Regression Concept Vectors for Bidirectional Explanations in Histopathology

Mara Graziani [1,2], Vincent Andrearczyk [1], and Henning Müller[1,2]

[1]University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[2]University of Geneva (UNIGE), Geneva, Switzerland

Abstract. Explanations for deep neural network predictions in terms of domain-related concepts can be valuable in medical applications, where explications are important for confidence in the decision-making. In this work, we propose a methodology to exploit continuous concept measures as Regression Concept Vectors (RCVs) in the activation space of a layer. The directional derivative of the decision function along the RCV represents the network sensitivity to increasing values of a given concept measure. When applied to breast cancer grading, nuclei texture emerges as a relevant concept in the detection of tumor tissue in breast lymph node samples. We evaluate score robustness and consistency by statistical analysis.

Keywords: interpretability · concept vector · histopathology.

1 Introduction

Understanding representations learned by deep neural networks is a core challenge in medical imaging. Recent work on Testing with Concept Activation Vectors (TCAV) proposed directional derivatives to quantify the influence of user-defined concepts on the network output. As a real application example, the presence of diagnostic concepts such as microaneurysms and aneurysms was used to explain network predictions for diabetic retinopaty levels [5]. However, diagnostic concepts are often continuous measures that might be counter intuitive to describe by their presence or absence.

Intense research on network interpretability defined the distinction between global and local interpretability and proposed a taxonomy of desiderata, methods and evaluation criteria [1, 9, 11]. The relevance, or saliency, of input factors to the network decision was proposed in several gradient-based methods [11, 13, 14, 16]. Outputs of these methods are typically local explanations that are gathered in attribution maps and overlayed to the original input image. The interpretability of these approaches, however, was shown to be limited and often inconsistent [6, 12]. Research in the linearity of the latent space showed that linear classifiers can learn meaningful directions. These directions were mapped to semantic word embeddings in [10] or human-friendly visual concepts in [5]. TCAV computes the direction representative of a concept as the normal to the hyperplane which

separates a set of concept images from a set of random images. The TCAV score estimates the influence of the user-defined concept on network decisions [5].

In this paper, we extend TCAV from a classification problem to a regression problem by computing Regression Concept Vectors (RCVs). Instead of seeking a discriminator between two concepts (or one concept and random inputs), we seek the direction of greatest increase of the measures for a single continuous concept. In particular, we compute RCVs by least squares linear regression of the concept measures for a set of inputs. We measure the relevance of a concept with bidirectional relevance scores, Br. The Br scores assume positive values when increasing values of the concept measures positively affect classification and negative in the opposite case.

We address breast cancer histopathology as an application for functionally grounded evaluation. The classification of high-resolution patches as tumorous and non-tumorous tissue is often used as a first step by state-of-the-art breast cancer classifiers [15]. Identifying the factors relevant to classification is essential to improve the physicians' trust in automated grading. For this reason, we referred to the Nottingham Histologic Grading system (NHG) [2] to select nuclear pleomorphism, and especially variations in nuclei size, shape and texture as concept measures.

The main contributions of this paper are (i) the expression of concept measures as RCVs; (ii) the development and evaluation of Br scores; (iii) the computation of nuclei pleomorphism relevance for breast cancer.

In the following, we clarify the notations adopted in the paper. We consider the set $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ of inputs and ground truth pairs and a deep convolutional neural network (CNN) for binary classification with prediction output $f(\mathbf{x}_i) \in [0, 1]$. The input $\mathbf{x}_i$ is a $224 \times 224 \times 3$ image patch and $y_i \in \{0, 1\}$ is the corresponding class label (with $y = 1$ for the tumor class). The disjoint set $\{\mathbf{x}_j, c_j\}_{j=1}^{K}$ is representative of a concept C, with measures $c_j \in \mathbb{R}$ for each image sample $\mathbf{x}_j$. In the activation space, the output of layer l for input $\mathbf{x}_i$ is $\Phi^l(\mathbf{x}_i)$ and the RCV for C is $\overrightarrow{v}_C^l$ (we will drop superscript l to simplify the notation). An overview of the method is presented in Figure 1.

