

# Concept Whitening for Interpretable Image Recognition

Zhi Chen<sup>1</sup> Yijie Bei<sup>2</sup> Cynthia Rudin<sup>1,2</sup>

## Abstract

What does a neural network encode about a concept as we traverse through the layers? Interpretability in machine learning is undoubtedly important, but the calculations of neural networks are very challenging to understand. Attempts to see inside their hidden layers can either be misleading, unusable, or rely on the latent space to possess properties that it may not have. In this work, rather than attempting to analyze a neural network posthoc, we introduce a mechanism, called *concept whitening* (CW), to alter a given layer of the network to allow us to better understand the computation leading up to that layer. When a concept whitening module is added to a CNN, the axes of the latent space are aligned with known concepts of interest. By experiment, we show that CW can provide us a much clearer understanding for how the network gradually learns concepts over layers. CW is an alternative to a batch normalization layer in that it normalizes, and also decorrelates (whitens) the latent space. CW can be used in any layer of the network without hurting predictive performance.

## 1. Introduction

An important practical challenge that arises with neural networks is the fact that the units within their hidden layers are not usually semantically understandable. This is particularly true with computer vision applications, where an expanding body of research has focused centrally on explaining the calculations of neural networks and other black box models. Some of the core questions considered in these posthoc analyses of neural networks include: “What concept does a unit in a hidden layer of a trained neural network represent?” or “Does this unit in the network represent a concept that a human might understand?”

<sup>1</sup>Department of Computer Science, Duke University, USA <sup>2</sup>Department of Electrical and Computer Engineering, Duke University, USA. Correspondence to: Zhi Chen <zhi.chen1@duke.edu>.

The questions listed above are important, but it is not clear that they would naturally have satisfactory answers when performing posthoc analysis on a pretrained neural network. In fact, there are several reasons why various types of posthoc analyses would not answer these questions.

Efforts to interpret individual nodes of pretrained neural networks (e.g. (Zhou et al., 2018a; 2014)) have shown that some fraction of nodes can be identified to be aligned with some high-level semantic meaning, but these special nodes do not provably contain the network’s full information about the concepts. That is, the nodes are not “pure,” and information about the concept could be scattered throughout the network.

Concept-vector methods also (Kim et al., 2018; Zhou et al., 2018b; Ghorbani et al., 2019) have been used to analyze pretrained neural networks. Here, vectors in the latent space are chosen to align with pre-defined or automatically-discovered concepts. While concept-vectors are more promising, they still make the assumption that the latent space of a neural network admits a posthoc analysis of a specific form. In particular, they assume that the latent space places members of each concept in one easy-to-classify portion of latent space. Since the latent space was not explicitly constructed to have this property, there is no reason to believe it holds.

Ideally, we would want a neural network whose latent space *tells us* how it is disentangling concepts, without needing to resort to extra classifiers like concept-vector methods (Kim et al., 2018; Ghorbani et al., 2019), without surveys to humans (Zhou et al., 2014), and without other manipulations that rely on whether the geometry of a latent space serendipitously admits analysis of concepts. Rather than having to rely on assumptions that the latent space admits disentanglement, we would prefer to constrain the latent space directly. We might even wish that the concepts align themselves along the axes of the latent space, so that each point in the latent space has an interpretation in terms of known concepts.

Let us discuss how one would go about imposing such constraints on the latent space. In particular, we introduce the possibility of what we call *concept whitening*. Concept whitening (CW) is a module inserted into a neural network. It constrains the latent space to represent target concepts and also provides a straightforward means to extract them. It

does not force the concepts to be learned as an intermediate step, rather it imposes the latent space to be aligned along the concepts.

For instance, let us say that, using CW on a lower layer of the network, the concept “airplane” is represented along one axis. By examining the images along this axis, we can find the lower-level characteristic that the network is using to best approximate the complex concept “airplane,” which might be white or silver objects with blue backgrounds. In the lower layers of a standard neural network, we cannot necessarily find these characteristics, because the relevant information of “airplane” might be spread throughout latent space rather than along an “airplane” axis.

By looking at images along the airplane axis at each layer, we see how the network gradually represents airplanes with an increasing level of sophistication and complexity.

Concept whitening could be used to replace a plain batch normalization step in a CNN backbone, because it combines batch whitening with an extra step involving a rotation matrix. Batch whitening usually provides helpful properties to latent spaces, but our goal requires the whitening to take place with respect to concepts; the use of the rotation matrix to align the concepts with the axes is the key to interpretability through disentangled concepts. Whitening decorrelates and normalizes each axis (i.e., transforms the post-convolution latent space so that the covariance matrix between channels is the identity).

Exploiting the property that a whitening transformation remains valid after applying arbitrary rotation, the rotation matrix strategically matches the concepts to the axes.

The concepts used in CW do not need to be the labels in the classification problem, they can be learned from an auxiliary dataset in which concepts are labeled. The concepts do not need to be labeled in the dataset involved in the main classification task (though they could be), and the main classification labels do not need to be available in the auxiliary concept dataset.

Through qualitative and quantitative experiments, we illustrate how concept whitening applied to the various layers of the neural network illuminates its internal calculations. We verify the interpretability and pureness of concepts in the disentangled latent space. Importantly for practice, we show that by replacing the batch normalization layer in pretrained state-of-the-art models with a CW module, the resulting neural network can achieve accuracy on par with the corresponding original black box neural network on large datasets, and it can do this within one additional epoch of further training. Thus, with fairly minimal effort, one can make a small modification to a neural network architecture (adding a CW module), and in return be able to easily visualize how the network is learning all of the different concepts

at any chosen layer.

CW can show us how a concept is represented at a given layer of the network. What we find is that at lower layers, since a complex concept cannot be represented by the network, it often creates lower-level *abstract concepts*. For example, an airplane at an early layer is represented by an abstract concept defined by white or gray objects on a blue background. A bed is represented by an abstract concept that seems to be characterized by warm colors (orange, yellow). In that sense, the CW layer can help us to discover new concepts that can be formally defined and built on, if desired.

## 2. Related work

There are several large and rapidly expanding bodies of relevant literature.

### Interpretability and explainability of neural networks:

There have been two schools of thought on improving the interpretability of neural networks: (1) learning an inherently *interpretable* model (Rudin, 2019); (2) providing post-hoc *explanations* for an exist neural network. CW falls within the first type, though it only enlightens what the network is doing, rather than providing a full understanding of the network’s computations. To provide a full explanation of each computation would lead to more constraints and thus a loss in flexibility, whereas CW allows more flexibility in exchange for more general types of explanations. The vast majority of current works on neural networks are of the second type, explainability. A problem with the terminology is that “explanation” methods are often summary statistics of performance (e.g., local approximations, general trends on node activation) rather than actual explanations of the model’s calculations. For instance, if a node is found to activate when a certain concept is present in an image, it does not mean that all information (or even the majority of information) about this concept is involved with that particular node.

