# Concept attribution: Explaining CNN decisions to physicians ☆

Graziani M. [a,b,*,1], Andrearczyk V. [a], Marchand-Maillet S. [b], Müller H. [a,b,2]

[a] University of Applied Sciences of Western Switzerland Hes-so Valais, Rue de Technopole 3, 3960 Sierre, Switzerland
[b] Department of Computer Science, University of Geneva, Battelle Building A, 7, Route de Drize, 1227 Carouge, Switzerland

## ARTICLE INFO

## ABSTRACT

Deep learning explainability is often reached by gradient-based approaches that attribute the network output to perturbations of the input pixels. However, the relevance of input pixels may be difficult to relate to relevant image features in some applications, e.g. diagnostic measures in medical imaging. The framework described in this paper shifts the attribution focus from pixel values to user-defined concepts. By checking if certain diagnostic measures are present in the learned representations, experts can explain and entrust the network output. Being post-hoc, our method does not alter the network training and can be easily plugged into the latest state-of-the-art convolutional networks. This paper presents the main components of the framework for attribution to concepts, in addition to the introduction of a spatial pooling operation on top of the feature maps to obtain a solid interpretability analysis. Furthermore, regularized regression is analyzed as a solution to the regression overfitting in high-dimensionality latent spaces. The versatility of the proposed approach is shown by experiments on two medical applications, namely histopathology and retinopathy, and on one non-medical task, the task of handwritten digit classification. The obtained explanations are in line with clinicians' guidelines and complementary to widely used visualization tools such as saliency maps.

## 1. Introduction

Deep Neural Networks (DNNs) operate on raw input values and internal neural activations that appear rather incomprehensible to humans [1]. The black-box decision-making process of DNNs is a limiting factor for their deployment in high-risk daily practices, where end-users are not necessarily familiar with deep learning [2–5]. For example, a DNN outputting a strong medical diagnosis can motivate the need for more aggressive treatment, highly impacting the life of the patient. While physicians can provide the reasons for a decision in clinical terms, the network output can only be explained in terms of its internal status. The values of weights, internal layer activations (latents) and input pixels are, however, far from the semantics of the physicians, who focus on affected region size, shape and aspect. Can we align the two representations? Can we find the representation of a concept inside the CNN and use it to interpret its relationship to the network output?

Numerous explainability techniques were developed to generate heatmaps of salient regions [6,7]. Most of these are obtained from the backpropagation of the gradients, for instance, by checking how individual pixel perturbations affect the decision. In medical imaging,

however, perturbations of biomarkers are more meaningful than those of individual pixels. *How would the decision change if there was less stroma in the tissue? What if the nuclei appeared larger and with less regular texture?* Some of these questions were shown as useful to the clinicians using the image retrieval system in [8]. The hallmark of concept attribution, compared to existing explainability tools [1,9–11], is that clinically relevant measures are directly used to produce explanations that match the semantics of the end-users. CNN decisions are explained by directly relating to well-known prognostic factors and clinical guidelines. This promotes a more intuitive interaction between the physicians and black-box systems aiding the diagnosis, with a consequent increase of confidence in automated support tools [8]. Besides, concept attribution is a complementary technique to saliency heatmaps [6,10,11], giving concept-based explanations rather than pixel-based ones.

As a concrete example, one of the applications in this paper focuses on finding tumorous tissue in histopathology slides, a tedious operation to perform manually [12,13]. Clinicians alternate several zoom-in and zoom-out phases to identify prognostic factors such as those in the Nottingham system (NHG), namely a low degree of tubular formation,
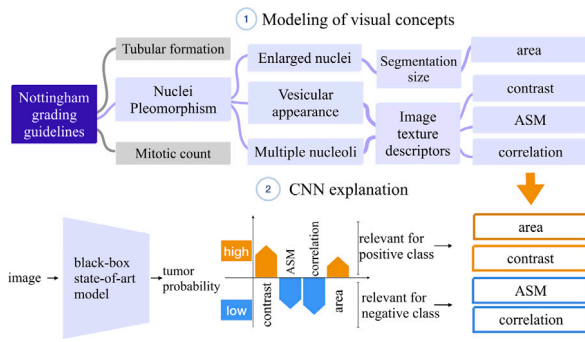
---

**Fig. 1.** Example of concept attribution for a breast cancer classifier. In phase 1, visual concepts are modeled on the basis of well-established guidelines for cancer diagnosis. In phase 2, this knowledge is used to explain the CNN trained to automatically diagnose a tumor.

alterations in the nuclei morphology and high mitotic activity [14]. Convolutional Neural Networks (CNNs) aid the detection process by suggesting tumorous regions, but are they relying on the same criteria to distinguish the tissue abnormalities? This can be investigated by concept attribution as in Fig. 1.

In the first step, indicators of nuclei pleomorphism are modeled as numeric features (top row in Fig. 1). For these features, called *concept measures*, a relevance score is computed that explains their importance in the network output.

Since the explanations are sought only in a post-training phase, our method can potentially be applied to any network without the need of retraining. The interpretability analysis can be applied to the latest state-of-the-art models, without impacting the network performance. Introducing the concept-based explanations in a computer-assisted tumor localization pipeline can improve the interaction between clinicians and the CNN model, for example by explaining why a certain area was highlighted as tumorous (see Fig. 5 in Section 3.5). Moreover, this framework can be helpful in a variety of application fields beyond medicine, as suggested by our experiments on handwritten digits. Other applications are, for example, highlighting the causes of eventual faults in assisted driving or robotic systems. Our primary focus in this paper, however, is the application in the medical context.

The main contribution of this paper is the formulation of a framework for concept-based attribution that generates explanations for CNNs decisions. Besides, we address important limitations of previous works on concept-based explainability [1,15,16]. Our contributions can be summarized as follows:

1. The framework of concept attribution is defined for multiclass classification tasks in Section 3.5.
2. The learning of concepts in the CNN is improved by removing spatial dependencies of the convolutional feature maps and introducing regularization. Experimental results show more accurate RCVs in Section 4.4.1.
3. The well-known dataset of handwritten digits is added to the analysis, broadening the applicative focus. The experiments in Section 4.3 show that the network outcome can be explained by characteristics such as digit shape and extension.

Besides, we test the computational complexity of interpreting the network with RCVs and we present an in-depth discussion about the potential of concept-based explanations for automated diagnosis systems.

## 2. Related work

### 2.1. Attribution to features

Several efforts have been made to unify the definition of deep learning interpretability [2,3]. Clear distinctions were made between models that introduce interpretability as a built-in additional task [4,17–21] and post-hoc methods. While the former may result in decreased performance on the main task due to the interpretability constraint, the latter can explain any machine learning algorithm (including better-than-human networks [1,3,5]) without altering the original performance. Several post-hoc techniques [6,9–11] highlight the most influential set of features in the input space, a technique known as attribution to features [22]. Given a CNN with decision function $f : \mathbb{R}^n \to [0,1]$ and an input image $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, each $a_i$ in the attribution vector $A_f(\mathbf{x}, \mathbf{x}') = (a_1, \ldots, a_n) \in \mathbb{R}^n$ explains the contribution of each pixel $x_i$ to $f(\mathbf{x})$, $i = 1, \ldots, n$. Such contribution is computed with respect to a baseline input $\mathbf{x}'$ with neutral predictions, for example a black image [22]. The attribution method identifies the pixels responsible for the classification of the input image, which is then overlaid with a heatmap. This was successfully applied to the medical field, for example highlighting the contours of colorectal polyps in [23]. The explanations generated by feature attribution are only true for a single input and may change for another data point of the same class. Such point-wise explanations are called *local explanations* [3]. This paper proposes concept-attribution as a complementary technique to feature attribution. Besides, concept-based interpretations that hold true for all inputs of the same class can describe global relationships between input and outputs known as *global explanations*.

### 2.2. Learning concepts inside CNNs

An important step in concept attribution is the learning of concepts in the internal activations of CNNs. Concept learning can be traced back to machine learning theory. Formally, it is defined as the binary classification problem of inferring a Boolean-valued function from input examples of the concept and the relative model output [24]. This was implemented by Concept Activation Vectors (CAV) to interpret the activations of CNNs. The main purpose of CAVs was, in fact, to verify the presence of human-friendly binary concepts (e.g. striped texture) inside CNNs [1]. Linear classifiers were also used as probes to interpret neural activations, being inherently interpretable and thus constituting a baseline of the linear interpretability of deep networks [1,15,16,25]. The performance of the linear classifier is indicative of how well the concept is learned in the network representation. Regression Concept Vectors (RCVs) [15] extended CAVs to model not only the presence or absence of a concept, but also continuous-valued measures. These are important in the medical domain since often the diagnosis is made on the basis of observed measurements, such as tumor growth or patient history, e.g. patient age. The applicability of RCVs to tasks with more than two classes is missing in their original formulation [15,16]. This prevents many future applications, for example on other histological types or medical tasks, e.g. the five-grades Gleason system for prostate cancer. Among others, this important limitation is addressed by the framework presented in this paper.

