




Concept Basis Extraction for Latent Space Interpretation of Image Classifiers

Alexandros Doumanoglou^{1,2}^a, Dimitrios Zarpalas¹^b and Kurt Driessens²^c

¹Information Technologies Institute, Centre for Research and Technology Hellas, 1st Km Charilaou - Thessaloniki, Thessaloniki, Greece

²Department of Advanced Computing Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands
{aldoum, zarpalas}@iti.gr; kurt.driessens@maastrichtuniversity.nl

Keywords: Concept Basis, Interpretable Basis, Unsupervised Learning, Explainable AI, Interpretability, Computer Vision, Deep Learning

Abstract: Previous research has shown that, to a large extent, deep feature representations of image-patches that belong to the same semantic concept, lie in the same direction of an image classifier’s feature space. Conventional approaches compute these directions using annotated data, forming an interpretable feature space basis (also referred as concept basis). *Unsupervised Interpretable Basis Extraction (UIBE)* was recently proposed as a novel method that can suggest an interpretable basis without annotations. In this work, we show that the addition of a classification loss term to the unsupervised basis search, can lead to bases suggestions that align even more with interpretable concepts. This loss term enforces the basis vectors to point towards directions that maximally influence the classifier’s predictions, exploiting concept knowledge encoded by the network. We evaluate our work by deriving a concept basis for three popular convolutional networks, trained on three different datasets. Experiments show that our contributions enhance the interpretability of the learned bases, according to the interpretability metrics, by up-to +45.8% relative improvement. As additional practical contribution, we report hyper-parameters, found by hyper-parameter search in controlled benchmarks, that can serve as a starting point for applications of the proposed method in real-world scenarios that lack annotations.

1 Introduction


A crucial finding of post-hoc explainable artificial intelligence (XAI) in computer vision, is that, in standard convolutional image classifier architectures (Simonyan and Zisserman, 2015; Szegedy et al., 2015; He et al., 2016), deep representations of image patches corresponding to the same human understandable concept (e.g. image patches of a *cat face*, a *car window*, or patches of a *wall*) lie on the same feature space direction (Szegedy et al., 2013; Kim et al., 2018; Zhou et al., 2018). Furthermore, quantitatively, this property is more explicit towards the top layers of the networks (Alain and Bengio, 2016).


The conventional approach to discover concept directions in the feature space of a pre-trained convolutional neural network (CNN) requires annotated data (Zhou et al., 2018; Kim et al., 2018). Each concept direction is defined by a *concept vector* which coincides


with the normal of the hyperplane that separates representations of images (or image-patches) that depict the concept from representations of images (or image-patches) that depict other (negative) concepts. This approach is applicable alike, whether the annotations are available at the image-level (Kim et al., 2018) or at pixel level (Zhou et al., 2018).

In (Zhou et al., 2018), the authors used the learned set of concept vectors, to construct an interpretable feature space basis (here also referred as *concept basis*). A concept basis can provide answers to several questions regarding the CNN and its predictions (Doumanoglou et al., 2023). For instance, it can be used to explain the relationship between concepts and filters (Fong and Vedaldi, 2018), to provide local explanations by interpreting predictions of individual examples (Zhou et al., 2018), or to provide global explanations by quantifying the class sensitivity of the CNN with respect to a concept (Kim et al., 2018).

Apart from the conventional approaches (Kim et al., 2018; Zhou et al., 2018), recently a novel, unsupervised, post-hoc, method was proposed that is able

^a <https://orcid.org/0000-0002-4337-1720>

^b <https://orcid.org/0000-0002-9649-9306>

^c <https://orcid.org/0000-0001-7871-2495>

to derive and suggest a concept basis without the need of annotations (Doumanoglou et al., 2023). In that case, the search for a concept basis is guided based on the explanation that the basis provides for deep representations. In particular, it was shown that projecting a representation to all the concept vectors of an interpretable basis and hard-thresholding, leads to sparse binary representations and the search for the concept basis is guided by this criterion.

To suggest an interpretable basis, prior work (Doumanoglou et al., 2023) exploits structure in the feature population of the studied CNN. However, the suggested *concept vectors* are not explicitly linked to the prediction strategy of the studied network. Here, we argue that an explicit link between the concept vectors and the CNN classifier’s output has three benefits. First, in case the network makes predictions based on human interpretable features, the suggested basis’ interpretability can be improved by this link. Second, in case the network uses a concept that is not known to humans, this link can reveal the use of this new concept which can be understood by inspecting dataset samples that maximally activate the concept vector (Kim et al., 2018). And third, in case the network is *cheating* through spurious correlations in the data, this link can expose this malfunction. Overall, linking the basis search with the CNN classifier’s output produces a basis that can provide explanations related to the prediction strategy of the model. Eventually, the interpretation that this basis can provide can be used to encourage trust to the model or debug it.