Saliency-based methods are the most common form of post-hoc explanations for neural networks (Zeiler & Fergus, 2014; Simonyan et al., 2014; Smilkov et al., 2017; Selvaraju et al., 2017). These methods assign importance weights to each pixel of the input image to show the importance of each pixel to the image’s predicted class. Saliency maps are problematic for well-known reasons: they often provide highlighting of edges in images, regardless of the class. Thus, very similar explanations are given for multiple classes, and often none of them are useful explanations (Rudin, 2019). Saliency methods can be unreliable and fragile (Adebayo et al., 2018).

Other work provides explanations of how the network’s latent features operate. Some measure the alignment of an

individual internal unit, or a filter of a trained neural network, to a predefined concept and find some units have relatively strong alignment to that concept (Zhou et al., 2018a; 2014). While some units (i.e., filters) may align nicely with pre-defined concepts, the concept can be represented diffusely through many units (the concept representation by individual nodes is impure); this is because the network was not trained to have concepts expressed purely through individual nodes. To address this weakness, several concept-based post-hoc explanation approaches have recently been proposed that do not rely on the concept aligning with individual units (Kim et al., 2018; Zhou et al., 2018b; Ghorbani et al., 2019; Yeh et al., 2019). Instead of analyzing individual units, these methods try to learn a linear combination of them to represent a predefined concept (Kim et al., 2018) or to automatically discover concepts by clustering patches and defining the clusters as new concepts (Ghorbani et al., 2019). Although these methods are promising, they are based on assumptions of the latent space that may not hold. For instance, these methods assume that a classifier (usually a linear classifier) exists on the latent space such that the concept is correctly classified. Since the network was not trained so that this assumption holds, it may not hold. More importantly, since the latent space is not shaped explicitly to handle this kind of concept-based explanation, unit vectors (directions) in the latent space may not represent concepts purely. We will give an example in the next section to show why latent spaces built without constraints may not achieve concept separation.

CW avoids these problems because it shapes the latent space through training. In that sense, CW is closer to work on inherently interpretable neural networks, though its use-case is in the spirit of concept vectors, in that it is useful for providing important directions in the latent space.

There are emerging works trying to build inherently interpretable neural networks. Like CW, they alter the network structure to encourage different forms of interpretability. For example, neural networks have been designed to perform case-based reasoning (Chen et al., 2019; Li et al., 2018), to incorporate logical or grammatical structures (Li et al., 2017; Granmo et al., 2019; Wu & Song, 2019), to do classification based on hard attention (Mnih et al., 2014; Ba et al., 2014; Sermanet et al., 2015; Elsayed et al., 2019), or to do image recognition by decomposing the components of images (Saralajew et al., 2019). These models all have different forms of interpretability than we consider (understanding how the latent spaces of each layer can align with a known set of concepts). Other work also develops inherently interpretable deep learning methods that can reason based on concepts, but are different from our work in terms of field of application (Bouchacourt & Denoyer, 2019), types of concepts (Zhang et al., 2018a;b) and ways to obtain concepts (Adel et al., 2018).

In the field of deep generative models, many works have been proposed to make the latent space more interpretable by forcing disentanglement. However, works such as InfoGAN (Chen et al., 2016) and  $\beta$ -VAE (Higgins et al., 2017), all use heuristic interpretability losses like mutual information, while in CW we have actual concepts that we use to align the latent space.

**Whitening and orthogonality:** Whitening is a linear transformation that transforms the covariance matrix of random input vectors to be the identity matrix. It is a classical pre-processing step in data science. In the realm of deep learning, batch normalization (Ioffe & Szegedy, 2015), which is widely used in many state-of-the-art neural network architectures, retains the standardization part of whitening but not the decorrelation. Earlier attempts whiten by periodically estimating the whitening matrix (Desjardins et al., 2015; Luo, 2017), which leads to instability in training. Other methods perform whitening by adding a decorrelation loss (Cogswell et al., 2016). A whitening module for ZCA has been developed that leverages the fact that SVD is differentiable, and supports backpropagation (Huang et al., 2018b; 2019). Similarly, others have developed a differentiable whitening block based on Cholesky whitening (Siarohin et al., 2018). The whitening part of our CW module borrows techniques from IterNorm (Huang et al., 2019) because it is differentiable and accelerated. CW is different from previous methods because its whitening matrix *is multiplied by an orthogonal matrix and maximizes the activation of known concepts along the latent space axes*.

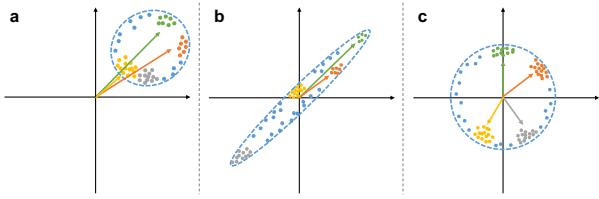
In the field of deep learning, many earlier works that incorporate orthogonality constraints are targeted for RNNs (Vorontsov et al., 2017; Mhammedi et al., 2017; Wisdom et al., 2016), since orthogonality could help avoid vanishing gradients or exploding gradients in RNNs. Other work explores ways to learn orthogonal weights or representations for all types of neural networks (not just RNNs) (Harandi & Fernando, 2016; Huang et al., 2018a; Lezcano-Casado & Martínez-Rubio, 2019; Lezama et al., 2018). For example, some work (Lezama et al., 2018) uses special loss functions to force orthogonality. The optimization algorithms used in the above methods are all different from ours. For CW, we optimize the orthogonal matrix by Cayley-transform-based curvilinear search algorithms (Wen & Yin, 2013). While some deep learning methods also use a Cayley transform (Vorontsov et al., 2017), they do it with a fixed learning rate that does not work effectively in our setting. More importantly, the goal of doing optimization with orthogonality constraints in all these works are completely different from ours. None of them try to align columns of the orthogonal matrix with any type of concept.

### 3. Methodology

Suppose  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathcal{X}$  are samples in our dataset and  $y_1, y_2, \dots, y_n \in \mathcal{Y}$  are their labels. From the latent space  $\mathcal{Z}$  defined by a hidden layer, a DNN classifier  $f : \mathcal{X} \rightarrow \mathcal{Y}$  can be divided into two parts, a feature extractor  $\Phi : \mathcal{X} \rightarrow \mathcal{Z}$ , with parameters  $\theta$ , and a classifier  $g : \mathcal{Z} \rightarrow \mathcal{Y}$ , parameterized by  $\omega$ . Then  $\mathbf{z} = \Phi(\mathbf{x}; \theta)$  is the latent representation of the input  $\mathbf{x}$  and  $f(\mathbf{x}) = g(\Phi(\mathbf{x}; \theta); \omega)$  is the predicted label. Suppose we are interested in  $k$  concepts called  $c_1, c_2, \dots, c_k$ . We can then pre-define  $k$  auxiliary datasets  $\mathbf{X}_{c_1}, \mathbf{X}_{c_2}, \dots, \mathbf{X}_{c_k}$  such that samples in  $\mathbf{X}_{c_j}$  are the most representative samples of concept  $c_j$ . Our goal is to learn  $\Phi$  and  $g$  simultaneously, such that (a) the classifier  $g(\Phi(\cdot; \theta); \omega)$  can predict the label accurately; (b) the  $j^{\text{th}}$  dimension  $z_j$  of the latent representation  $\mathbf{z}$  aligns with concept  $c_j$ . In other words, samples in  $\mathbf{X}_{c_j}$  should have larger values of  $z_j$  than other samples. Conversely, samples not in  $\mathbf{X}_{c_j}$  should have smaller values of  $z_j$ .