## 3. Methods

### 3.1. Notation

We first clarify the notation adopted in the paper. We consider a neural network of $L$ layers. The function $f(\mathbf{x})$ is the network output for an input image $\mathbf{x}$. The activation of layer $l$ is $\Phi^l(\mathbf{x})$. For convolutional layers, $\Phi^l(\mathbf{x}) \in \mathbb{R}^{w \times h \times p}$, where $w$ is the width, $h$ the height and $p$ the number of channels. The dataset used to train the network on the main task is $X_{task}$, which is split into training ($X_{task}^{train}$) and testing ($X_{task}^{test}$). Note that $|X_{task}| = N$. The set of class labels $Y_{task}$ is available. In binary classification problems, $y \in \{0, 1\}$ and $f(\mathbf{x}) \in [0, 1]$. In classification problems with classes $k = 1, \ldots, K$, $y$ is a $K$-dimensional one-hot encoding of the class label, and $f(\mathbf{x})$ is a $K$-dimensional vector of the predicted class probabilities. A set of $M$ images $X_{concepts}$, from which it is possible to extract measures of the concepts, is used to learn the
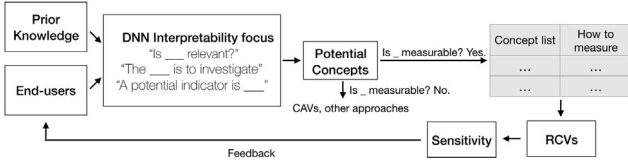
**Fig. 2.** Concept measure selection workflow. End-users such as experts with knowledge in the application domain are asked to provide a set of questions that determine the focus of the interpretability analysis. Similarly, domain-knowledge can be used to identify potential concepts. If a potential concept is measurable from the images, then this is kept in the analysis. If the concept is not measurable, CAVs or other approaches can be adopted.

continuous-valued concepts in the activation space. The set $X_{concepts}$ can be either disjoint or overlap with $X_{task}$. A list of $Q$ concepts is considered for the analysis, e.g. {"area", "contrast", "ASM"}. The function $c_i(\mathbf{x}) \in \mathbb{R}$ measures the value of the $i$th concept in the list on the input image $\mathbf{x}$. Thus, in the list of functions $\{c_1(\cdot), \dots, c_Q(\cdot)\}$, each item corresponds to a different concept, e.g. $c_1 = $ "area".

### 3.2. From expert knowledge to continuous concepts

The starting point for the concept attribution analysis is the formulation of concepts of interest as measurable attributes. This can be done by directly interacting with experts or by referring to the literature. The interaction between ophthalmologists and developers, for example, led to the use of pre-existent handcrafted features describing the appearance of retinal vessels in [16]. In addition to this, the existent guidelines for decision-making are based on many years of study and joint efforts by many experts to develop well-established practices, e.g. the TNM and the Nottingham grading system (NGH) in breast histopathology, the Gleason score in prostate cancer grading or RECIST for tumor grading in radiology. Besides, handcrafted visual features are successfully employed as image descriptors in radiomics [26,27], eye fundus analysis [28] and digital pathology [29]. Our framework exploits this kind of prior information to delimit the focus of the interpretability method, defining a list of $Q$ measurable concepts that should be part of the interpretability analysis. Concepts are chosen so that specific questions can be addressed, for instance by following the workflow in Fig. 2. They can be formulated to verify that domain-knowledge is reflected in the layer activations of the network. In our example for breast histopathology in Fig. 3, the analysis focus is on validating whether the network decisions are in line with the guidelines of clinical practice. A question of interest could be *"Is the nuclei shape relevant to the automatic classification as tumor?"*. Prior expectations on the network behavior can also be validated (e.g. *"Changes in color appearance do not influence the classification"*). The concept measures are computed on a small set of visual examples (i.e. around 30 images or more) that is $X_{concepts}$. For nuclei area and texture, that are representative of the NGH nuclear pleomorphism, the segmentation of the nuclei instances in the image can be obtained by either manual annotations [15] or by automatic segmentation [30]. Nuclei area is expressed as the sum of pixels in the nuclei contours, whereas the nuclei texture is described by Haralick's descriptors such as Angular Second Moment (ASM), contrast and correlation [31].

Some concepts can be chosen to allow cross-application analysis, as the general concept of *area* in handwritten digit recognition (Section 4.3) and in histopathology (Section 4.4). General concepts describing the image such as texture descriptors can be applied in several imaging applications [32]. Other concepts are specific to the type of data being analyzed, as undefined for some data types. RGB color measures, for instance, are undefined for single-channel image modalities, e.g. computed tomography (CT) scans [32]. In addition, the exhaustive evaluation of all possible concepts is unfeasible. For this reason, the selection of concepts is an iterative process that starts from

the more general concepts of texture and appearance and that is then updated with specific requests by the interaction with experts, as shown by the feedback branch in Fig. 2. Once extracted, the concept measures are used to explain the network decisions by CAVs or RCVs.

### 3.3. Attribution to concepts

The feature attribution problem described in Section 2.1 is changed into the problem of evaluating the relevance of each concept to the deep learning classification task. The interpretability analysis is a post-hoc step that does not need the retraining of the network parameters. The network being analyzed is therefore unchanged and it can be replaced at any time by newer architectures with better performance. A vector $\mathbf{v}_c$ representative of the presence or increase of a concept measure is found in the activation space of a layer. The attribution is changed into $A_f(\Phi^l(\mathbf{x}), \{\mathbf{v}_{c_i}\}_{i=1}^Q) = (a_1, \dots, a_Q)$ where each $a_i$ is the relevance of concept $c_i$ to $f(\mathbf{x})$.[3]

### 3.4. Regression concept vectors

RCVs are computed using the output of a CNN layer as input to a regression problem, as illustrated in Fig. 4. We consider the space of the activations of layer $l$, $\Phi^l(\mathbf{x})$. We extract $\Phi^l(\mathbf{x})$ for $\mathbf{x} \in X_{concepts}$. We seek the linear regression that can model the concept $c(\mathbf{x})$ as:

$$c(\mathbf{x}) = \mathbf{v}_c \cdot \Phi^l(\mathbf{x}) + error \qquad (1)$$

where $\mathbf{v}_c$ is the RCV for concept $c$. The RCV components can be found by applying linear least squares (LLS) estimation to $X_{concepts}$. If $l$ is a dense layer, $\mathbf{v}_c$ is a $p$-dimensional vector in the space of its activations. If $l$ is a convolutional layer the output of $\Phi^l(\mathbf{x})$ has spatial and channel dimensions (height, width, channels) represented as $w \times h \times p$. The simplest way of solving LLS in this space is to flatten $\Phi^l(\mathbf{x})$ to a one-dimensional array of $whp$ elements as in [1,15]. The number of dimensions of the unrolled convolutional maps may, however, easily grow to millions. Moreover, the 2D structure of the space is broken by assigning neighboring features to independent dimensions. In this paper, we apply spatial aggregation, i.e. global pooling, along the (height, width) of each feature map to obtain a representation of $\Phi^l(\mathbf{x})$ as a one-dimensional array of $p$ elements. This solution, only briefly mentioned in [25], actually improves the quality of the regression fit by considering the spatial dependencies in the representations.

A further solution proposed in this paper involves a regularization term that is added to the optimization:

$$\mathbf{v}_c^{ridge} = argmin_{\mathbf{v}_c}(\|c(\mathbf{x}) - \mathbf{v}_c\Phi^l(\mathbf{x})\|_2^2 + \lambda\|\mathbf{v}_c\|_2^2) \qquad (2)$$

The penalty term $\lambda$ controls the strength of the regularization. The larger the $\lambda$, the stronger the regularization. The experimental results in this paper compare the two solutions in Eq. (1) and Eq. (2). The RCV represents the direction of the strongest increase of the concept measures for the concept $c$ and it is normalized to obtain a unit vector $\mathbf{v}_c$.

### 3.5. Sensitivity to a concept

The conceptual sensitivity[4] $S_c$ represents how much the concept measure affects the network's output for the input image $\mathbf{x} \in X_{task}$. The sensitivity to a concept was defined for binary concepts in [1]. The same formula is applied to continuous concepts by projecting the derivative on the RCV direction rather than on the CAV direction. For a binary classification task, $S_c^l(\mathbf{x}) \in \mathbb{R}$ is defined as the directional

---

[3] In the following sections, we drop the subscript $i$ for simplicity and we refer to the vector $\mathbf{v}_c$ as the RCV for a concept $c$.