In this work, we make a simple, yet elegant, three-fold impactful extensions to (Doumanoglou et al., 2023). First, we empirically found that the number of hyper-parameters of the *Inactive Classifier Loss (ICL)* in (Doumanoglou et al., 2023) does not contribute much to the interpretability of the learned bases and thus, here we propose a simplification that reduces their count. Second, and most important, we propose a loss term that guides the search of concept vectors towards the directions that have stronger impact on the CNN’s prediction outcomes and quantitatively justify that the bases learned with this term score better in the interpretability metrics. Third, we provide indicative hyper-parameter values for the proposed method, that were found by extensive hyper-parameter search. Those values were experimentally found to work best in our benchmarks (for which annotations are available for quantitative evaluation) and may serve as a future reference for applying the proposed method in real-world cases, when annotations are absent. Finally, as a last contribution, inspired from (Zarlenga et al., 2023), we propose an additional basis interpretability metric which we also use for ba-

sis interpretability evaluation.

2 Background

2.1 Problem Statement

Considering an intermediate layer of a **pre-trained** CNN, let $D \in \mathbb{N}^+$ denote the dimensionality of its feature space. Let $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$ denote the representation of an image in this layer, and $\mathbf{x}_p \in \mathbb{R}^D$ a spatial element of this representation at location $\mathbf{p} = (x, y), x \in \{0, 1, \dots, W - 1\}, y \in \{0, 1, \dots, H - 1\}$. Previous work (Alain and Bengio, 2016; Zhou et al., 2018; Kim et al., 2018) has shown that, based on their semantic label, deep representations (either \mathbf{X} or \mathbf{x}_p), especially for layers near the top, are - up to a large extend - linearly separable. Based on this fact, (Zhou et al., 2018; Doumanoglou et al., 2023) were motivated to construct a feature space basis (referred as *interpretable basis* or *concept basis*) where each vector in the basis, points towards the direction of the feature space where representations of a human-understandable concept lies. Once an interpretable basis is learned, it can be used to attribute human-understandable meaning to intermediate representations \mathbf{x}_p , by projecting \mathbf{x}_p onto the interpretable basis. In both approaches (Zhou et al., 2018; Doumanoglou et al., 2023), the learned basis is extracted in a **post-hoc** manner by analyzing feature populations, and regards a specific layer of the pre-trained CNN image classifier, while the weights of the CNN classifier itself, are kept frozen. In this work, we contribute solution improvements to the problem of unsupervised concept basis extraction, by suggesting simple and elegant, yet impactful, adjustments to (Doumanoglou et al., 2023).

2.2 Supervised Approach to Concept Basis Extraction

Let $\mathbf{w}_i \in \mathbb{R}^D, i \in I, I = \{0, 1, \dots, I - 1\}$ denote a set of $I \leq D, I \in \mathbb{N}^+$ concept vectors that form a concept basis for the previously mentioned feature space. In the standard supervised approach to concept basis extraction (Zhou et al., 2018), each concept vector \mathbf{w}_i is learned by training a linear classifier $\{\mathbf{w}_i, b_i\}, b_i \in \mathbb{R}$ (also referred as *concept detector* (Bau et al., 2017; Doumanoglou et al., 2023)) with the objective to separate representations of examples depicting a concept (e.g. representations of patches depicting “car”) from representations of examples depicting other (negative) concepts (e.g. representations of patches depicting “wall”, “person”, “sky”, etc). For instance, after successfully training a concept-detector for the concept “car”, the vector \mathbf{w}_i , which coincides with the separating hyper-plane’s normal, points towards the

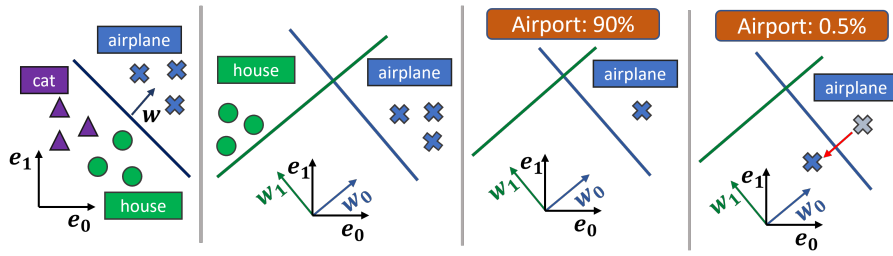


Figure 1: From left to right: a) An interpretable direction \mathbf{w} points towards the representations of a concept in the CNN’s feature space. The position of the separating hyper-plane delimits the subspace where those representations lie. b) UIBE finds a rotation of the feature space $\{\mathbf{w}_i\}$ along with the positions $\{b_i\}$ of the separating hyper-planes, such as, under the new rotated basis, each feature representation is classified positively by just one (or a few) of the classifiers $\{\mathbf{w}_i, b_i\}$. c) and d) In this work we complement UIBE with an additional loss term that enforces the new basis to point towards directions that maximally influence CNN’s predictions. An intermediate representation is manipulated towards the negative direction of the basis vector that classifies it positively, in an amount sufficient to surpass the position of the classifier’s separating hyper-plane, and require that under this manipulation the prediction of the CNN becomes maximally uncertain.

feature space direction where the representation of the concept (“car”) lies. Thus, projecting an arbitrary \mathbf{x}_p on \mathbf{w}_i measures *how much car* is contained in the representation, while in that case, the bias b_i , delimits *point zero*, i.e. the minimum *projected quantity* that is required in order to attribute the label “car” to \mathbf{x}_p .