2 Methods

2.1 Correlation to Network Prediction

As a prior analysis, we compute the Pearson product-moment correlation coefficient ρ between c_j and $f(\mathbf{x}_j)$ for $j = 1, .., K$. If c_j is not relevant for $f(\mathbf{x}_j)$, their correlation should be low. In this case, $\Phi^l(\mathbf{x}_j)$ should not encode information about c_j and it should be unlikely to find a good linear regression. A high correlation could, instead, suggest a positive (if $\rho > 0$) or negative ($\rho < 0$) influence of the concept on the prediction.

2.2 Regression Concept Vectors

We extract and flatten the $\Phi^l(\mathbf{x}_j)$ for each $\mathbf{x}_j$. The RCV $\overrightarrow{v}_C$ is the vector in the space of the activations which best fits the direction of the strongest increase

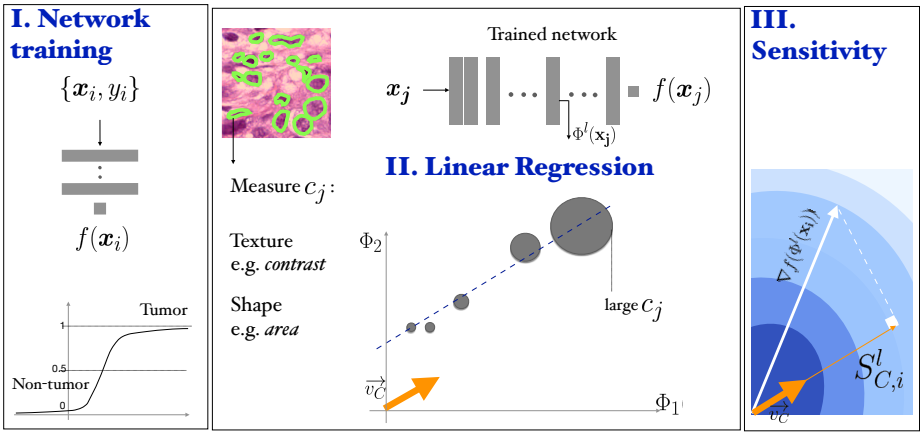


Fig. 1: *Method overview. I. Network training.* The last node of the CNN outputs a logistic regression function. The class *Tumor* is assigned to the input patch when $f(\mathbf{x}_i) > 0.5$. *II. Linear Regression.* We compute average measurements of morphological and texture features from each $\mathbf{x}_j$. Linear regression $c_j = \overrightarrow{v}_C \cdot \Phi^l(\mathbf{x}_j)$ is solved on each $(\Phi^l(\mathbf{x}_j), c_j)$ in the activation space of layer l. *III. Sensitivity.* Sensitivity is computed for the $\mathbf{x}_i$ as the derivative of $f(\mathbf{x}_i)$ along $\overrightarrow{v}_\mathbf{C}$.

of the concept measures. This direction can be computed as the least squares linear regression fit of $\{\Phi^l(\mathbf{x}_j), c_j\}_{j=1}^K$ (see Figure 1). In the NHG, for example, larger nuclei are assigned higher grades by pathologists. If we take *nuclei area* as a concept, we seek the vector in the activation space that points towards representations of larger nuclei.

2.3 Sensitivity to RCV

For each testing pair $(\mathbf{x}_i, y_i)$ we compute the sensitivity score $S^l_{C,i}$ along the direction of the RCV:

$$S^l_{C,i} = \nabla f(\Phi^l(\mathbf{x}_i)) \cdot \overrightarrow{v}_C \tag{1}$$