[4] Note that the term conceptual sensitivity was defined in [1] and it does not refer to the output classification sensitivity commonly known as recall.
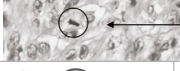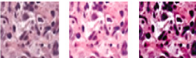
| Concept | Clinical reference | Visual examples | | Computation | Magnification | Source |
|---|---|---|---|---|---|---|
| Count of cavities | NGH tubular formation [11] | well formed | poorly formed | detection and count of cavities | low | annotation or automated |
| Nuclei area | NGH nuclear pleomorphism [11] | regular | enlarged | count of pixels inside nuclei segmentations | high | annotation or automated |
| Nuclei Texture | | | uneven stain | texture descriptors | | |
| Mitotic count | NGH mitotic count [11] | | mitosis | count of detected mitosis | high | annotation or automated |
| Nuclei density | Ki-67 protein expression | regular | overgrowth | count of nuclei | any | annotation or automated |
| Staining | Staining procedure [9] | | different appearance | color descriptors | any | metadata |

Fig. 3. Concept list derived for the breast histopathology application and how to compute the measures.
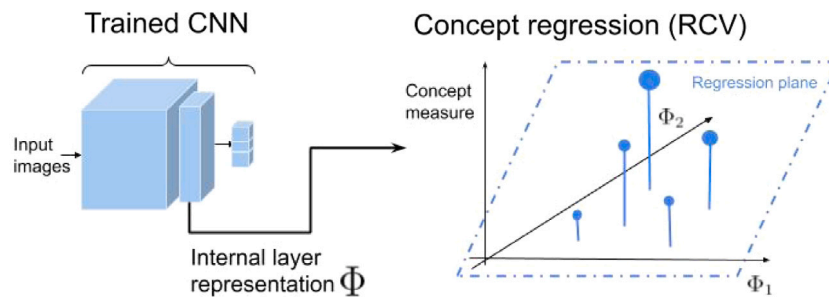


Fig. 4. The output of the CNN internal layer is used to find the RCVs. This does not require the retraining of the CNN parameters. In this two-dimensional example, the RCV is the direction represented by the regression plane. In higher dimensions, unwanted pixel dependencies are removed by an aggregating operation of the internal layer representation.

derivative of the network output $f(\mathbf{x})$ over the RCV direction $\mathbf{v}_c$, computed as a scalar product (see Eq. (3)).

$$S_c^l(\mathbf{x}) = \mathbf{v}_c \cdot \frac{\partial f(\mathbf{x})}{\partial \Phi^l(\mathbf{x})} \qquad (3)$$

$S_c^l(\mathbf{x})$ represents the network responsiveness to changes in the input along the direction of the increasing values of the concept measures. The sign of $S_c^l(\mathbf{x})$ represents the direction of change, while its magnitude represents the rate of change. When moving along the RCV direction, the output $f(\mathbf{x})$ may either increase (positive conceptual sensitivity), decrease (negative conceptual sensitivity) or remain unchanged (conceptual sensitivity equals zero). In a binary classification network with a single neuron in the decision layer, the decision function is a logistic regression over the activations of the penultimate layer. A positive value of the sensitivity to a concept can be interpreted as an increase of $p(y = 1|\mathbf{x})$ when the representation $\Phi^l(\mathbf{x})$ is moved towards the direction of the increasing values of the concept. Negative conceptual sensitivity can be interpreted as an increase in $p(y = 0|\mathbf{x})$ when the same shift in the representation is applied. Conceptual sensitivities scores are informative about the concept influence on the decision for the single input image (as shown, for example, in Fig. 5). These scores are gathered for all inputs of a class by the relevance scores in Section 3.6.

The derivation of the scores for multiclass classification tasks is straightforward. Given the class label $k$, we consider the corresponding $k$th neuron in layer $L$. The neuron activation before softmax, $\Phi^{L,k}(\mathbf{x})$, is a vector of real numbers representing the raw prediction values. These values are then squashed by the softmax into a probability distribution, namely the probability of the label $k$ to be assigned to the input data point $\mathbf{x}$. The conceptual sensitivity score for class $k$ is computed as:

$$S_c^{l,k}(\mathbf{x}) = \mathbf{v}_c \cdot \frac{\partial \Phi^{L,k}(\mathbf{x})}{\partial \Phi^l(\mathbf{x})} \qquad (4)$$

The sensitivity scores can be computed for each class $k$, thus obtaining a vector of $K$ elements. Large absolute values of the conceptual sensitivity for a single class correspond to a strong impact in the decision function when the activations are shifted along the direction of the RCV. In both Eqs. (3) and (4) the derivative of the decision function can be obtained by stopping gradient backpropagation at the $l$th layer of the network. The computational complexity for a single input data point is therefore given by the complexity of the backpropagation operation. In case the backpropagation is done on $(L - l)$ fully connected layers of input size $p$ and output size $d$, the complexity is $O((L - l)dp)$.

### 3.6. Relevance scores

#### 3.6.1. TCAV score

The global relevance of a concept can be computed from the individual sensitivity scores. Previous work on CAVs [1] proposed the TCAV score as the fraction of k-class inputs for which the activation vector of layer $l$ was positively influenced by concept $C$[5]:

$$\text{TCAV} = \frac{|\{\mathbf{x} \in X_k : S_c^{l,k}(\mathbf{x}) > 0\}|}{|X_k|} \qquad (5)$$

where $X_k \subset X_{task}$ is the set of inputs with label $k$. The TCAV score is bounded between zero and one. If there are no images influencing the decision with a positive gradient, TCAV is zero. The TCAV score (computed from the concept sensitivities as in Eq. (5)) is used as a baseline for comparison with the proposed Br scores in our experiments. In the original paper, however, TCAV was only defined for binary concepts [1].

---

[5] The TCAV score for concept $c$ is TCAV$_c$. For simplicity, we drop the subscript $c$ in the rest of the paper for both TCAV and $Br$.
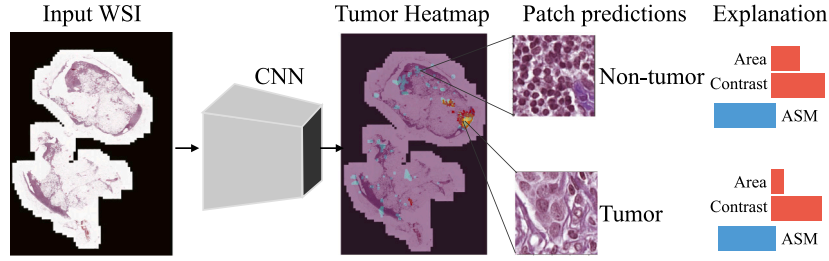
**Fig. 5.** Integration of patch-wise concept-based explanations in the system for assisted diagnosis of breast cancer. The explanations are the concept sensitivities scores for the individual input images. This visualization can help physicians to understand the reasons for a certain classification rather than another. Similarly, developers can inspect and debug the model by looking at misclassification errors and edge-cases.

### 3.6.2. Bidirectional scores

Bidirectional relevance ($Br$) scores were proposed for computing the relevance of concept measures in a binary classification task of histopathology tissue in [15]. $Br$ scores are defined as:

$$Br = R^2 \times \left( \frac{\hat{\mu}}{\hat{\sigma}} \right) \qquad (6)$$

The coefficient of determination $R^2 \leq 1$ indicates how well the RCV represents the concept in the internal CNN activations. It measures whether the concept vector is actually representative of the concept by evaluating their predictive performance on unseen data. The coefficient of variation $\hat{\sigma}/\hat{\mu}$ is the standard deviation of the scores over their average. It describes their relative variation around the mean. Note that $R^2$ evaluates how a concept is present in the internal activations in the form of a linear correlation between the features and the concept values. This, however, is not linked to the final prediction as it does not quantify the influence of the concept for a specific decision. This information is given by the mean average of the concept sensitivity scores $\mu$ in Eq. (6). $Br$ is large when two conditions are met, namely $R^2$ is 1 and the coefficient of variation is small (the values of the sensitivity scores lie closely concentrated near their sample mean). $Br$ explodes, for instance, to infinite if $\hat{\sigma} = 0$. After computing $Br$ for multiple concepts, we scale the scores to the range $[-1, 1]$ by dividing by the maximum absolute value. Such scaling permits a fair comparison among concepts since these are represented by different RCVs. With the set of analyzed concepts being reasonably large, a score close to the absolute value of one can be considered as large. This means that the concept has a considerable impact on the increase (in case of positive sign) of the outcome probability.