#### 3.1. Standard Neural Networks May Not Achieve Concept Separation

Some posthoc explanation methods have looked at unit vectors in the direction of data where a concept is exhibited; this is done to measure how different concepts contribute to a classification task (Zhou et al., 2018b). Other methods consider directional derivatives towards data exhibiting the concept (Kim et al., 2018), for the same reason. There are important reasons why these types of approaches may not work.



**Figure 1.** Possible data distributions in the latent space. **a**, the data are not mean centered; **b** the data are standardized but not decorrelated; **c** the data are whitened. In both **a** and **b**, unit vectors are not valid for representing concepts.

First, suppose the latent space is not mean-centered. This alone could cause problems for posthoc methods that compute directions towards concepts. Consider, for instance, a case where all points in the latent space are far from the origin. In that case, *all* concept directions point towards the same part of the space: the part where the data lies (see Figure 1(a)). This situation might be fairly easy to solve since the users can just analyze the latent space of a batch normalization layer or add a bias term. But then other problems could arise.

Even if the latent space is mean-centered and standardized, the latent space of standard neural networks may not separate concepts. Consider, for instance, an elongated latent space similar to that illustrated in Figure 1(b), by the green and orange clusters. Here, two unit vectors pointing to different groups of data (perhaps exhibiting two separate concepts) may have a large inner product, suggesting that they may be part of the same concept, when in fact, they may be not be similar at all, and may not even lie in the same part of the latent space. Thus, even if the latent space is standardized, multiple unrelated concepts can still appear similar because, from the origin, their centers point towards the same general direction. For the same reason, taking derivatives towards the parts of the space where various concepts tend to appear may yield similar derivatives for very different concepts.

For the above reasons, a latent space in which unit vectors can effectively represent different concepts should have small inter-concept similarity (as illustrated in Figure 1(c)). That is, samples of different concepts should be near orthogonal in the latent space. In addition, for better concept separation, the ratio between inter-concept similarity and intra-concept similarity should be as small as possible.

The CW module we introduce in this work can make the latent space *mean-centered and decorrelated*. This module can align predefined concepts in orthogonal directions. More details of the proposed module will be discussed in Section 3.2 and Section 3.3. The experimental results in Section 4.3 compare the inter-concept and intra-concept similarity of the latent space of standard NNs with and without the proposed CW module. The results validate that the previously mentioned problems of standard neural networks do exist, and that the proposed method successfully avoids these problems.

#### 3.2. Concept Whitening Module

Let  $\mathbf{Z}_{d \times n}$  be the latent representation matrix of  $n$  samples, in which each column  $\mathbf{z}_i \in \mathbb{R}^d$  contains the latent features of the  $i^{\text{th}}$  sample. Our Concept Whitening module (CW) consists of two parts, whitening and orthogonal transformation. The whitening transformation  $\psi$  decorrelates and standardizes the data by

$$\psi(\mathbf{Z}) = \mathbf{W}(\mathbf{Z} - \mu\mathbf{1}_{n \times 1}^T) \quad (1)$$

where  $\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i$  is the sample mean and  $\mathbf{W}_{d \times d}$  is the whitening matrix that obeys  $\mathbf{W}^T \mathbf{W} = \Sigma^{-1}$ . Here,  $\Sigma_{d \times d} = \frac{1}{n} (\mathbf{Z} - \mu\mathbf{1}^T)(\mathbf{Z} - \mu\mathbf{1}^T)^T$  is the covariance matrix. The whitening matrix  $\mathbf{W}$  is not unique and can be calculated in many ways such as ZCA whitening and Cholesky decomposition. Another important property of the whitening matrix is that it is rotation free; suppose  $\mathbf{Q}$  is an orthogonal

matrix, then

$$\mathbf{W}' = \mathbf{Q}^T \mathbf{W} \quad (2)$$

is also a valid whitening matrix. In our module, after whitening the latent space to endow it with the properties discussed above, we still need to rotate the samples in their latent space such that the data from concept  $c_j$ , namely  $\mathbf{X}_{c_j}$ , are highly activated on the  $j^{th}$  axis. Specifically, we need to find an orthogonal matrix  $\mathbf{Q}_{d \times d}$  whose column  $\mathbf{q}_j$  is the  $j^{th}$  axis, by optimizing the following objective:

$$\begin{aligned} & \max_{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_k} \sum_{j=1}^k \frac{1}{n_j} \mathbf{q}_j^T \psi(\mathbf{Z}_{c_j}) \mathbf{1}_{n_j \times 1} \\ & \text{s.t. } \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_d \end{aligned} \quad (3)$$

where  $\mathbf{Z}_{c_j}$  is a  $d \times n_j$  matrix denoting the latent representation of  $\mathbf{X}_{c_j}$  and  $c_1, c_2, \dots, c_k$  are concepts of interest. An optimization problem with an orthogonality constraint like this can be solved by gradient-based approaches on the Stiefel manifold (e.g., the method of (Wen & Yin, 2013)).

This whole procedure constitutes CW, and can be done for any given layer of a neural network as part of the training of the network. The forward pass of the CW module, which makes predictions, is summarized in Algorithm 1.

### 3.3. Optimization and Implementation Detail

Whitening has not (to our knowledge) been previously applied to align the latent space to concepts. In the past, whitening has been used to speed up back-propagation. The specific whitening problem for speeding up back-propagation is different from that for concept alignment—the rotation matrix is not present in other work on whitening, nor is the notion of a concept—however, we can leverage some of the optimization tools used in that work on whitening (Huang et al., 2019; 2018a; Siarohin et al., 2018). Specifically, we adapt ideas underlying the IterNorm algorithm (Huang et al., 2019), which employs Newton’s iterations to approximate ZCA whitening, to the problem studied here. Let us now describe how this is done.

The whitening matrix in ZCA is

$$\mathbf{W} = \mathbf{D} \Lambda^{-\frac{1}{2}} \mathbf{D}^T \quad (4)$$

where  $\Lambda_{d \times d}$  and  $\mathbf{D}_{d \times d}$  are the eigenvalue diagonal matrix and eigenvector matrix given by the eigenvalue decomposition of the covariance matrix,  $\Sigma = \mathbf{D} \Lambda \mathbf{D}^T$ . Like other normalization methods, we calculate a  $\mu$  and  $\mathbf{W}$  for each mini-batch of data, and average them together to form the model used in testing.

As mentioned in Section 3.2, the challenging part for CW is that we also need to learn an orthogonal matrix by solving an optimization problem. To do this, we will optimize the

objective while strictly maintaining the matrix to be orthogonal by performing gradient descent with a curvilinear search on the Stiefel manifold (Wen & Yin, 2013) and adjust it to deal with mini-batch data.