For the supervised approach to work, the availability of a concept-dataset containing dense per-pixel annotations is required. Correspondence between the pixel-level annotations, which are provided in the image input space, and the deep image representations at the various spatial locations (x, y) is established via calculations related to the receptive-field of each spatial feature \mathbf{x}_p .

2.3 Unsupervised Approach to Concept Basis Extraction

Unsupervised Interpretable Basis Extraction (UIBE) (Doumanoglou et al., 2023), has been proposed as a novel method to suggest a concept basis without the need of expensive annotations. At its core, UIBE utilizes the same concept detector model (binary classifiers $\{\mathbf{w}_i, b_i\}$) as the supervised approach, with the additional constraint that the vectors \mathbf{w}_i form an orthonormal basis (for the reasons behind this, the reader is advised to refer to (Doumanoglou et al., 2023)). For each concept detector in the basis, the sigmoid classifier rule $y_{p,i} = \sigma(s_i(\mathbf{w}_i^T \mathbf{x}_p - b_i)) \in (0, 1)$, $s_i \in \mathbb{R}^+$, $\|\mathbf{w}_i\|_2 = 1$ classifies the representation \mathbf{x}_p positively whenever $y_{p,i} > 0.5$.

UIBE is learning a concept basis for a CNN’s intermediate layer, by analyzing the feature population $\mathbf{x}_p \forall p$ in a concept dataset. During basis learning, and due to the lack of annotations, the name of the concept that each classifier can detect is unknown. This name is going to be determined later, using a procedure to label the basis, as described in Section 2.4. Since

annotations are absent, training individual concept-detectors is impossible. Instead of that, UIBE takes a holistic approach, by considering the set of concept-detectors in entirety. The basis that UIBE suggests, satisfies a sparsity property that an interpretable basis meets (Doumanoglou et al., 2023). In particular, in UIBE, basis search is driven by the observation that the semantic label(s) of each image-patch is (are) just one (or just a few) over a plethora of other possible semantic labels. Under the assumption that the pre-trained studied CNN, has managed to disentangle the different semantic concepts, this implies that for the representation \mathbf{x}_p of each image-patch, only a fraction of the concept-detectors (binary classifiers) associated with the concept basis would classify each \mathbf{x}_p positively for the presence of the basis’ concepts. This implies sparsity on the L1-normalized vectors $\mathbf{y}_p = [y_{p,0}, y_{p,1}, \dots, y_{p,l-1}]$. For a graphical illustration please refer to Figure 1.

To find a concept basis, UIBE utilizes four kinds of losses. The first one, is the **Sparsity Loss (SL)** which, for the reasons described above, dictates entropy minimization on the L1-normalized $y_{p,i}, \forall i$. The second one, is the **Maximum Activation Loss (MAL)**, which enforces the most dominant positive classification j , ($j : y_{p,j} > y_{p,i} \forall i \neq j$) to be the most confident prediction in an absolute sense, i.e. $y_{p,j} \rightarrow 1.0$. **SL** together with **MAL** guide basis search towards directions where each \mathbf{x}_p is attributed only a small fraction of concept labels, compared to the plethora of concept labels that the basis vectors are associated with. The third one, is the **Inactive Classifier Loss (ICL)** which ensures that each one of the concept-detectors in the basis, is actually meaningful, by classifying positively a minimum amount of \mathbf{x}_p in the concept dataset. Finally, the **Maximum Margin Loss (MML)** minimizes s_i , and thus im-

poses the largest possible margin around the separating concept-detector’s hyper-plane. In the rest of the paper we say that a pixel \mathbf{p} is **explained by the basis**, when $y_{\mathbf{p},i} > 0.5$ for at least one i . We also say that a pixel \mathbf{p} is **not explained** by the basis when $y_{\mathbf{p},i} < 0.5 \forall i$, i.e. none of the concept detectors in the basis classify pixel \mathbf{p} positively, as a sample of a concept. Finally, we only consider $I = D$ (Sections 2.1, 2.2), leaving study and experiments for other values of I for future work.