Bidirectional scores were not defined in the previous work on RCVs for multi-class classification tasks. We present their extension in the following. Given a concept $c$, the mean and the standard deviation of the sensitivities are computed on the set of inputs belonging to the $k$th class, $X_k$. In multiclass classification, the vector of sensitivities has a value for neuron $m = 1, \ldots, K$ in the decision layer. We use the notation $Br_k^m$ to indicate the bidirectional relevance at the $m$th neuron of the decision layer, for inputs belonging to class $k$. Therefore, given the set of k-class inputs $X_k = \{\mathbf{x}_i\}_{i=1}^{|X_k|}$ we define the mean $\hat{\mu}_k^m$ and standard deviation $\hat{\sigma}_k^m$ of the sensitivities of the $m$th neuron as in Eqs. (7) and (8).

$$\hat{\mu}_k^m = \frac{1}{|X_k|} \sum_{i=1}^{|X_k|} S_c^{l,m}(\mathbf{x}_i) \qquad (7)$$

$$\hat{\sigma}_k^m = \sqrt{\frac{\sum_{i=1}^{|X_k|} (S_c^{l,m}(\mathbf{x}_i) - \hat{\mu}_k^m)^2}{|X_k| - 1}} \qquad (8)$$

$Br_k^m$ is then computed as:

$$Br_k^m = R^2 \times \left( \frac{\hat{\mu}_k^m}{\hat{\sigma}_k^m} \right) \qquad (9)$$

Given a concept measure, $Br_k^m$ represents its relevance in the classification of inputs belonging to class $k$ with respect to the $m - th$ neuron of

the decision layer. We can organize the $Br$ scores in a square matrix $B$ of dimensions $K \times K$:

$$B = \begin{bmatrix} Br_1^1 & \cdots & & Br_K^1 \\ & \ddots & & \\ \vdots & & Br_k^k & \vdots \\ & & & \ddots & \\ Br_1^K & \cdots & & Br_K^K \end{bmatrix}$$

In the $B$ matrix, the rows correspond to the $m$th neuron in the decision layer. The columns correspond to computing the input classes (inputs belonging to the $k$th class). The elements on the diagonal ($Br_k^k$) explain the relevance of the concept measures on test inputs of class $k$ with respect to the neuron responsible for the classification of this same class.

These can be interpreted as the relevance of the concept to the correct classification of each class. The off-diagonal elements measure the relevance of the concept to the softmax output of the neurons that are not responsible for the classification of the true input class $k$. These values can be interpreted as the impact on misclassification of increasing the values of the concept measures. Similarly to the scalar $Br$ for binary classification, the $B$ matrix can be computed for multiple concept measures. The individual matrices can then be concatenated on a third axis to form a tensor with three axes: (neuron, class, concept). To allow a proper comparison between the scores for multiple concepts, the third axis can be scaled in the range $[-1, 1]$.[6] The computational complexity of the $Br$, as for TCAV, is a function of the number of samples used to evaluate the sensitivity scores. If only the testing split is used, the complexity is $O(|X_{task}^{test}|)$, which has to be multiplied by the complexity of computing the sensitivity scores of each point. Hence the final complexity, for a network with $L$ fully-connected layers of input size $p$ and output size $d$, is $O(|X_{task}^{test}|(L - l)dp)$.

### 3.6.3. Alternative scores

TCAV and $Br$ evaluate the influence of a concept on the model's decision for all images of an input class, gathering the information given by the individual conceptual sensitivity scores. The global concept influence cannot be explained by the $R^2$, as this evaluates only the predictive performance of the RCV on test data, giving a measure of how the RCV is representative of the concept. The TCAV and $Br$ score add information to the analysis, by considering different properties of the sensitivity values. TCAV estimates whether the concept increases the probability of a class in the decision function. This is done by counting the number of positive directional derivatives for inputs of class $k$ and a given concept. $Br$ scores introduce information about the magnitude and variation of the influence of the concept measures. Besides, Br scores are informative on the influence of a concept to the direction of the decision, namely by showing if the increase of a concept measure results in an increase or decrease in the likelihood of a specific class.

---

[6] The scaling is performed as a division of the $Br$ for one concept over the maximum value of the scores for all concepts.

Different scores can explore other characteristics of the gradients such as the largest variation of the gradient (i.e. with a max operation on the directional derivatives) or the ratio between positive and negative derivatives. One example is the layer-agnostic metric proposed in [27], which allows comparing scores across all the network layers.

## 4. Experiments and results

### 4.1. Architectures

We use the following architectures in the experiments:

- A multilayer perceptron (MLP) with one hidden layer of 512 nodes is used to introduce the methodology on binary and multiclass classification tasks. For the former model, logistic regression and binary cross-entropy loss are used. For the latter, we introduce an additional hidden layer of 512 nodes and use softmax and categorical cross-entropy.
- For the breast cancer histopathology application, we use a ResNet101 pretrained on ImageNet [33]. The last layer of the network is replaced by a single node with a sigmoid activation for binary classification. High-resolution patches of tumor regions are distinguished from patches of nontumor regions.[7]
- For the retinopathy application, the last layer of InceptionV1 (pretrained on ImageNet) is replaced with a dense layer with softmax activation, which is fine-tuned for the classification of three classes, namely *normal*, *pre-plus* and *plus* [34].[8]

### 4.2. Datasets

We report the datasets used for the experiments. We first introduce the method on the classification of handwritten digits as a very simple example to explain the concepts. We then extend the experiments on two medical applications, namely the classification of tumor patches in histopathology images and the classification of the plus disease in ROP images.

#### 4.2.1. Handwritten digits

The MLP was trained on the dataset of handwritten digits MNIST [35]. The concept measures are automatically extracted from the digit images that are binarized by applying a threshold of 0.5. From the resulting binary maps, we extract concept measures describing the shape of the digit, such as *eccentricity* (deviation of a curve from circularity), *perimeter* (length of the digit contour) and *area* (number of pixels in the digit). The input images to the network are the original images and not the binary masks.

#### 4.2.2. Breast histopathology

Three datasets are used for the breast histopathology experiments. Two of them, namely Camelyon16 and Camelyon17[9] are used to fine-tune the decision layer of ResNet101. More than 40,000 patches at the highest resolution level are extracted from random locations of the Whole Slide Images (WSIs) in Camelyon16 and 17 and used as $X_{task}$. We use only the WSIs for which the annotation of the tumor area is given. Staining normalization and online data augmentation (random flipping, brightness, saturation and hue perturbation) are used to reduce the domain shift between the different centers. A dataset with manual segmentation of the nuclei [36][10] is used to extract the concept measures and learn the regression. This dataset contains WSIs

of several organs with more than 21,000 annotated nuclei boundaries. From this data, we select the WSIs of breast tissue, from which we extract 300 patches which constitute the $X_{concepts}$. Concept measures describing the morphology of the nuclei are extracted from the manual segmentation of the nuclei. For these patches, the labels of tumor or non-tumor regions are not available.

#### 4.2.3. Retinopathy of prematurity

Images from a private dataset of 4800 de-identified posterior retinal images constitute the $X_{task}$ for the application on Retinopathy of Prematurity (ROP). A commercially available camera was used to capture the images (namely RetCam; Natus Medical Incorporated, Pleasanton, CA). A total of 3024 images (1084 for *normal*; 1074 for *pre-plus*; 1080 for *plus*) were used as the training split $X_{task}^{train}$. The testing split $X_{task}^{test}$ contains 985 samples (817 for *normal*; 148 for *pre-plus*; 20 for *plus*). The high class imbalance between *plus* and *normal* cases is due to the fact that ROP is a disease with a low prevalence (only 3%). The preprocessing pipeline of the images is as in Brown et al. [37].[11] A CNN is used to segment the retinal vasculature. After segmentation, the images are resized to 224 x 224 pixels and data augmentation is applied, i.e. right-angle rotations and horizontal and vertical flipping. $X_{concepts}$ is built by gathering the samples of class *plus* and *normal* from $X_{task}^{train}$. Samples of class *pre-plus* represent the transition from *normal* to *plus* and can be excluded from $X_{concepts}$ to keep the size of this dataset smaller than $X_{task}^{train}$. The interpretability analysis is performed on $X_{task}^{test}$.

### 4.3. Experiments on MNIST digits

Experiments on MNIST were performed because it is a widely studied dataset in computer vision that can be used as a well-controlled problem to evaluate the principles of our method and to verify that expected results are obtained. In the first part of this section, we consider the binary classification of zero and one digits, a trivial task to show the application of concept attribution on networks with a single decision node. In the second part, we extend the application to all classes, showing the application of multiclass scores and the comparison with the TCAV baseline. We select two concept measures that are complementary to each other by construction, such as *count of black pixels* ($n_{black}$) and *count of white pixels* ($n_{white}$). These measures are perfectly linearly correlated since their summation is equal to the total number of pixels in the image. Hence, in such a controlled experiment we expect to find two parallel RCVs that point in opposite directions. As expected, the angle between the two RCVs is indeed 180 degrees.