**The two step alternating optimization:** During training, our procedure must handle two types of data: data for calculating the main objective and the data representing the predefined concepts. The model is optimized by alternating optimization: the mini-batches of the main dataset and the following two objectives are optimized in turns. The first objective is the main objective (usually related to classification accuracy):

$$\min_{\theta, \omega, \mathbf{W}, \mu} \frac{1}{n} \sum_{i=1}^n \ell(g(\mathbf{Q}^T \psi(\Phi(\mathbf{x}_i; \theta); \mathbf{W}, \mu); \omega), y_i) \quad (5)$$

where  $\Phi$  and  $g$  are layers before and after the CW module parameterized by  $\theta$  and  $\omega$  respectively.  $\psi$  is a whitening transformation parameterized by sample mean  $\mu$  and whitening matrix  $\mathbf{W}$ .  $\mathbf{Q}$  is the orthogonal matrix. The combination  $\mathbf{Q}^T \psi$  forms the CW module (which is also a valid whitening transformation).  $\ell$  is any differentiable loss. We use cross-entropy loss for  $\ell$  in our implementation to do classification, since it is the most commonly used. The second objective is the concept alignment loss:

$$\begin{aligned} & \max_{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_k} \sum_{j=1}^k \frac{1}{n_j} \sum_{x_i^{(c_j)} \in X_{c_j}} \mathbf{q}_j^T \psi(\Phi(\mathbf{x}_i^{(c_j)}; \theta); \mathbf{W}, \mu) \\ & \text{s.t. } \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_d. \end{aligned} \quad (6)$$

The orthogonal matrix  $\mathbf{Q}$  is fixed when training for the main objective and the other parameters are fixed when training for  $\mathbf{Q}$ . The optimization problem is a linear programming problem with quadratic constraints (LPQC) which is generally NP-hard. Since directly solving for the optimal solution is intractable, we optimize it by gradient methods on the Stiefel manifold. At each step  $t$ , in which the second objective is handled, the orthogonal matrix  $\mathbf{Q}$  is updated by Cayley transform

$$\mathbf{Q}^{(t+1)} = \left( I + \frac{\eta}{2} \mathbf{A} \right)^{-1} \left( I - \frac{\eta}{2} \mathbf{A} \right) \mathbf{Q}^{(t)}$$

where  $\mathbf{A} = \mathbf{G}(\mathbf{Q}^{(t)})^T - \mathbf{Q}^{(t)} \mathbf{G}^T$  is a skew-symmetric matrix,  $\mathbf{G}$  is the gradient of the loss function and  $\eta$  is the learning rate. The optimization procedure is accelerated by curvilinear search on the learning rate at each step (Wen & Yin, 2013). Note that, in the Cayley transform, the stationary points are reached when  $\mathbf{A} = 0$ , which has multiple solutions. Since the solutions are in high-dimensional space, these stationary points are very likely to be saddle points, which can be avoided by SGD. Therefore, we use

---

 Algorithm 1. Forward Pass of CW Module

- 1: **Input:** mini-batch input  $\mathbf{Z} \in \mathbb{R}^{d \times m}$
  - 2: **Optimization Variables:** orthogonal matrix  $\mathbf{Q} \in \mathbb{R}^{d \times d}$  (learned in Algorithm 2)
  - 3: **Output:** whitened representation  $\hat{\mathbf{Z}} \in \mathbb{R}^{d \times m}$
  - 4: calculate batch mean:  $\mu = \frac{1}{m} \mathbf{Z} \cdot \mathbf{1}$ , and center the activation:  $\mathbf{Z}_C = \mathbf{Z} - \mu \cdot \mathbf{1}^T$
  - 5: calculate ZCA-whitening matrix  $\mathbf{W}$ , for details see Algorithm 1 of (Huang et al., 2019)
  - 6: calculate the whitened representation:  $\hat{\mathbf{Z}} = \mathbf{Q}^T \mathbf{W} \mathbf{Z}_C$ .
- 

## Algorithm 2. Alternating Optimization Algorithm for Training

- 1: **Input:** main objective dataset  $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ , concept datasets  $\mathbf{X}_{c_1}, \mathbf{X}_{c_2}, \dots, \mathbf{X}_{c_k}$
  - 2: **Optimization Variables:**  $\theta, \omega, \mathbf{W}, \mu, \mathbf{Q}$ , whose definitions are in Section 3.2
  - 3: **Parameters:**  $\beta, \eta$
  - 4: **for**  $t = 1$  to  $T$  **do**
  - 5:     randomly sample a mini-batch  $\{\mathbf{x}_i, y_i\}_{i=1}^m$  from  $\mathcal{D}$
  - 6:     do one step of SGD w.r.t.  $\theta$  and  $\omega$  on the loss  $\frac{1}{m} \sum_{i=1}^m \ell(g(\mathbf{Q}^T \psi(\Phi(\mathbf{x}_i; \theta); \mathbf{W}, \mu); \omega), y_i)$
  - 7:     update  $\mathbf{W}$  and  $\mu$  by exponential moving average
  - 8:     **if**  $t \bmod 20 = 0$  **then**
  - 9:         sample mini-batches  $\{\mathbf{x}_i^{(c_1)}\}_{i=1}^m, \{\mathbf{x}_i^{(c_2)}\}_{i=1}^m, \dots, \{\mathbf{x}_i^{(c_k)}\}_{i=1}^m$  from  $\mathbf{X}_{c_1}, \mathbf{X}_{c_2}, \dots, \mathbf{X}_{c_k}$
  - 10:         calculate  $\mathbf{G} = \nabla_{\mathbf{Q}}$ , with columns  $\mathbf{g}_j = -\frac{1}{m} \sum_{i=1}^m \psi(\Phi(\mathbf{x}_i^{(c_j)}; \theta); \mathbf{W}, \mu)$  when  $1 \leq j \leq k$ , else  $\mathbf{g}_j = \mathbf{0}$
  - 11:         calculate the exponential moving average of  $\mathbf{G}$ :  $\mathbf{G}' = \beta \mathbf{G}' + (1 - \beta) \mathbf{G}$
  - 12:         obtain learning rate  $\eta$  by curvilinear search, for details see Algorithm 1 of (Wen & Yin, 2013)
  - 13:         update  $\mathbf{Q}$  by Cayley transform:  $\mathbf{Q} \leftarrow (I + \frac{\eta}{2} (\mathbf{G}' \mathbf{Q}^T - \mathbf{Q} \mathbf{G}'^T))^{-1} (I - \frac{\eta}{2} (\mathbf{G}' \mathbf{Q}^T - \mathbf{Q} \mathbf{G}'^T)) \mathbf{Q}$
- 

the stochastic gradient calculated by a mini-batch of samples to replace  $\mathbf{G}$  at each step. To accelerate and stabilize the stochastic gradient, we also apply momentum to it during implementation. Algorithm 2 provides details for the two-step alternating optimization.