2.4 Basis Labeling

The term *basis labeling* refers to the process of attributing a concept label name to each one of the vectors in a concept basis. This process is applicable only to the bases learned in an unsupervised way, since for the bases that were learned in a supervised manner, the label of each concept vector is known even before training the respective concept-detector. In cases where annotations are available, the basis labeling process can be accomplished in a systematic way. Basis labeling methods (Bau et al., 2017; Mu and Andreas, 2020) can be used in tandem with an annotated concept dataset, such as Broden (Bau et al., 2017) and Broden Action (Ramakrishnan et al., 2019).

In this work we use (Bau et al., 2017) to label the bases extracted with the proposed method. In particular, let $\phi_i(c, \mathcal{K}) \in [0, 1]$ denote a metric score function that is used to measure the *suitability* of the classifier i ($\{\mathbf{w}_i, b_i\}$) to accurately detect concept c in the annotated concept dataset \mathcal{K} . The basis vector i is assigned the concept label c^* , which is the label of the concept dataset that maximizes $\phi_i(c, \mathcal{K}_{rain})$, among all c , in the training split \mathcal{K}_{rain} of the concept dataset. For the choice of ϕ we use Intersection Over Union (IoU), as originally proposed in (Bau et al., 2017) and also used in (Fong and Vedaldi, 2018; Mu and Andreas, 2020; Doumanoglou et al., 2023)

2.5 Basis Interpretability Metrics

To measure the interpretability of a basis, we actually need to measure how well each concept-detector performs in classifying positively the concept’s representations, and negatively, the representations of concept counter-examples (Bau et al., 2017; Zhou et al., 2018). Thus, an **annotated** concept dataset is **required**.

In (Doumanoglou et al., 2023) two basis interpretability metrics were proposed that were based on ideas from other previous approaches (Bau et al., 2017; Losch et al., 2021). In this work, we use those exact same metrics, which are, subsequently, described briefly. Using the validation split of the concept dataset \mathcal{K}_{val} , each classifier $\{\mathbf{w}_i, b_i\}$ is as-

signed a validation score $\phi_i(c_i^*, \mathcal{K}_{val})$, with c_i^* denoting the concept label assigned to the classifier during the previously mentioned basis labeling procedure. Then the two interpretability scores \mathcal{S}^1 and \mathcal{S}^2 are defined as:

$$\mathcal{S}^1 = \int_0^1 \sum_{i=0}^{I-1} \mathbb{1}_{x \geq \xi}(\phi_i(c_i^*, \mathcal{K}_{val})) d\xi \quad (1)$$

$$\mathcal{S}^2 = \int_0^1 \psi(\xi) d\xi \quad (2)$$

with $\mathbb{1}(x)$ denoting the indicator function. The first metric \mathcal{S}^1 , counts the number of concept detectors in the basis with a validation score better than a threshold ξ . In the second metric \mathcal{S}^2 , $\psi(\xi)$ is defined as $\psi(\xi) = |\{c_i^* | \exists i : \phi(i, c^*, \mathcal{K}_{val}) \geq \xi\}|$, i.e. the number of unique concept detectors exhibiting performance better than ξ (Bau et al., 2017). In both cases, we make the scores, threshold agnostic, by integrating across all $\xi \in [0, 1]$, as it was proposed in (Losch et al., 2021; Doumanoglou et al., 2023).

2.6 Other Related Work

The proposed approach belongs in the same area as other concept-based, post-hoc explainability methods that do not require annotations for the discovery of concepts encoded by a network. Although our approach builds upon a certain line of research (Szegedy et al., 2013; Alain and Bengio, 2016; Bau et al., 2017; Zhou et al., 2018; Kim et al., 2018; Doumanoglou et al., 2023), which studies concepts as directions in a CNN classifier’s feature space, there are other, recent, approaches that study concepts from different viewpoints. For instance, in (Zhang et al., 2021), a concept vector points to the concept cluster’s center and (Vielhaben et al., 2022) uses spectral clustering to find meaningful subspaces in the CNN’s feature space. Other than that, (Achtibat et al., 2023) extends (Bach et al., 2015), by effectively turning a pixel attribution method, to a concept discovery method. Finally, (Chormai et al., 2022) discovers disentangled concept subspaces by taking into account the relevance of neurons to the CNN’s prediction outcomes. Our work is faithful to finding concepts that are relevant to the CNN’s prediction outcome. In that sense, our work also shares common ground with (Achtibat et al., 2023) and (Chormai et al., 2022) although approaching concept discovery in a different way.