We expand the analysis with concept measures of digit *area, perimeter, eccentricity*. The distribution of the concept measures between the zero and ones classes is shown in Fig. 6. In Fig. 7, we visualize the t-Distributed Stochastic Neighbor Embedding (t-SNE) projection of the representations of the two classes with the corresponding values of the concept measures for the concept *eccentricity* [38]. Measures of eccentricity are larger for inputs of class one. Pearsons' correlation coefficient between *eccentricity* and network output is 0.84 while, for *perimeter*, it is $-0.96$ ($p$-value $< 0.0001$ for both). The RCVs of *eccentricity* and *perimeter* form an angle of 174 degrees. The $R^2$ of the RCVs are 0.91 for *eccentricity*, 0.91 for *perimeter* and 0.95 for *area*. By contrast, the regression of an irrelevant concept such as random values of the concept measures returns non-positive or zero values of the $R^2$.

The *Br* scores (second row in Table 1) show that an increase in the *eccentricity* shifts the prediction towards the one class, while an increase in the *perimeter* or *area* shifts the decision towards the zero class. The TCAV scores only identify eccentricity as a relevant concept.

The next experiments show the extension from binary to multiclass classification, with all digit classes being considered. The Pearsons' correlation coefficient between the network neurons and the concept

---

[7] We refer to the first convolutional layer as conv1 and to the merge layers at the end of residual blocks of increasing depth as res2a, res2b, etc.

[8] We refer to the first convolutional layer as conv1, and to the filter concatenation layers of increasing depth as Mixed3b, Mixed4b, etc.

[9] Downloadable at https://camelyon17.grand-challenge.org/.

[10] Dataset available at https://monuseg.grand-challenge.org/Data/.

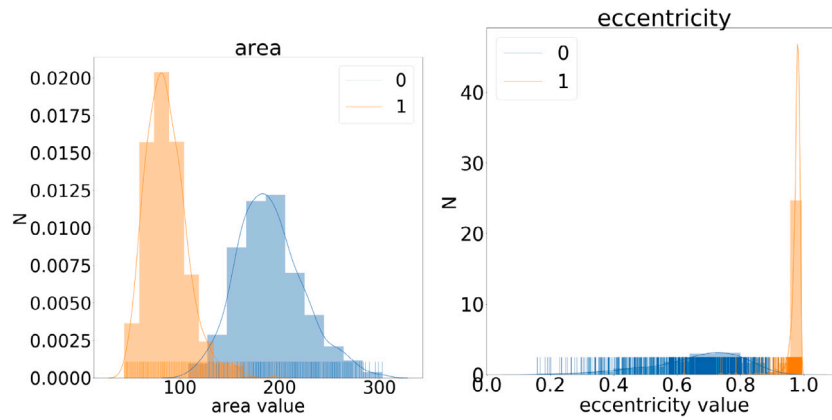[11] Source code: https://github.com/QTIM-Lab/qtim_ROP.

**Fig. 6.** Distribution plots of the concept measures *area, and eccentricity* for the zero and one MNIST digits. Concept measures of *perimeter* show a close distribution to the one for *area*.
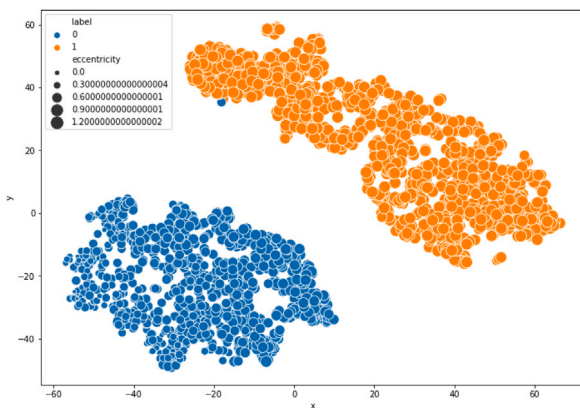


**Fig. 7.** t-SNE projection of the concept measures for *eccentricity* on zero and one inputs from MNIST digits. Best seen in color.
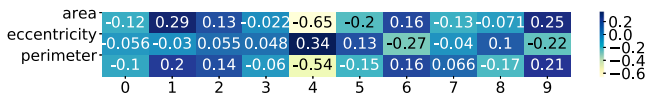


**Fig. 8.** Pearsons' correlation coefficient of the concept measures *area, eccentricity and perimeter* and the network logits (*p*-value < 0.0001). Best seen in color.

**Table 1**
$R^2$ and $Br$ scores for the binary classification of handwritten zeros and ones.

| Concept | Area | Eccentricity | Perimeter |
|---------|------|--------------|-----------|
| $R^2$   | 0.95 | 0.91 | 0.91 |
| $Br$    | −1.0 | 0.92 | −0.92 |
| TCAV    | 0    | 1.0  | 0 |

measures is shown in Fig. 8. The $R^2$ are 0.86, 0.88, and 0.95 for the RCVs of respectively *eccentricity, perimeter* and *area*. Relevance scores are computed for the concept *eccentricity* on a test set of 30 samples. The B matrix is compared to the confusion matrix in Fig. 9. A comparison with the TCAV baseline is in Fig. 10.

*4.4. Experiments on medical data*

We report the experiments on breast cancer histopathology and ROP. This analysis presents additional experiments on: (1) applying global pooling to $\Phi^l(\mathbf{x})$; (2) evaluating the RCVs with $R^2_{adj}$, and (3) using ridge regression. These experiments address the technical limitations of the method in [15]. Moreover, they prove the applicability of RCVs to multiple architectures, datasets and tasks.
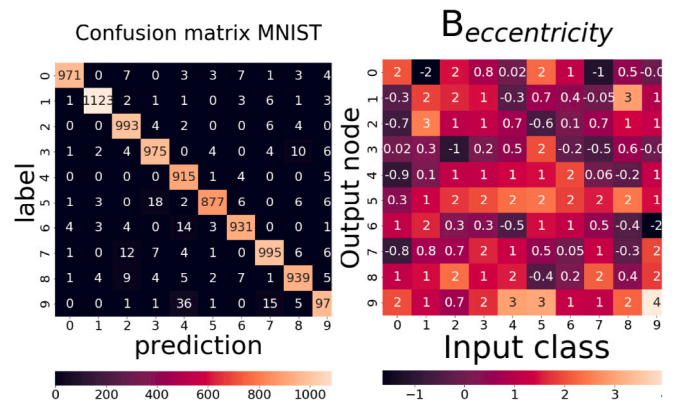


**Fig. 9.** Confusion matrix and $B$ matrix comparison for the multiclass classification of MNIST digits. In the $B$ matrix each cell is $Br_k^m$. The rows correspond to the $m$th neuronal activation and the columns to input of class $k$. Best seen in color.

The accuracy of ResNet101 is close to the state-of-the-art of the challenge, at 92%. More details on the network training and performance are described in [15]. Six concept measures, namely *area, eccentricity, Euler number, contrast, Angular Second Moment (ASM)* and *correlation*, are extracted from patches for which manual segmentations of the nuclei are available (more details in [15]). The concepts are selected in order to mirror aspects that are considered by the grading system conventions for breast cancer. The first four rows in Table 2 show the $R^2$ of the RCVs for the best performing concepts, namely *area, contrast, ASM* and *correlation*. The concepts *eccentricity* and *Euler* are excluded from the remaining analysis because they are not learned successfully in the activation space, as shown in [15]. These concepts were, in fact, not appropriate to the task. Our method, however, highlights non-robust concepts by the evaluation of the RCVs. For instance, both *eccentricity* and *Euler* had low $R^2$ over multiple repetitions on validation data, with broad confidence intervals. The comparison between $Br$ scores and the TCAV baseline is shown in Table 3.

For the ROP application, the input images of the network are binary masks of the segmentation of the vessels in the retina, obtained following the pipeline in [39]. Handcrafted visual features used in state-of-the-art machine learning approaches for ROP classification [28] are used as concepts. Six concept measures are selected from 143 hand-crafted features of vessel curvature, tortuosity and dilation, namely *curvature mean* (mnCURV), *curvature median* (mdCURV), *avg point diameter mean* (mnAPD), *avg segment diameter median* (mnASD), *cti mean* (mnCTI) and *cti median* (mdCTI) (more details in [16]). In our previous work [16], we compute the $R^2$ of the RCVs at different layers in the
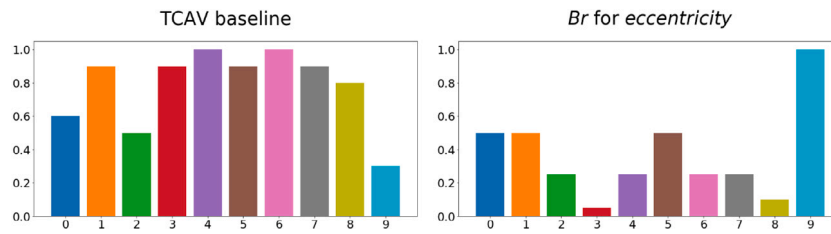
**Fig. 10.** TCAV baseline and $Br$ scores for each class in MNIST digits (represented as numbers), giving an outlook of the influence of *eccentricity* on the CNN. Best seen in color.