**Dealing with the convolution outputs:** In the previous description of our optimization algorithm, we assume that the activations in the latent space form a vector. However, in CNNs, the output of the layer is a tensor instead of a vector. In CNNs, a feature map (a channel within one layer, created by a convolution of one filter) contains the information of how activated a part of the image is by a single filter, which may be a detector for a specific concept. Let us reshape the feature map into a vector, where each element of the vector represents how much one part of the image is activated by the filter. Thus, if the feature map for one filter is  $h \times w$  then a vector of length  $hw$  contains the activation information for that filter around the whole feature map. We do this reshaping procedure for each filter, which reshapes the output of a convolution layer  $Z_{h \times w \times d \times n}$  into a matrix  $Z_{d \times (hwn)}$ , where  $d$  is the number of channels. We then perform CW on the reshaped matrix. After doing this, the resulting matrix is still size  $d \times (hwn)$ . If we reshape this matrix back to its original size as a tensor, one feature map of the tensor now (after training) represents whether a

meaningful concept is detected at each location in the image for that layer. Note that, now the output of a filter is a feature map which is a  $h \times w$  matrix but the concept activation score we used in the optimization problem is a scalar. Therefore, we need to get an activation value from the feature map. There are multiple ways to do this. We try the following calculations to define activation based on the feature map: (a) mean of all feature map values; (b) max of all feature map values; (c) mean of all positive feature map values; (d) mean of down-sampled feature map obtained by max pooling. We use (d) in our experiments since it is good at capturing both high-level and low-level concepts. Detailed analysis and experiments about the choice of different activation calculations are discussed in Supplementary Information A.

**Warm start with pretrained models:** Let us discuss some aspects of practical implementation. The CW module can substitute for other normalization modules such as BatchNorm in an hidden layer of the CNN. Therefore, one can use the weights of a pretrained model as a warm start. To do this, we might leverage a pretrained model (for the same main objective) that does not use CW, and replace a BatchNorm layer in that network with a CW layer. The model usually converges in one epoch (one pass over the data) if a pretrained model is used.

Note that CW does strictly more work than BatchNorm. CW

alone will achieve desirable outcomes of using BatchNorm, therefore, there is no need to use BatchNorm when CW is in place.

**Computational Efficiency:** The CW module involves two iterative optimization steps: one for whitening normalization and one for concept alignment. The efficiency of iterative whitening normalization is justified experimentally in (Huang et al., 2019); the concept alignment optimization is performed only every 20 batches, usually costing less than 20 matrix multiplications and 10 matrix inversions, which do not notably hurt the speed of training. Indeed, our experiments show that there is no significant training speed slowdown using CW compared to using vanilla BN.

## 4. Experiments

In this section, we first show that after replacing one batch norm (BN) layer with our CW module, the accuracy of image recognition is still on par with the original model (4.1). After that, we visualize the concept basis we learn and show that the axes are aligned with the concepts assigned to them. Specifically, we display the images that are most activated along a single axis (4.2.1); we then show how two axes interact with each other (4.2.2); and we further show how the same concept evolves in different layers (4.2.3), where we have replaced one layer at a time. Then we validate the problems standard neural network mentioned in Section 3.1 through experiments and show that CW can solve these problems (4.3). Moreover, we quantitatively measure the interpretability of our concept axes and compare with other concept-based neural network methods (4.4). We also show how we can use the learned representation to measure the contributions of the concepts (4.5). Finally, we show the practicality of the CW module through a case study of skin lesion diagnosis (4.6).

### 4.1. Main Objective Accuracy

We evaluate the image recognition accuracy of the CNNs before and after adding a CW module. We show that simply replacing a BN module with a CW module and training for a single epoch leads to similar main objective performance. Specifically, after replacing the BN module with the CW module, we trained popular CNN architectures including VGG16+BN (Simonyan & Zisserman, 2015), ResNet with 18 layers and 50 layers (He et al., 2016) and DenseNet161 (Huang et al., 2017) on the Places365 (Zhou et al., 2017) dataset. The auxiliary concept dataset we used is MS COCO (Lin et al., 2014). Each annotation, e.g., ‘‘person’’ in MS COCO, was used as one concept, and we selected all the images with this annotation (images having ‘‘person’’ in it), cropped them using bounding boxes and used the cropped images as the data representing the concept. The concept bank has 80 different concepts corresponding to 80 anno-

tations in MS COCO. In order to limit the total time of the training process, we used pretrained models for the popular CNN architectures (discussed above) and fine-tuned these models after BN was replaced with CW.

Table 1 shows the average test accuracy on the validation set of Places365 over 5 runs. We randomly selected 3 concepts from the concept bank to learn using CW for each run, and used the average of them to measure accuracy. We repeated this, applying CW to different layers and reported the average accuracy among the layers. The accuracy does not change much when CW is applied to the different layers and trained on different number of concepts, as shown in Supplementary Information B.

Table 1. Top-1 and top-5 test accuracy on Places365 dataset. Our results show that CW does not hurt performance.

	Top-1 acc.		Top-5 acc.	
	Original	+CW	Original	+CW
VGG16-BN	53.6	53.3	84.2	83.8
ResNet18	54.5	53.9	84.6	84.2
ResNet50	54.7	54.9	85.1	85.2
DenseNet161	55.3	55.5	85.2	85.6

Because we have leveraged a pretrained model, when training with CW, we conduct only one additional epoch of training (one pass over the dataset) for each run. As shown in Table 1, the performance of these models using the CW module is on par with the original model: the difference is within 1% with respect to top-1 and top-5 accuracy. This means in practice, if a pretrained model (using BN) exists, one can simply replace the BN module with a CW module and train it for one epoch, in which case, the pretrained black-box model can be turned into a more interpretable model that is approximately equally accurate.

### 4.2. Visualizing the Concept Basis

In order to demonstrate the interpretability benefits of models equipped with a CW module, we visualize the concept basis in the CW module and validate that the axes are aligned with their assigned concepts. In detail, (a) we check the most activated images on these axes; (b) we look at how images are distributed in a 2D slice of the latent space; (c) we show how realizations of the same concept change if we apply CW on different layers. All experiments in 4.2 were done on ResNet18 equipped with CW trained on Places365 and three simultaneous MS COCO concepts.

#### 4.2.1. TOP-10 ACTIVATED IMAGES

We sort all validation samples by their activation values (discussed in Section 3.3) to show how much they are related to the concept. Figure 2 shows the images that have the top-

10 largest activations along three different concepts' axes. Note that all these concepts are trained together using one CW module.

From Figure 2(b), we can see that all of the top activated images have the same semantic meaning when the CW module is located at a higher layer (i.e., the 16th layer). Figure 2(a) shows that when the CW module is applied to a lower layer (i.e., the 2nd layer), it tends to capture low level information such as color or texture characteristic of these concepts. For instance, the top activated images on the “airplane” axis generally has a blue background with a white or gray object in the middle. It is reasonable that the lower layer CW module cannot extract complete information about high-level concepts such as “airplane” since the model complexity of the first two layers is limited.

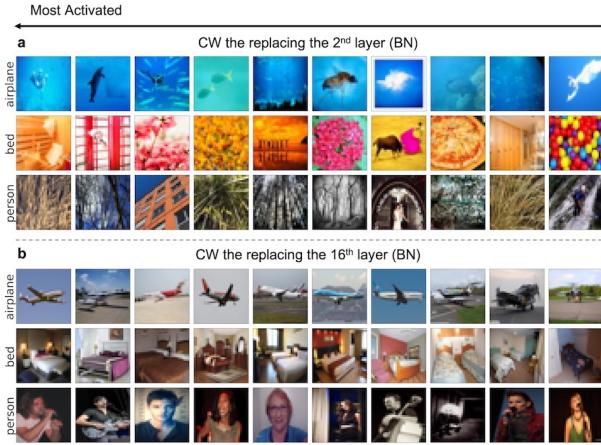


Figure 2. Top-10 Image activated on axes representing different concepts. **a**, results when the 2<sup>nd</sup> layer (BN) is replaced by CW; **b**, results when the 16<sup>th</sup> layer (BN) is replaced by CW.

