. Kim, B., Wattenberg, M., Gilmer, J., Cai, C.J., Wexler, J., Viégas, F.B., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: ICML (2017)
21. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
22. Lucieri, A., Bajwa, M.N., Braun, S.A., Malik, M.I., Dengel, A., Ahmed, S.: On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In: IJCNN (2020)
23. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
24. Qureshi, M.A., Greene, D.: Eve: explainable vector based embedding technique using wikipedia. Journal of Intelligent Information Systems **53**(1), 137–165 (2019)
25. Rehse, J.R., Mehdiyev, N., Fettke, P.: Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory. KI-Künstliche Intelligenz **33**(2), 181–187 (2019)
26. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? arXiv preprint arXiv:1611.07450 (2016)

27. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps (2014)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
29. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3319–3328. JMLR.org (2017)
31. Tieleman, T., Hinton, G.: Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning (2012)
32. Waltl, B., Vogl, R.: Explainable artificial intelligence the new frontier in legal informatics. Jusletter IT **4**, 1–10 (2018)
33. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
34. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 119–134 (2018)