3 Proposed Method

3.1 Inactive Classifier Loss (ICL) Simplification

In UIBE, the Inactive Classifier Loss (ICL) was proposed as a loss term that would enforce the concept

Table 1: a) Hyper-parameter names, bounds, step and initial value that were used in hyper-parameter tuning. The table also provides the best values (suggestions) returned by the optimizer. λ^s , λ^{ma} , λ^{ic} , λ^{mm} , and λ^{cc} are the loss weight parameters for Sparsity Loss, Max Activation Loss, Inactive Classifier Loss, Maximum Margin Loss and CNN Classifier Loss. τ refers to the hyper-parameter of equation 3. b) Comparing interpretabilities of the natural feature space basis, the basis extracted with UIBE and the basis extracted with the proposed method for the latent space of ResNet50. UIBE’s scores are set as the reference scores against which relative percentage scores are calculated. The loss term introduced in this work leads to learning a basis that is significantly more interpretable than the reference, across all metrics, and especially in terms of S^2 .

(a)							(b)			
	λ^s	λ^{ma}	λ^{ic}	λ^{mm}	τ	λ^{cc}	ResNet50 / MiT			
Lower	1.0	1.0	1.0	0.1	0.5	0.05	Basis	S^1 (\uparrow)	S^2 (\uparrow)	S^3 (\downarrow)
Upper	10.0	10.0	10.0	10.0	1.0	0.5	Natural	90.82 (-27.1%)	16.24 (-12.0%)	0.023 (+0.44%)
Step	0.5	0.5	0.5	0.5	0.1	0.01	UIBE	124.73 (+0.0%)	18.47 (+0.0%)	0.0229 (+0.0%)
Init	2.0	5.0	5.0	0.5	1.0	0.2	Proposed	131.73 (+5.61%)	26.94 (+45.8%)	0.0225 (-1.75%)
	ResNet18 / Places365									
Best	2.6	2.8	4.8	0.6	0.9	0.25				
	VGG16BN / ImageNet									
Best	3.7	6.7	5.0	1.0	0.95	0.05				

detectors in the basis to *classify positively* (Section 2.3) a non-trivial amount of \mathbf{x}_p . In ICL’s absence, some of the concept detectors in the basis might point towards directions where none of \mathbf{x}_p is classified positively, introducing redundancy and a less useful basis suggestion, since SL can be easier be fulfilled when $y_{p,i} \rightarrow 0$, for most i .

Experiments with UIBE have shown that the quality of the extracted bases are mostly insensitive to the intuitive choice of hyper-parameters for ICL. Thus, in this work, we opt to reduce the complexity of the approach and simplify ICL, by removing the notion of partitions. Let $\tau \in (0, 1]$ denote a percentage over the number of pixels p in the concept dataset, implicitly controlling the minimum amount of pixels to be *explained by the basis* (Section 2.3). Let also $v_i = \tau/I$ and $\gamma > 1, \gamma \in \mathbb{R}^+$ a sharpening factor.

For the inactive classifier loss we use the following equation:

$$\mathcal{L}^{ic} = \mathbb{E}_{i \in I} \left[\frac{1}{v_i} \text{ReLU}(v_i - \mathbb{E}_{\mathbf{p}}[y_{p,i}^\gamma]) \right] \quad (3)$$

which is the same formula as in (Doumanoglou et al., 2023) but with a different definition for v_i , eliminating the hyper-parameters α_μ, ω_μ and the intermediate variables n_μ . Essentially, under the marginal case where $y_{p,i}$ is close to zero or one (as urged by the combination of SL and MAL), \mathcal{L}^{ic} equals zero, when each one of the concept detectors classify positively at least v_i percent of the pixels \mathbf{x}_p and equals one in the case none of the concept detectors classifies positively any of them.

3.2 CNN Classifier Loss (CCL)

The concept vectors in a basis derived with (Doumanoglou et al., 2023) are not necessarily related to the CNN’s predictions. However, as mentioned in

Section 1, this work proposes that such a link could improve the suggested basis’ interpretability, reveal new concepts exploited by the network, or aid model debugging. Here we focus on the first benefit, while leaving the study of the other two for future work.

The motivation of our second contribution is based on the fact that, in the absence of annotations, we may try to exploit the knowledge of concepts that are encoded into the network, to aid the discovery of interpretable directions. Our hypothesis is that, in many cases, interpretable directions might maximally influence the classifier’s predictions. For instance, the classification of an image as the scene “park” might be influenced by the presence of concepts “person”, “tree” and “bench”. Thus, the latent representation of an image depicting a “park” could have strong components across the directions of those interpretable concepts. In that case, and as a consequence, we would expect that if we manipulate the representation of the “park” image by attenuating its components across the previously mentioned concept directions, the classifier would find it hard to classify the image and thus its prediction would become very much uncertain.