In that sense, the CW layer has *discovered* lower-level characteristics of a more complex concept; namely it has discovered that the blue images with white objects are primitive characteristics that can approximate the “airplane” concept. Similarly, the network seems to have discovered that the appearance of warm colors is a lower-level characteristic of the “bedroom” concept, and that a dark background with vertical light streaks is a characteristic of the “person” concept.

Interestingly, when different definitions of activation are used (namely the options discussed in Section 3.3), the characteristics discovered by the network often look different. Some of these are shown in Supplementary Information A.2.

Moreover, similar visualizations show that CW can deal with various types of concepts. Top activated images on more concepts, including concepts defined as objects and concepts defined as general characteristics, can be found

in Supplementary Information D. Top activated images visualized with empirical receptive fields can be found in Supplementary Information E.

#### 4.2.2. 2D-REPRESENTATION SPACE VISUALIZATION

Let us consider whether joint information about different concepts is captured by the latent space of CW. To investigate how the data are distributed in the new latent space, we pick a 2D slice of the latent space, which means we select two axes  $q_i$  and  $q_j$  and look at the subspace they form.

The data's joint distribution on the two axes is shown in Figure 3. To visualize the joint distribution, we first compute the activations of all validation data on the two axes, then divide the latent space into a  $50 \times 50$  grid of blocks, where the maximum and minimum activation value are the top and bottom of the grid. For the grid shown in Figure 3(a), we randomly select one image that falls into each block, and display the image in its corresponding block. If there is no image in the block, the block remains black. From Figure 3(a), we observe that the axes are not only aligned with their assigned concepts, they also incorporate joint information. For example, a “person in bed” has high activation on both the “person” axis and “bed” axis.

We also include a 2D histogram of the number of images that fall into each block. As shown in Figure 3(b), most images are distributed near the center (which is the origin) suggesting that the samples' feature vector has high probability to be nearly orthogonal to the concept axes we picked (meaning that they do not exhibit the two concepts), and consequently the latent features have near 0 activation on the concept axes themselves.

#### 4.2.3. TRAJECTORY OF CONCEPTS IN DIFFERENT LAYERS

Although our objective is the same when we apply the CW module to different layers in the same CNN, the latent space we get might be different. This is because different layers might be able to express different levels of semantic meaning. Because of this, it might be interesting to track how the representation of a single image will change as the CW module is applied to different layers of the CNN.

In order to better understand the latent representation, we plot a 2D slice of the latent space. Unlike in the 2d-representation space visualization (Figure 3), here, a point in the plot is not specified by the activation values themselves but by their rankings. For example, the point (0.7, 0.1) means the point is at the 70<sup>th</sup> percentile for the first axis and the 10<sup>th</sup> percentile in the second axis. We use the percentage instead of using the value, because as shown in the 2d-representation space visualization (Figure 3), most points are near the center of the plot, so the rankings spread

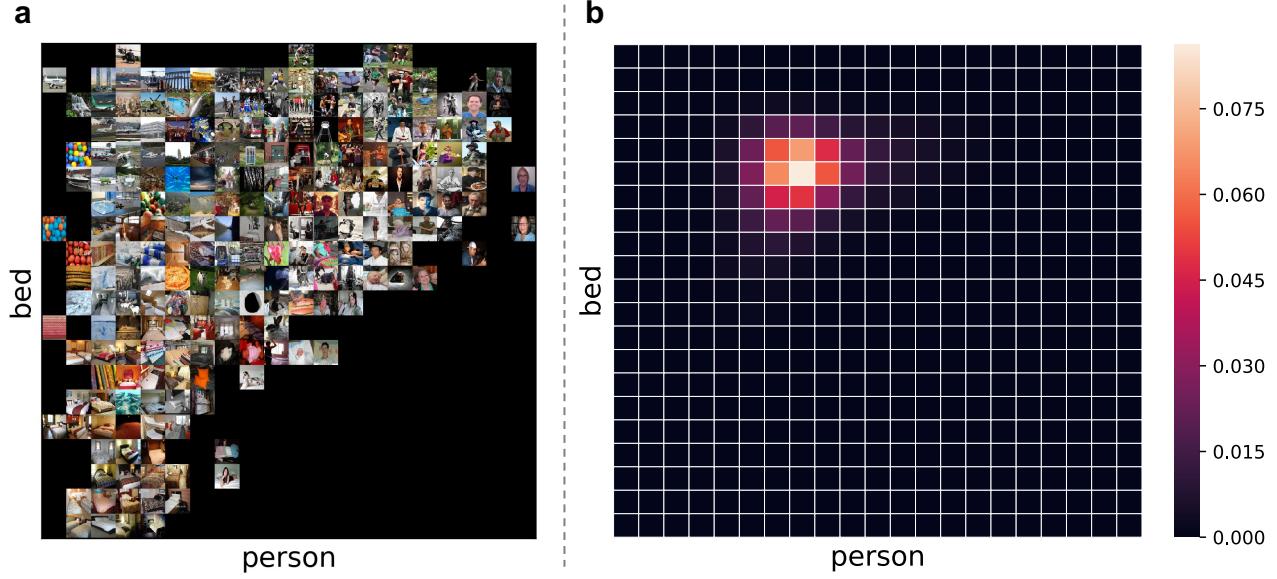


Figure 3. Joint distribution of the bed-person subspace. The bounding box given by projected values in the subspace is evenly divided into  $20 \times 20$  blocks. **a**, Plotting a random test image fall into each block; **b**, Density map of test image representation

the values for plotting purposes.

Figure 4 shows the 2D representation plot of two representative images. Each point in the plot corresponds to the percentile rank representation of the image when the CW module is applied to different layers. The points are connected by arrows according to the depth of the layer. These plots confirm that the abstract concepts learned in the lower layers tend to capture lower-level meaning (such as colors or shapes) while the higher layers capture high-level meaning (such as types of objects). For example, in the left image in Figure 4(a), the bed is blue, where blue is typical low level information about the “airplane” class but not about the “bed” class since bedrooms are usually warm colors. Therefore, in lower layers, the bed image has higher ranking in the “airplane” axis than the “bed” axis. However, when CW is applied to deeper layers, high level information is available, and thus the image becomes highly ranked on the “bed” axis and lower on the “airplane” axis.

In Figure 4(b), traversing through the networks’ layers, the image of a sunset does not have the typical blue coloring of a sky. Its warm colors put it high on the “bedroom” concept for the second layer, and low on the “airplane” concept. However, as we look at higher layers, where the network can represent more sophisticated concepts, we see the image’s rank grow on the “airplane” concept (perhaps the network uses the presence of skies to detect airplanes), and decrease on the “bed” concept.