$S^l_{C,i}$ represents the network sensitivity to changes in the input along the direction of increasing values of the concept measures. When moving along this direction, $f(\mathbf{x}_i)$ may either increase, decrease or remain unchanged ($S^l_{C,i}$=0). The sign of $S^l_{C,i}$ represents the direction of change, while the magnitude of $S^l_{C,i}$ represents the rate of change. TCAV can be used to obtain global explanations from the N sensitivities although it does not consider their magnitude. Therefore, we propose Br as an alternative measure. Br scores were formulated by taking into account the principles of *explanation continuity* and *selectivity* proposed in [11]. For the former, we consider whether the sensitivity scores are similar for similar data samples. For the latter, we redistribute the final relevance to concepts with the strongest impact on the decision function. We define Br scores as the ratio

between the coefficient of determination of the least squares regression, R^2, and the coefficient of variation $\hat{\sigma}/\hat{\mu}$ of the N sensitivity scores:

$$Br = R^2 \times \left(\frac{\hat{\mu}}{\hat{\sigma}}\right) \tag{2}$$

$R^2 \leq 1$ indicates how closely the RCV fits the $\{\Phi^l(\mathbf{x}_i), c_i\}_{i=1}^N$. The coefficient of variation is the standard deviation of the scores over their average, and describes their relative variation around the mean. For the same value of R^2, the Br for spread scores will be lower than for scores that lay closely concentrated near their sample mean. After computing Br for multiple concepts, we scale the scores to the range [-1, 1] by dividing by the maximum absolute value.

2.4 Evaluation of Explanations

Explanations are evaluated on the basis of their statistical significance as proposed in [5]. We compute TCAV and Br scores for 30 repetitions and perform a two-tailed t-test with Bonferroni correction (with significance level $\alpha = 0.01$). If we can reject the null hypothesis of TCAV of 0.5 for random scores and Br of 0, we accept the result as statistically significant.

3 Experiments and Results

3.1 Datasets

We trained the network on the challenging Camelyon16 and Camelyon17 datasets[1]. More than 40,000 patches at the highest resolution level were extracted from Whole Slide Images (WSIs) with ground truth annotation. To extract concepts, we used the nuclei segmentation dataset in [8], for which no labels of tumorous and non-tumorous regions were available. The dataset contains WSIs of several organs with more than 21,000 annotated nuclear boundaries. From this dataset, we extracted 300 training patches only from the WSIs of breast tissue.

3.2 Network Architecture and Training

A ResNet101[4] pretrained on ImageNet was finetuned with binary cross-entropy loss for classification of tumor and non-tumor patches. For each input, the network outputs its probability to be tumorous with a logistic regression function. We trained for 30 epochs with Nesterov momentum stochastic gradient descent and standard hyperparameters (initial learning rate 10^{-4}, momentum 0.9). Staining normalization and online data augmentation (random flipping, brightness, saturation and hue perturbation) were used to reduce the domain shift between the different centers. Statistics on network performance were computed from five random splits with unseen test patients.

[1] `https://camelyon17.grand-challenge.org/` as of June 2018

3.3 Results

Classification Performance The validation accuracy of our classifier is just below the performance of the patch classifier used to get state-of-the-art results on the Camelyon17 challenge [15], as reported in Table 1. We report the per-patch validation accuracy for both models, although details about the training setup in [15] are unknown. Bootstrapping of the false positives was not performed and the training set size was kept limited (with 40K patches instead of 600K). The obtained accuracy is sufficient for a meaningful model interpretation analysis which may be used to boost the network accuracy and generalization. Besides, this analysis could be itself used as an alternative to bootstrapping for detecting mislabeled examples [7].

Table 1: Network accuracy % for binary classification of Camelyon17 patches.

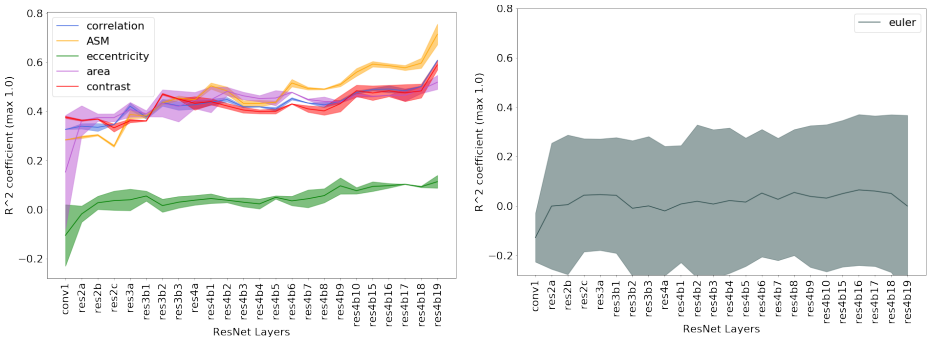
model	validation accuracy
Zanjani et al.	**98.7**
ResNet101	92.43 ± 0.657