More formally, we introduce the CNN Classifier Loss (CCL), as a complement to the losses of UIBE. The principal idea behind CCL is to manipulate each of the representations \mathbf{x}_p , in such a way, that the resulting representation \mathbf{x}'_p is not explained by the basis (Section 2.3), i.e. $\mathbf{y}'_{p,i} < 0.5 \forall i$ and require that the CNN’s prediction for the manipulated image representation \mathbf{X}' becomes maximally uncertain. Let $d_{p,i} = \mathbf{w}_i^T \mathbf{x}_p - b_i$ denote the signed distance of \mathbf{x}_p to the i -th concept’s separating hyperplane. On the one hand, if $d_{p,i}$ is positive, this means that \mathbf{x}_p is a sample of concept i . Thus, according to our principal idea, we need to attenuate the presence of this concept in \mathbf{x}_p by manipulating the representation in the direction $-\mathbf{w}_i$

and for a distance equal to $d_{p,i}$. On the other hand, if $d_{p,i}$ is negative, this implies that the concept i is not present in \mathbf{x}_p and we do not need to perform any manipulation to remove it. Let also $\alpha_{p,i} = \sigma(s'd_{p,i})$ a coefficient in $(0,1)$ that approaches 1 whenever we need to perform a manipulation in \mathbf{x}_p for concept i . The hyper-parameter $s' \in \mathbb{R}^+$ is considered fixed and its meaning is similar to s that was defined in Section 2. The manipulation \mathbf{x}'_p is given by the following formula:

$$\mathbf{x}'_p = \mathbf{x}_p - \mathbf{W}\mathbf{v} \quad (4)$$

with the columns of \mathbf{W} being equal to the concept vectors of the basis and $\mathbf{v} \in \mathbb{R}^D$ a vector with elements $v_i = \alpha_{p,i}d_{p,i}$. Let the function f^+ denote the part of the CNN after the layer of study. If $f^+(\mathbf{X}') \in (0,1)^K$ denotes the final vector of probabilities of the CNN, the CNN Classifier Loss is defined as:

$$\mathcal{L}^{cc} = -\mathbb{E}_{\mathbf{X}'} \mathcal{H}(f^+(\mathbf{X}')) \quad (5)$$

with \mathcal{H} denoting the entropy of the CNN's prediction vector for the $K \in \mathbb{N}^+$ classes. An illustration of the procedure is provided in Figure 1.

3.3 Basis Impurity Score

Inspired from the Oracle Impurity Score (OIS) that was introduced in (Zarlenga et al., 2023), and to save computational time, this work proposes a third basis interpretability metric, that complements the previous two metrics described in Section 2.5. This metric aims to capture how distinguished is the suitability of each classifier $\{\mathbf{w}_i, b_i\}$ in the basis, in detecting the concept associated with its assigned label. More formally, let $A \in [0,1]^{I \times C}$ a matrix with elements $a_{i,c} = \phi_i(c, \mathcal{K}_{val})$ with $c \in \mathcal{C}$ and \mathcal{C} denoting the set of concept labels in the annotated concept dataset. Let $P \in \{0,1\}^{I \times C}$ the matrix with elements $p_{i,c} = 1$ if $c = c_i^*$ and 0 otherwise. The **Basis Impurity Score** is defined as:

$$\mathcal{S}^3 = \frac{1}{\sqrt{I|C|}} \|P - A\|_F \quad (6)$$

In eq (6) $|C|$ denotes the number of concepts in \mathcal{C} and $\|\cdot\|_F$ the Frobenius norm of a matrix. A value of zero for \mathcal{S}^3 implies that the concept detectors associated with the basis do not share suitability in detecting concepts other than the concept of their assigned label, while values closer to 1 imply less exclusivity in this suitability.

4 Experimental Results

4.1 Experimental Setup

Overall Evaluation Approach We choose to evaluate the bases extracted with the proposed method

based on the three basis interpretability metrics that were described in Sections 2.5 and 3. For the evaluation, we derive bases for the last convolutional layers (after ReLU) of three different CNN models: ResNet18 (He et al., 2016) trained on Places365 (Zhou et al., 2017), VGG16BN (Simonyan and Zisserman, 2015) (VGG16 with batch normalization layers) trained on ImageNet (Deng et al., 2009) and ResNet50 (He et al., 2016) trained on Moments in Time (MiT) (Monfort et al., 2019). Overall, we follow the evaluation approach that was described in (Doumanoglou et al., 2023). First we learn the bases using the training split of the respective concept dataset. Subsequently, we use (Bau et al., 2017) to label the bases using the training split of the same dataset and finally we use its validation split to compute concept detector scores and the basis evaluation metrics. For the networks trained on Places365 and ImageNet we used the Broden (Bau et al., 2017) concept dataset, while for the network trained on MiT we used Broden Action (Ramakrishnan et al., 2019).

Parameter Initialization In all of our experiments we initialize the basis vectors with the vectors of the natural feature space basis (Doumanoglou et al., 2023). We also deviate from (Doumanoglou et al., 2023) regarding the parameterization of margin and bias. In this work, we use exponential parameterization (e^t) for both margin $M = 1/s$ and bias b (in the standardized space). Both M and b , are initialized to $t = \log(0.5)$. We experimentally found that the exponential parameterization stabilizes learning.