### 4.3. Separability of Latent Representations

In this subsection, we evaluate properties of the spatial distribution of the concepts in the latent space. By experimentally comparing such properties across latent representations produced by the CW module and other methods, we demonstrate that the issues arising in standard methods, as outlined in Section 3.1, do not occur when using CW. We also investigate such properties on a non-posthoc neural network, trained with an auxiliary loss that aims to classify different concepts in the latent space (that is, in the objective, there are classification losses for each axis, using each axis’ assigned concept as its label). Interestingly, we find that such issues mentioned in Section 3.1 may also exist in that network. The experiments in Section 4.3 were all done on ResNet18. The CW module was trained with seven simultaneous MS COCO concepts.

Specifically, for each concept image, we first extract its latent space representation. The representation for instance  $j$  of concept  $i$  is denoted  $\mathbf{x}_{ij}$ . Then, intra-concept similarity for concept  $i$ , denoted  $d_{ii}$ , is defined to be:

$$d_{ii} = \frac{1}{n^2} \left( \sum_{j=1}^n \sum_{k=1}^n \frac{\mathbf{x}_{ij} \cdot \mathbf{x}_{ik}}{\|\mathbf{x}_{ij}\|_2 \|\mathbf{x}_{ik}\|_2} \right) \quad (7)$$

where  $n$  is the total number of instances of concept  $i$ .

Inter-concept similarity between concept  $p$  and  $q$  is similarly

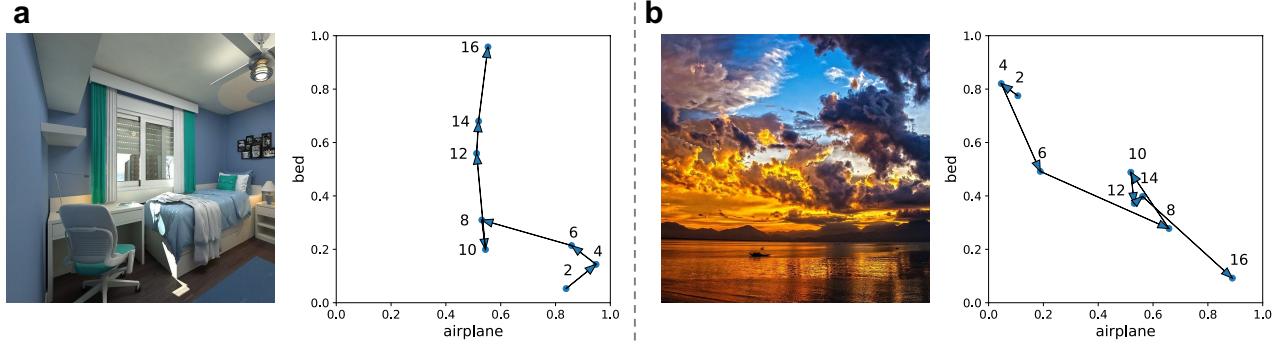


Figure 4. 2D representation plot of two representative images. Each point in the right trajectory plot corresponds to the percentile rank for the activation values on each axis. The number labeling each point on the plot provides the layer depth of the CW module. The trajectory shows how the percentile rank of the left image changes when CW is applied to different layers.

defined as:

$$d_{pq} = \frac{1}{nm} \left( \sum_{j=1}^n \sum_{k=1}^m \frac{\mathbf{x}_{pj} \cdot \mathbf{x}_{qk}}{\|\mathbf{x}_{pj}\|_2 \|\mathbf{x}_{qk}\|_2} \right) \quad (8)$$

where  $n$  and  $m$  are the number of instances of concepts  $p$  and  $q$  respectively. Indeed, intra-concept similarity is the average pairwise cosine similarity between instances of the same concept, and inter-concept similarity is the average pairwise cosine similarity between instances of two different concepts.

With those defined, we plot heat maps in Figure 5 where value in cell at row  $i$  column  $j$  is computed as:

$$Q_{ij} = \frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}}. \quad (9)$$

From Figure 5, we notice that with the CW module, latent representations of concepts achieve greater separability: the ratios between inter-concept and intra-concept similarities (average 0.35) are notably smaller than of standard CNNs (average 0.94). In addition, without normalization, the CW module has very small inter-concept similarities (average 0.05) while analogous values for a standard neural network are around 0.74. This means that in the latent space of CW, two concepts are nearly orthogonal, while in a standard neural network, they are generally not. *This indicates that some of the problems we identified in Section 3.1 occur in standard neural networks, but they do not occur with CW.*

In this experiment, as mentioned earlier, we also trained a standard neural network with a concept-distinction auxiliary loss. The auxiliary loss is the cross entropy of the first several dimensions in the latent space with respect to the concepts we investigated. Shown in Figure 5(b), the latent representations do not naturally help concept separation. The average ratio between inter-concept and intra-concept

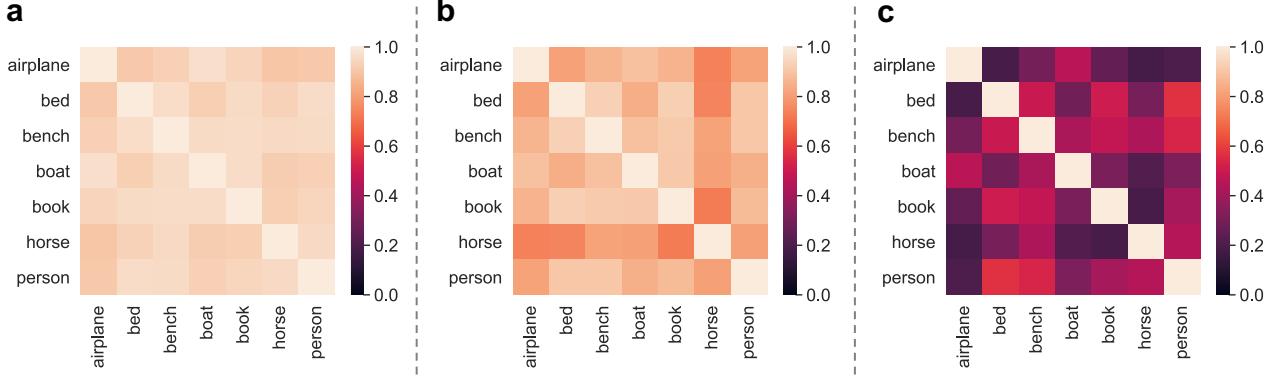
similarities is 0.85. Without normalization, the average inter-concept similarity is also around 0.74, similar to that of the standard neural network without the auxiliary loss. This has important implications: *good discriminative power in the latent space does not guarantee orthogonality of different concepts. Thus, the whitening step is crucial for representing pure concepts.*

#### 4.4. Quantitative Evaluation of Interpretability

In this subsection, we measure the interpretability of the latent space quantitatively and compare it with other concept-based methods.