Correlation Analysis We expressed the NHG criteria for nuclei pleomorphism as average statistics of the nuclei morphology and texture features. From the patches ($\mathbf{x}_j$) with ground truth segmentation, we computed average nuclei area, Euler coefficient and eccentricity of the ellipses that have the same second-moments as the nuclei segmented contours. We extracted three Haralick texture features inside the segmented nuclei, namely Angular Second Moment (ASM), contrast and correlation [3]. The Pearson correlation between the concept measurements and the relative network prediction is shown in Table 2. The concept measures for *contrast* had the largest correlation coefficient, $\rho = 0.41$.

Table 2: Pearson correlation between the concept measurements and the network prediction.

	correlation	ASM	eccentricity	Euler	area	contrast
ρ	$\mathbf{-0.2985}$	-0.1869	-0.1460	0.1534	0.2820	$\mathbf{0.4119}$
p-value	≤ 0.001	≤ 0.001	≤ 0.01	≤ 0.001	≤ 0.001	≤ 0.001

Are We Learning the Concepts? The performance of the linear regression was used to check if the network is learning the concepts and in which layers. The determination coefficient of the regression R^2 expresses the percentage of variation that is captured by the regression. We computed R^2 for all x_j patches over multiple reruns to analyze the learning dynamics. Almost all the concepts were learned in the early layers of the network (see Figure 2a), with *eccentricity* and *Euler* being the only two exceptions. Figure 2b shows that the concept *Euler* is highly unstable and has almost zero mean, suggesting that the learned RCVs might be random directions.



(a) (b)

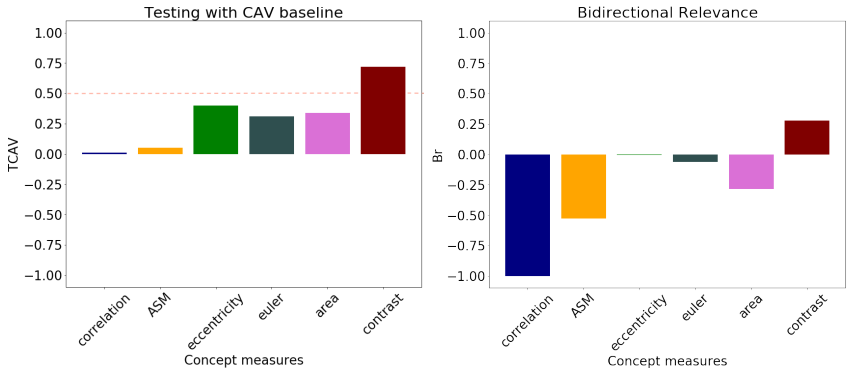
Fig. 2: (a) Linear regression determination coefficient at different layers in the network for five of the six diagnostic measurements. Results were averaged over three reruns. 95% confidence intervals are reported. (b) The RCVs for the concept *Euler* show high instability of the determination coefficient.

Sensitivity and Relevance Sensitivity scores were computed on $N = 300$ patches ($\mathbf{x_i}$) from Camelyon17. The global relevance was tested with $TCAV$ and Br, as reported in Figure 3. *Contrast* is relevant for the classification, with $TCAV = 0.75$ and $Br = 0.25$. *Correlation* has mostly negative sensitivities, leading to $TCAV = 0.1$ and $Br = -1$. These scores mirror the preliminary analysis of Pearson correlation in Table 2. Unstable concepts, such as *Euler* and *eccentricity*, lead to almost zero Br scores, in accordance with the initial hypothesis that the RCVs for these concepts might just be random vectors.