**Table 2**
Impact of global pooling on the $R^2$ of the RCVs for breast histopathology. The pooling strategy is on the top left of each block. The labels in the other columns refer to the CNN layers, as in the Keras implementation of ResNet50.

| no pooling | conv1 | res2a | res2b | res2c | res3a | res4a | res5a |
|---|---|---|---|---|---|---|---|
| Area | 0.32 | 0.32 | 0.36 | 0.36 | 0.43 | 0.47 | 0.46 |
| Contrast | 0.37 | 0.36 | 0.37 | 0.34 | 0.37 | 0.45 | 0.43 |
| ASM | 0.28 | 0.29 | 0.31 | 0.26 | 0.38 | 0.44 | 0.50 |
| Correlation | 0.33 | 0.35 | 0.32 | 0.35 | 0.41 | 0.42 | 0.48 |
| max pool | conv1 | res2a | res2b | res2c | res3a | res4a | res5a |
| Area | 0.34 | 0.00 | 0.10 | 0.06 | 0.46 | 0.60 | 0.60 |
| Contrast | 0.24 | 0.0 | 0.27 | 0.21 | 0.33 | 0.52 | 0.63 |
| ASM | **0.49** | **0.24** | 0.28 | 0.24 | 0.52 | **0.65** | **0.70** |
| Correlation | 0.43 | 0.19 | **0.53** | **0.54** | **0.58** | **0.65** | 0.64 |
| avg pool | conv1 | res2a | res2b | res2c | res3a | res4a | res5a |
| Area | 0.0 | 0.0 | 0.15 | 0.24 | 0.03 | 0.32 | 0.52 |
| Contrast | 0.0 | 0.0 | 0.34 | 0.18 | 0.02 | 0.42 | 0.57 |
| ASM | 0.0 | 0.0 | 0.18 | 0.39 | 0.28 | 0.52 | 0.62 |
| Correlation | 0.0 | 0.0 | 0.35 | 0.34 | 0.18 | 0.54 | 0.62 |

**Table 3**
$Br$ and TCAV scores for breast histopathology.

|  | Area | Contrast | ASM | Correlation |
|---|---|---|---|---|
| TCAV | 0.34 | 0.72 | 0.05 | 0.01 |
| $Br$ | 0.10 | 0.27 | −0.53 | −1 |

**Table 4**
$R^2$ and $R^2_{adj}$ for breast histopathology (conv1).

| max pool | Area | Contrast | ASM | Correlation |
|---|---|---|---|---|
| $R^2$ | 0.34 | 0.24 | 0.49 | 0.43 |
| $R^2_{adj}$ | 0.16 | 0.04 | 0.35 | 0.28 |

**Table 5**
Comparison of $R^2$ and $R_a dj^2$ for the ROP concepts in [16]. The pooling strategy is on the top left of each block. The labels of the other columns refer to the layers of Inception V1 [34].

| max pool | conv1 | Mixed3b | Mixed4b | Mixed4c | Mixed5c |
|---|---|---|---|---|---|
| mdCTI $R^2$ | 0.59 | 0.66 | 0.64 | 0.63 | 0.67 |
| mdCTI $R^2_{adj}$ | 0.58 | 0.66 | 0.63 | 0.63 | 0.66 |
| mnCTI $R^2$ | 0.49 | 0.56 | 0.50 | 0.47 | 0.56 |
| mnCTI $R^2_{adj}$ | 0.48 | 0.56 | 0.49 | 0.46 | 0.56 |
| mdCURV $R^2$ | 0.65 | 0.72 | 0.69 | 0.67 | 0.71 |
| mdCURV $R^2_{adj}$ | 0.64 | 0.72 | 0.69 | 0.67 | 0.71 |
| mnCURV $R^2$ | 0.65 | 0.70 | 0.61 | 0.57 | 0.72 |
| mnCURV $R^2_{adj}$ | 0.64 | 0.70 | 0.61 | 0.57 | 0.71 |
| mnASD $R^2$ | 0.55 | 0.66 | 0.58 | 0.56 | 0.64 |
| mnASD $R^2_{adj}$ | 0.54 | 0.65 | 0.57 | 0.56 | 0.64 |
| mdAPD $R^2$ | 0.69 | 0.76 | 0.69 | 0.66 | **0.76** |
| mnASD $R^2_{adj}$ | 0.68 | 0.76 | 0.69 | 0.65 | **0.76** |
| avg pool | conv1 | Mixed3b | Mixed4b | Mixed4c | Mixed5c |
| mdCTI $R^2$ | 0.68 | 0.75 | 0.70 | 0.72 | 0.72 |
| mdCTI $R^2_{adj}$ | 0.67 | 0.75 | 0.71 | 0.70 | 0.69 |
| mnCTI $R^2$ | 0.56 | 0.63 | 0.54 | 0.55 | 0.56 |
| mnCTI $R^2_{adj}$ | 0.55 | 0.62 | 0.53 | 0.55 | 0.56 |
| mdCURV $R^2$ | 0.62 | 0.73 | 0.75 | 0.76 | 0.71 |
| mdCURV $R^2_{adj}$ | 0.61 | 0.73 | 0.75 | 0.76 | 0.71 |
| mnCURV $R^2$ | 0.65 | 0.74 | 0.68 | 0.69 | 0.71 |
| mnCURV $R^2_{adj}$ | 0.64 | 0.74 | 0.68 | 0.69 | 0.71 |
| mdASD $R^2$ | 0.69 | 0.74 | 0.67 | 0.67 | 0.64 |
| mdASD $R^2_{adj}$ | 0.68 | 0.73 | 0.67 | 0.67 | 0.64 |
| mnAPD $R^2$ | **0.72** | **0.80** | **0.76** | **0.77** | 0.76 |
| mnAPD $R^2_{adj}$ | 0.71 | **0.80** | **0.76** | **0.77** | **0.76** |

network only with examples of one single class at a time. The results for the classes *plus* and *normal* are provided in [16]. In this paper, we combine the images of class *normal* and *plus* in a single set of examples $X_{concept}$, from which we regress the concept measures. The class *pre-plus* was excluded to keep the size of $X_{concept}$ to a smaller size than the whole training set $X_{task}$. The $R^2$ is shown in Fig. 11 (on the left).

In the following, we report the experimental results obtained by extending the original framework.

### 4.4.1. Global pooling of the features

We compare the results of the aggregation of the features at the spatial level at different layers. For the histopathology application, the results of the regression on the flattened feature vectors are shown in the first four rows of Table 2 and compared to the results with average-pooling and max-pooling (last eight rows).

For the ROP application, the results of max- and average-pooling are compared to the results without pooling in Table 5 and Fig. 11.

### 4.4.2. Adjusted $R^2$

The $R^2$ is compared to $R^2_{adj}$, which penalizes unnecessary variables. The latter shows how much variation is explained by more than one independent variable in the regression model. Table 4 shows the $R^2_{adj}$ of the RCVs computed on the pooled features of the first convolutional layer (conv1) for the histopathology application (with 300 samples and 64-dimensional data). For the other layers of the network, the dimensionality of $\Phi^l(\mathbf{x})$ is much larger than the number of samples used to learn the RCV and it is not possible to consider $R^2_{adj}$ a valid statistic. Table 5 shows the values of $R^2_{adj}$ of the RCVs for the ROP application. Global pooling was applied to the convolutional filters (max-pooling is shown in the top rows, average-pooling in the bottom rows).

### 4.4.3. Regularized regression

We introduce the regularization of the L2 norm of the RCV by using ridge regression. Tables 6 and 7 show the $R^2$ of $\mathbf{v}_C^{ridge}$ for the histopathology and ROP applications. The regularization term $\lambda$ is tuned with grid-search over the range $[0, 10^6]$, as shown in Fig. 12.

## 5. Discussion

The experiments show the versatility of concept attribution in handling different tasks. The first experiment in Section 4.3 analyzes the RCVs in a controlled setting, i.e. the binary classification of zero and one handwritten digits. As expected, the RCVs for the complementary pixel counts (when $n_{black}$ increases, $n_{white}$ decreases) are parallel with opposite pointing directions, forming an angle of 174 degrees.