Hyper-parameters The efficacy of the proposed method, like UIBE's, relies on minimizing the individual loss terms with the right balance. The proposed method linearly combines five individual loss terms (SL, MAL, ICL, MML, CCL) with coefficients λ . In order to evaluate the potential of the proposed approach, we perform two hyper-parameter tuning experiments, by exploiting the availability of annotations in the concept datasets. For these experiments we take the 70% of the Broden's training set and further split it in 70%-30% train/validation splits. For the validation metric we use \mathcal{S}^2 . The hyper-parameter tuning experiments regard bases learned for ResNet18 and VGG16BN. We don't perform any hyper-parameter tuning regarding bases learned for ResNet50. For hyper-parameter optimization we use the Nevergrad (Rapin and Teytaud, 2018) optimization platform. We perform tuning in steps. In the first step, we use the Two Points Differential Evolution (TwoPointsDE) algorithm with budget 100, to tune all parameters except λ^{cc} which is set to zero. (This is the same as tuning UIBE, with the modification of the hereby proposed ICL loss). Subsequently,

Table 2: Comparing interpretabilities of the natural feature space bases and the bases extracted with either UIBE or the proposed method. We set the metric scores extracted with UIBE, as the references against which relative percentage scores are calculated. According to all the metrics, the loss term introduced in the proposed work consistently leads to basis extraction of improved interpretability compared to UIBE, both for bases extracted for ResNet18 trained on Places365 as well as for bases extracted for VGG16BN trained on ImageNet.

ResNet18 / Places365				VGG16BN / ImageNet		
Basis	\mathcal{S}^1 (\uparrow)	\mathcal{S}^2 (\uparrow)	\mathcal{S}^3 (\downarrow)	\mathcal{S}^1 (\uparrow)	\mathcal{S}^2 (\uparrow)	\mathcal{S}^3 (\downarrow)
Natural	35.52 (-41.7%)	18.26 (-35.6%)	0.0272 (+5.0%)	34.72 (-27.0%)	11.86 (+1.4%)	0.0271 (+2.2%)
UIBE	60.93 (+0.0%)	28.39 (+0.0%)	0.0259 (+0.0%)	47.58 (+0.0%)	11.69 (+0.0%)	0.0265 (+0.0%)
Proposed	69.43 (+13.9%)	31.53 (+11.6%)	0.0256 (-1.1%)	48.96 (+2.9%)	12.11 (+2.9%)	0.0264 (-0.3%)

Table 3: Comparing Network Dissection results for all bases of ResNet18 and ResNet50. Two numbers are reported for each concept category. The first being the number of concept detectors that can detect a concept from the category, and second, the number of (unique) concepts from the category that can be detected by the detectors.

ResNet 18 / Places365							
Basis	Object	Part	Scene	Material	Texture	Color	Action
Natural	116 / 45	10 / 7	261 / 121	2 / 2	50 / 26	0 / 0	-
UIBE	248 / 53	21 / 10	117 / 110	8 / 6	28 / 21	1 / 1	-
Proposed	174 / 46	11 / 8	258 / 135	7 / 6	33 / 25	1 / 1	-
ResNet50 / MiT							
Natural	295 / 35	20 / 5	126 / 43	1 / 1	357 / 27	0 / 0	336 / 86
UIBE	289 / 39	49 / 4	332 / 42	8 / 3	665 / 39	99 / 3	140 / 73
Proposed	403 / 44	26 / 3	104 / 52	5 / 5	37 / 25	1 / 1	1082 / 120

we kept the rest of the hyper-parameters frozen from step one and did an initial search for λ^{cc} with the same optimizer and budget 60 using a broad range of values in 0.5 – 5.0. (We did this step only once for ResNet18 - and not for VGG16BN - to have an indication of λ^{cc} 's magnitude). Finally, we used the One-PlusOne optimizer in order to further fine-tune λ^{cc} in the range 0.05 – 0.5 with a budget of 20 trials, for both ResNet18 and VGG16BN. The exact details and the best hyper-parameter values are given in Table 1. In all experiments we set $s' = 5.0$.

Basis Learning Details We use the Adam optimizer to learn the concept bases, and the learning lasts for a maximum of 300 epochs. We use a learning rate scheduler to reduce the learning rate by a factor of 0.1 when the loss term does not improve for 10 epochs and enable early stopping with patience of 15 epochs and absolute minimum delta of 0.01. For the batch size, we use the maximum number allowed by the available GPU memory. We use the Broden (Bau et al., 2017) and Broden Action (Ramakrishnan et al., 2019) concept datasets to collect CNN feature populations and extract the concept bases. We use the ICL and CCL losses proposed in this work, in addition to the other losses from (Doumanoglou et al., 2023). For basis learning, we linearly combine the various loss terms based on the values suggested by the hyper-parameter search and learn the bases using the whole training split of the respective concept dataset (as opposed to the 70% of the training split that we used for hyper-parameter search). For ResNet50, we used the hyper-parameters suggested by the hyper-parameter

search for ResNet18.