Statistical Evaluation We performed a two-tailed t-test to compare the distributions of the scores against the null hypothesis of learning a random direction for the TCAV (mean 0.5) and Br (mean 0) scores. The results are presented in Table 3. There was a significant difference (with $p\text{-}value \leq 0.01$) in the scores for all the relevant concepts, namely *correlation*, *ASM*, *area* and *contrast*. The statistical significance of *correlation* improves for Br scores. From the sensitivity and relevance analysis, we do not expect the *Euler* and *eccentricity* concepts to be statistically different from random directions. The analysis of both TCAV and Br scores confirms this hypothesis ($p\text{-}value \leq 0.01$) for the *eccentricity*, although the confidence to not reject the null hypothesis is higher with Br. The *Euler* concept, however, is not rejected from the null hypothesis with the TCAV score, but it is with the Br one as the latter takes into account the sensitivities variation.

4 Discussion and Future Work

This paper proposed RCVs as the direction of the greatest increase of concept measures in the activation space of a network layer. RCVs allow the computation



(a) (b)

Fig. 3: Comparison of TCAV ($\in [0,1]$) and Br ($\in [-1,1]$) scores. *Contrast* is relevant according to both measurements. Br scores show that higher *correlation* reduces the prediction output. Scores for the unstable *Euler* are approximately flattened to zero by Br.

Table 3: Statistical significance of the scores. The p-values are reported for two-tailed t-tests evaluating the difference between the distributions of the obtained scores against a normal distribution of the scores for random concepts, i.e. mean 0.5 for $TCAV$ and 0 for Br.

	correlation	ASM	eccentricity	Euler	area	contrast
TCAV	0.002	0.001	0.02	0.01	0.001	0.001
Br	0.001	0.001	0.30	1.0	0.001	0.001

of relevance scores for non-discrete concepts. We proposed Br scores, which assign positive and negative relevance and use different information from TCAV, such as R^2 and the variance of the individual sensitivities. As a real-application example, we proposed to measure the relevance of NHG diagnostic measures. We selected six different concept measures and we computed their relevance with TCAV and Br.

Our analysis evidenced that nuclei *contrast* and *correlation* are relevant to classification. This is in accordance with the NHG grading system, which identifies hyperchromatism as a signal of nuclear atypia. Moreover, homogeneous textures inside the nuclei (high *correlation* of the pixel values) negatively affect predictions. This suggests that the network is using this concept to identify non-tumor patches.

RCVs could allow pathologists to select concepts of interest to interpret the network output. Moreover, RCVs could be used during model development to get insights about network training. Outliers in the values of the sensitivity scores could identify challenging training inputs or highlight domain mismatches (e.g. differences across hospitals, staining techniques, etc.). Ad-hoc bootstrapping and domain adaptation could then be developed on top of this method.

Acknowledgements This work was possible thanks to the project PROCESS, part of the European Unions Horizon 2020 research and innovation program (grant agreement No 777533).

References

1. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arxiv:1702.08608 (2017)
2. Elston, C.W., Ellis, I.O.: Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology **19**(5), 403–410 (1991)
3. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics (6), 610–621 (1973)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Kim, B., Gilmer, J., Viegas, F., Erlingsson, U., Wattenberg, M.: TCAV: Relative concept importance testing with linear concept activation vectors. arXiv preprint arXiv:1711.11279 (2017)
6. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (un) reliability of saliency methods. arXiv preprint arXiv:1711.00867 (2017)
7. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. arXiv preprint arXiv:1703.04730 (2017)
8. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging **36**(7), 1550–1560 (2017)
9. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
11. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing (2017)
12. Ribeiro, M.T., Singh, S., 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)
13. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
15. Zanjani, F.G., Zinger, S., de With, P.H.N.: Automated detection and classification of cancer metastases in whole-slide histopathology images using deep learning (2017)
16. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)