This result reflects the organization of *eccentricity* measures in Fig. 6, which is also kept in the latent space as shown in Fig. 7. RCVs, however, seem to more intuitively represent the direction of increase of the concept measures than the direct visualization of the latents.
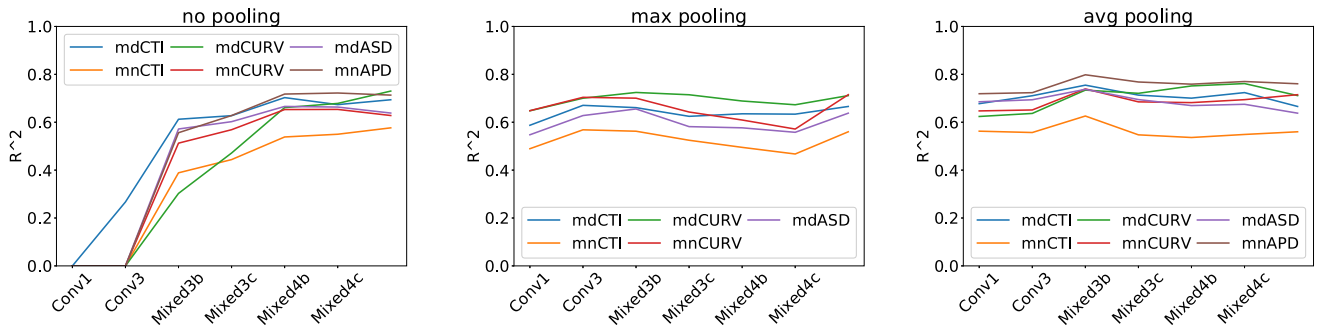
**Fig. 11.** $R^2$ of the RCVs for the regression of concepts of *curvature (mdCURV and mnCURV), dilation (mdASD, mnAPD) and tortuosity (mdCTI and mnCTI)* in ROP images of class *normal* and *plus*. The lines for each clinical factor are specifically representing the mean (mn-) and median values (md-). Best seen in color.
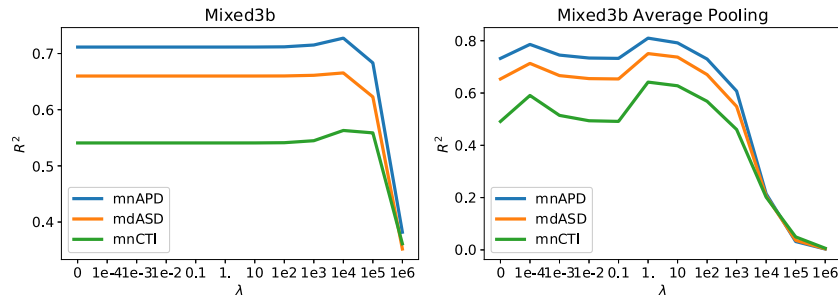


**Fig. 12.** Impact of $\lambda$ on the ridge regression with (on the left) and without (on the right) global average pooling for the ROP concepts as in [16]. The pooling operation reduces the need for regularization and leads to higher values of $R^2$. The lines represent respectively dilation (blue and orange lines) and tortuosity (green line). For clarity reasons, only a subset of ROP concepts is shown, representing dilation (mnAPD and mdASD) and tortuosity (mnCTI) as in [16]. Best seen in color.

**Table 6**
Comparison of unregularized and ridge regression for mnAPD and mdASD on ROP. $\lambda$ is set to $10^4$ for unpooled features and to 1 for pooled features.

| Mixed3b, mnAPD | Unregularized | L2 |
|---|---|---|
| no pooling | 0.54 | 0.63 |
| avg pooling | 0.73 | 0.81 |
| Mixed3b, mdASD | Unregularized | L2 |
| no pooling | 0.56 | 0.59 |
| avg pooling | 0.65 | 0.75 |

**Table 7**
$R^2$ of ridge regression on histopathology application with global pooling ($\lambda = 10^2$). Each column refers to the layers of ResNet50.

| | res3a | res4a | res5a |
|---|---|---|---|
| ASM | 0.64 | 0.66 | 0.71 |
| Correlation | 0.64 | 0.67 | 0.68 |

Fig. 10 compares *Br* and TCAV scores for the concept *eccentricity* in the classification of the ten digit classes. This is an additional contribution to [15], where only binary classification problems were considered. The *Br* scores on the diagonal of the *B* matrix in Fig. 9 show the relevance of the concept *eccentricity* for the output node matching the input class, while the off-diagonal values consider the nodes that are different from the input class. These scores can help developers to analyze the influence of a concept to misclassification errors.

The concept *eccentricity*, for example, influences the misclassification of the inputs of class three. An image of an *eccentric* three, i.e. stretched along the vertical axis, is more likely to be misclassified by the model as a nine, a seven or a five. This insight can be used to reduce the importance of eccentricity when classifying this digit, for example by learning adversarial representations to the concept eccentricity. Future applications in medical imaging could use this to introduce concept-based adversarial learning to remove the influence of the domain-shifts in medical imaging data.

Besides, the extensions to the work in [15,16] improve the quality of the RCVs on both medical applications, as shown by Tables 2 and 5 and Fig. 11. The pooling operation leads to more robust RCVs for all concepts, reducing also the need for regularization (see Table 6 and Fig. 12). Ridge regression further improves the $R^2$ in both applications. The adjusted determination coefficients in (Table 4) shows that correlated variables in the regression should be removed by an additional dimensionality reduction, further motivating our pooling of the feature maps. The large size of the ROP dataset, however, seems to lead to smaller difference between the $R^2_{adj}$.

Concept attribution can give multiple benefits to automated decision support tools. Nor the model accuracy nor the AUC can explain why the CNNs predicts certain region as cancerous and not another, for example. Conceptual sensitivity scores can explain, without any need of model retraining, why the CNN assigns an input image to a certain class. Br scores, moreover, give a global picture of the influence of clinically relevant factors on the model decisions. This can reduce the black-box perception of CNN and promote their integration in daily practice. The concept-based explanations of different CNNs could be compared by their *Br* scores to see if diagnostic factors have the same relevance in different architectures or parameter initializations. Physicians could use concept-based explanations to verify whether the correct prognostic factors are used for the decision and to inspect if cause–effect links are maintained in the CNN. Finally, a list of clinical concepts that the CNN should prioritize can be used in future work to introduce concept learning as an extra-task. This could improve the model performance and generalization on unseen data.

## 6. Conclusions

This paper presents an in-depth analysis of the interpretability framework of concept attribution for deep learning, which is complementary to the widely used heatmaps of salient pixels. Differently from previous works [1,15,22], concept-based explanations are used to quantify the contribution of features of interest to the network's

decision-making. Physicians can use this tool to compare their procedures and guidelines to the network's internal process and evaluate whether factors that are generally relevant in their practice are also used to influence the network decision. Explanations in terms of pre-existing guidelines in the applicative field can help to reduce the gap between the representation of the task in the deep network and in the mind of domain experts. This method lays at the frontier of medical sciences and computer scientists, aiming at bridging two different worlds with the purpose of making the automated solutions of deep learning more understandable and less intimidating for physicians. Our framework can be plugged-in on top of an existent network to verify that prior beliefs are mirrored by the network decisions. Undesired behavior can be spotted and new hypotheses can be tested. For example, this method could highlight if the presence of watermarks and text annotations, often present in medical images, is affecting the network decision and thus corrupting the learning process. This, among others, could be an important check before deploying a CNN for daily practice. Future developments can automatically extract concepts, following the line of work of [40,41]. The optimization function, moreover, can be further modified to amplify or reduce the attention given to particular concepts.

Adversarial training can be used to discard information about unwanted concepts, for example. In addition, a user-evaluation analysis can be performed to measure the impact of the generated explanations in the clinical daily-routine.