4.2 Evaluation Results

In evaluation, we compare against the natural feature space basis and the basis extracted with UIBE. Comparison with the supervised approach (Zhou et al., 2018) yields similar findings as the ones reported in (Doumanoglou et al., 2023) and is omitted. In these experiments we use the full set of labels available in the concept datasets (i.e. labels from all categories: object / part / scene / material / texture / color and action - when applicable -). For the natural feature space basis we use $\mathbf{w}_i = \mathbf{e}_i$ and compute b_i according to (Bau et al., 2017). Tables 1, 2 depict basis interpretability results for all cases. For **all** the three different networks, the bases suggested by the proposed method are scoring higher than UIBE across **all** the interpretability metrics. The most prominent case regards ResNet50 and \mathcal{S}^2 , in which the proposed method suggests a basis that is more interpretable than UIBE by a relative factor of +45.8%.

Regarding ResNet18 and ResNet50, we find it interesting to share statistics that are reported by Network Dissection (Bau et al., 2017) for all the concept detectors in the bases participating in the benchmark. Those statistics were obtained using a fixed value of $\xi = 0.04$, which is the value that was proposed in the respective paper. Two numbers are provided per concept category in Table 3. The first being the number of concept detectors that can detect a concept from the category, and second, the number of (unique) concepts from the category that can

be detected by the detectors. We find noteworthy to highlight that for ResNet18, the basis suggested by the proposed method, has more than twice as much concept detectors in the category *scene* than the basis suggested by UIBE. Additionally its concept detectors are able to detect 25 *scene* concepts more than the detectors of UIBE. For a network that is trained to do scene classification this suggestion looks highly plausible. An even more prominent result regards ResNet50. Compared to UIBE, the proposed method suggests a basis with **10 times** more concept detectors, for concepts in the *action* category. Additionally, the respective basis' concept detectors can detect 47 more action concepts than the detectors of UIBE. This suggestion aligns better with the goal of a network that is trained to perform action recognition.

5 Conclusion

In this work we proposed to complement previous work (*UIBE*) with a novel loss term, that exploits the knowledge encoded in CNN image classifiers and suggests more interpretable bases. The proposed method demonstrates up to 45.8% interpretability improvements in the extracted bases, when using optimal hyper-parameters that were suggested for learning a basis regarding a different classifier trained on another task. Future work may study applications of the proposed method to debug and improve model performance.

Acknowledgement This work has been supported by the EC funded Horizon Europe Framework Programme: CAVAA Grant Agreement 101071178.

REFERENCES

- Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., and Lapuschkin, S. (2023). From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9).
- Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10(7).
- Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*.
- Chormai, P., Herrmann, J., Müller, K.-R., and Montavon, G. (2022). Disentangled explanations of neural network predictions by finding relevant subspaces.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE.
- Doumanoglou, A., Asteriadis, S., and Zarpalas, D. (2023). Unsupervised interpretable basis extraction for concept-based visual explanations. *Accepted at IEEE TAI, currently available on arXiv preprint arXiv:2303.10523*.
- Fong, R. and Vedaldi, A. (2018). Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *CVPR*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*. PMLR.
- Losch, M., Fritz, M., and Schiele, B. (2021). Semantic bottlenecks: Quantifying and improving inspectability of deep representations. *IJCV*, 129(11).
- Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al. (2019). Moments in time dataset: one million videos for event understanding. *TPAMI*, 42(2).
- Mu, J. and Andreas, J. (2020). Compositional explanations of neurons. In *NIPS*, volume 33. Curran Associates, Inc.
- Ramakrishnan, K., Monfort, M., McNamara, B. A., Lascelles, A., Gutfreund, D., Feris, R. S., and Oliva, A. (2019). Identifying interpretable action concepts in deep networks. In *CVPR Workshops*.
- Rapin, J. and Teytaud, O. (2018). Nevergrad - A gradient-free optimization platform. <https://GitHub.com/FacebookResearch/Nevergrad>.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *ICLR*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *CVPR*.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Vielhaben, J., Blücher, S., and Strodthoff, N. (2022). Sparse subspace clustering for concept discovery (SSCCD).
- Zarlenga, M. E., Barbiero, P., Shams, Z., Kazhdan, D., Bhatt, U., Weller, A., and Jamnik, M. (2023). Towards robust metrics for concept representation evaluation. *arXiv preprint arXiv:2301.10367*.
- Zhang, R., Madumal, P., Miller, T., Ehinger, K. A., and Rubinstein, B. I. P. (2021). Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *AAAI*.
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE TPAMI*.
- Zhou, B., Sun, Y., Bau, D., and Torralba, A. (2018). Interpretable basis decomposition for visual explanation. In *ECCV*.