## CRediT authorship contribution statement

**Graziani M.:** Conceptualization, Formal analysis, Methodology, Software, Writing - original draft. **Andrearczyk V.:** Conceptualization, Formal analysis, Methodology. **Marchand-Maillet S.:** Conceptualization, Validation, Supervision. **Müller H.:** Conceptualization, Methodology, Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: International Conference on Machine Learning, 2018, pp. 2673–2682.
[2] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R.M. Rao, et al., Interpretability of deep learning models: a survey of results, in: IEEE Smart World Congress 2017 Workshop: DAIS, 2017.
[3] Z.C. Lipton, The mythos of model interpretability, Commun. ACM 61 (10) (2018) 36–43, http://dx.doi.org/10.1145/3233231.
[4] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1721–1730.
[5] B. Goodman, S. Flaxman, European Union regulations on algorithmic decision-making and a "right to explanation", AI Mag. 38 (3) (2017) 50–57.
[6] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013, CoRR abs/1312.6034.
[7] P.J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K.T. Schütt, S. Dähne, D. Erhan, B. Kim, The (Un)reliability of saliency methods, 2017, CoRR abs/1711.00867.
[8] C.J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G.S. Corrado, M.C. Stumpe, et al., Human-centered tools for coping with imperfect algorithms during medical decision-making, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 2019, p. 4.
[9] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, 2013, CoRR abs/1311.2, URL: http://arxiv.org/abs/1311.2901.
[10] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: ICCV, 2017, pp. 618–626.
[11] R.C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.
[12] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, et al., 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset, GigaScience 7 (6) (2018).
[13] Y. Liu, K. Gadepalli, M. Norouzi, G.E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P.Q. Nelson, G.S. Corrado, et al., Detecting cancer metastases on gigapixel pathology images, 2017, arXiv preprint arXiv:1703.02442.
[14] H. Bloom, W. Richardson, Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years, Br. J. Cancer 11 (3) (1957) 359, http://dx.doi.org/10.1038/bjc.1957.43.
[15] M. Graziani, V. Andrearczyk, H. Muller, Regression concept vectors for bidirectional explanations in histopathology, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, 2018.
[16] M. Graziani, J. Brown, V. Andrearczyck, V. Yildiz, J.P. Campbell, D. Erdogmus, S. Ioannidis, M.F. Chiang, J. Kalpathy-Kramer, H. Muller, Improved interpretability for computer-aided severity assessment of retinopathy of prematurity, in: Proceedings Volume 10950, Medical Imaging 2019: Computer-Aided Diagnosis, 2019, http://dx.doi.org/10.1117/12.2512584.
[17] A.A. Freitas, Comprehensible classification models: a position paper, ACM SIGKDD Explorations Newsl. 15 (1) (2014) 1–10, http://dx.doi.org/10.1145/2594473.2594475.
[18] B. Kim, J.A. Shah, F. Doshi-Velez, Mind the gap: A generative approach to interpretable feature selection and extraction, in: Advances in Neural Information Processing Systems, 2015, pp. 2260–2268.
[19] K. Cho, A. Courville, Y. Bengio, Describing multimedia content using attention-based encoder-decoder networks, IEEE Trans. Multimed. 17 (11) (2015) 1875–1886, http://dx.doi.org/10.1109/TMM.2015.2477044.
[20] D. Alvarez-Melis, T.S. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc., 2018, pp. 7786–7795.
[21] S. Shen, S.X. Han, D.R. Aberle, A.A. Bui, W. Hsu, An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification, Expert Syst. Appl. 128 (2019) 84–95, http://dx.doi.org/10.1016/j.eswa.2019.01.048, URL: http://www.sciencedirect.com/science/article/pii/S0957417419300545.
[22] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, pp. 3319–3328.
[23] K. Wickstrøm, M. Kampffmeyer, R. Jenssen, Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps, Med. Image Anal. 60 (2020) 101619, http://dx.doi.org/10.1016/j.media.2019.101619, URL: http://www.sciencedirect.com/science/article/pii/S1361841519301574.
[24] T.M. Mitchell, Machine Learning, first ed., McGraw-Hill, Inc., New York, NY, USA, 1997.
[25] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, 2016, CoRR abs/1610.01644.
[26] A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, et al., The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology (2020) 191145, http://dx.doi.org/10.1148/radiol.2020191145.
[27] H. Yeche, J. Harrison, T. Berthier, UBS: A dimension-agnostic metric for concept vector interpretability applied to radiomics, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer, 2019, pp. 12–20.
[28] E. Ataer-Cansizoglu, V. Bolon-Canedo, J.P. Campbell, A. Bozkurt, D. Erdogmus, J. Kalpathy-Kramer, S. Patel, K. Jonas, R.P. Chan, S. Ostmo, et al., Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the "i-ROP" system and image features associated with expert diagnosis, Transl. Vis. Sci. Technol. 4 (6) (2015) 5, http://dx.doi.org/10.1167/tvst.4.6.5.
[29] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, A. Madabhushi, Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features, J. Med. Imaging 1 (3) (2014) http://dx.doi.org/10.1117/1.JMI.1.3.034003.

[30] S. Otálora, M. Atzori, A. Khan, O. Jimenez-del Toro, V. Andrearczyk, H. Müller, A systematic comparison of deep learning strategies for weakly supervised Gleason grading, in: Medical Imaging 2020: Digital Pathology, Vol. 11320, International Society for Optics and Photonics, 2020, p. 113200L.

[31] R.M. Haralick, I. Dinstein, K. Shanmugam, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (6) (1973) 610–621, http://dx.doi.org/10.1109/TSMC.1973.4309314.

[32] M. Graziani, H. Müller, V. Andrearczyk, Interpreting intentionally flawed models with linear probes, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, Statistical Deep Learning for Computer Vision, 2019.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[35] Y. LeCun, C. Cortes, The MNIST database of handwritten digits, 1998.

[36] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, A. Sethi, A dataset and a technique for generalized nuclear segmentation for computational pathology, IEEE Trans. Med. Imaging 36 (7) (2017) 1550–1560, http://dx.doi.org/10.1109/TMI.2017.2677499.

[37] J.M. Brown, J.P. Campbell, A. Beers, K. Chang, S. Ostmo, R.P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, J. Kalpathy-Cramer, et al., Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks, JAMA Ophthalmol, http://dx.doi.org/10.1001/jamaophthalmol.2018.1934.

[38] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.

[39] J.M. Brown, J.P. Campbell, A. Beers, K. Chang, K. Donohue, S. Ostmo, R.P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, et al., Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning, in: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, Vol. 10579, International Society for Optics and Photonics, 2018, p. 105790Q.

[40] B. Zhou, Y. Sun, D. Bau, A. Torralba, Interpretable basis decomposition for visual explanation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 119–134.

[41] Z. Zhang, Q. Wu, Y. Wang, F. Chen, High-quality image captioning with fine-grained and semantic-guided visual attention, IEEE Trans. Multimed. 21 (7) (2019) 1681–1693, http://dx.doi.org/10.1109/TMM.2018.2888822.

**Mara Graziani** is a third-year PhD student with double affiliation at the computer science faculty at the University of Geneva and at the University of Applied Sciences of Western Switzerland (Hes-so). Her research aims at improving the interpretability of machine learning systems for healthcare. She was a visiting student at the Martinos Center, part of Harvard Medical School in Boston, MA, USA to focus on the interaction between clinicians and deep learning systems. From a background of IT Engineering, she was awarded the Engineering Department Award for completing the MPhil in Machine Learning, Speech and Language at the University of Cambridge (UK) in 2017.



**Vincent Andrearczyk** received a double Masters degree in electronics and signal processing from ENSEEIHT, France and Dublin City University, in 2012 and 2013 respectively. He completed his PhD degree on deep learning for texture and dynamic texture analysis at Dublin City University in 2017. He is currently a post-doctoral researcher at the University of Applied Sciences and Arts Western Switzerland with a research focus on deep learning for medical image analysis and texture feature extraction.



**Prof. Stéphane Marchand-Maillet** has obtained his PhD in Applied Mathematics at Imperial College (London, UK) in 1997. He is Associate Professor in the Department of Computer Science at University of Geneva since 2011. His research group (Viper) specializes in large-scale, high-dimensional distributed machine learning and information retrieval, mining and indexing, with applications to data modeling and prediction, including social network analysis. He has authored, co-authored or edited a number of publications on these topics. He and his group are part of several national and European and international projects in the domain. He is Senior PC Member of the International Joint Conference on AI (IJCAI, one of the oldest established conferences in AI). He was general co-chair of the International Conference of the ACM-SIG on Information Retrieval in 2010 and general co-chair of the 16th IEEE Conference in Business Informatics in 2014.



**Henning Müller** studied medical informatics at the University of Heidelberg, Germany, then worked at Daimler-Benz research in Portland, OR, USA. From 1998–2002 he worked on his PhD degree at the University of Geneva, Switzerland with a research stay at Monash University, Melbourne, Australia. Since 2002, Henning has been working for the medical informatics service at the University Hospital of Geneva. Since 2007, he has been a full professor at the HES-SO Valais and since 2011 he is responsible for the eHealth unit of the school. Since 2014, he is also professor at the medical faculty of the University of Geneva. In 2015/2016 he was on sabbatical at the Martinos Center, part of Harvard Medical School in Boston, MA, USA to focus on research activities. Henning is coordinator of the ExaMode EU project, was coordinator of the Khresmoi EU project, scientific coordinator of the VISCERAL EU project and is initiator of the ImageCLEF benchmark that has run medical tasks since 2004. He has authored over 500 scientific papers with more than 13,000 citations and is in the editorial board of several journals.