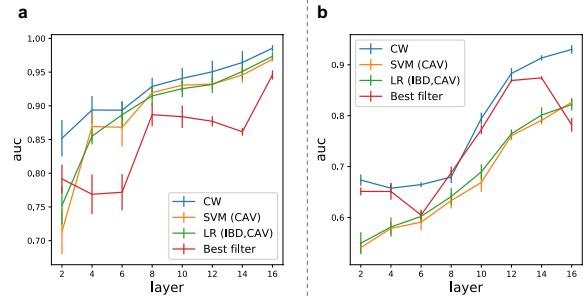
First, we measure the purity of learned concepts by the AUC (of classifying the concept, not classifying with respect to the label for the overall prediction problem) calculated from the activation values. To calculate the test AUC, we divide the concept bank, containing 80 concepts extracted from MS COCO, into training sets and test sets. After training the CW module using the training set, we extract the testing samples' activation values on the axis representing the concept. For the target concept, we assign samples of this concept to the label 1 while giving samples of the other 79 concepts label 0. In this way, we calculate the one-vs-all test AUC score of classifying the target concept in the latent space. The AUC score measures whether the samples belonging to a concept are ranked higher than other samples. That is, the AUC score indicates the purity of the concept axis. Specifically, we randomly choose 14 concepts from the concept bank for the purity comparison. Since our CW module can learn multiple concepts at the same time, we divide the 14 concepts into two groups and train CW with 7 simultaneous concept datasets.

We compared the AUC concept purity of CW with the concept vectors learned by TCAV (Kim et al., 2018) from black box models, IBD (Zhou et al., 2018b) from black box mod-



**Figure 5.** Normalized intra-concept and inter-concept similarities. The diagonal values are normalized average similarities (see definition in Section 4.3) between latent representations of images of the same concept; off-diagonal values are normalized average similarities between latent representations of images of different concepts. **a**, The 16<sup>th</sup> layer is a BN module; **b**, The 16<sup>th</sup> layer is a BN module with auxiliary loss to classify these concepts; **c**, The 16<sup>th</sup> layer is a CW module.

els, and filters in standard CNNs (Zhou et al., 2014). Since TCAV and IBD already find concept vectors, we use the samples’ projections on the vectors to measure the AUC score. Note that in their original papers, the concept vectors are calculated for only one concept each time; therefore, we calculated 14 different concept vectors, each by training a linear classifier in the black box’s latent space, with the training set of the target concept as positive samples and samples randomly drawn from the main dataset as negative samples. For standard CNNs, we measure the AUC score for the output of all filters and choose the best one to compare with our method, separately for each concept (denoted “Best Filter”). Figure 6 shows the AUC concept purity of “airplane” and “person” of these methods across different layers. The error bars on Figure 6 were obtained by splitting the testing set into 5 parts and calculating AUC over each of them. The AUC plots for the other 12 concepts are shown in Supplementary Information D.1.2. From the plots, we observe that *concepts learned in the CW module are generally purer than those of other methods*. This is accredited to the orthogonality of concept representations as illustrated in Section 3.1, as a result of CW’s whitening of the latent space and optimization of the loss function. We perform another quantitative evaluation that aims to measure the correlation of axes in the latent space before and after the CW module is applied. For comparison with posthoc methods like TCAV and IBD, we measure the output of their BN modules in the pretrained model, because the output of these layers are mean centered and normalized, which, as we discussed, are important properties for concept vectors. Shown by the absolute correlation coefficients plotted in Figure 7(a), the axes still have relatively strong correlation after passing through the BN module. If CW were applied instead of BN, they would instead be decorrelated as shown in Figure 7(b). This figure shows the correlation matrices for



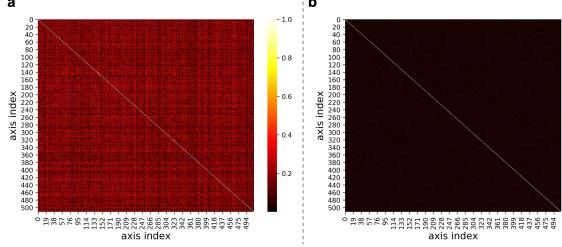
**Figure 6.** Concept purity measured by AUC score. **a**, concept “airplane”; **b**, concept “person.” Concept purity of CW module is compared to several posthoc methods on different layers. The error bar is the standard deviation over 5 different test sets, and each one is 20% of the entire test set.

the 16<sup>th</sup> layer. The same correlation comparison is shown in Supplementary Information C when CW is applied to other layers. These results reflect why purity of concepts is important; *when the axes are pure, the signal of one concept can be concentrated only on its axis, while in standard CNNs, the concept could be strewn throughout the latent space*.

#### 4.5. Concept Importance

In order to obtain practical insights for how the concepts contribute to the classification results, we can measure the concept importance. The concept importance of the  $j^{\text{th}}$  axis is defined as the ratio of a “switched loss” to the original loss:

$$CI_j = \frac{e_{\text{switch}}^{(j)}}{e_{\text{orig}}} \quad (10)$$

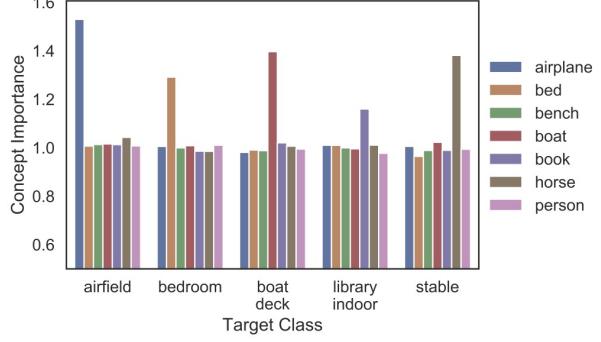


**Figure 7.** Absolute correlation coefficient of every feature pair in the 16<sup>th</sup> layer. **a**, when the 16<sup>th</sup> layer is a BN module; **b**, when 16<sup>th</sup> layer is a CW module.

where the switched loss  $e_{\text{switch}}^{(j)}$  is the loss calculated when the sample values of  $j^{\text{th}}$  axis are randomly permuted, and  $e_{\text{orig}}$  is the original loss without permutation. The expression for  $CI_j$  is similar to classical definitions of variable importance (Breiman, 2001; Fisher et al., 2019). Specifically:

- To measure the contribution of a concept to the entire classifier, the training loss function can be used in the variable importance calculation, which is the multi-class cross entropy in this case.
- To measure the contribution of a concept to a target class, e.g., how much “bed” contributes to “bedroom,” one can use a balanced binary cross entropy loss in the variable importance calculation, calculated on the softmax probability of the target class. The concept importance score is measured on the test set to prevent overfitting.

In our experiments, we measure concept importance scores of the learned concepts to different target classes in the Places365 dataset (corresponding to the second of the bullets above). Figure 8 shows the results in a grouped bar plot. The target classes we choose relate meaningfully to a specific concept learned in CW (e.g., “airplane” and “airfield”). We apply CW on the 16<sup>th</sup> layer since the concepts are generally purer in the layer, as shown in Figure 6. As shown in Figure 8, the irrelevant concepts have concept importance scores near 1.0 (no contribution), e.g., “airplane” is not important to the detection of “bedroom.” For the concepts that relate meaningfully to the target class, e.g., “airplane” to “airfield,” the concept importance scores are much larger than those for other concepts. *Thus, the concept importance score measured on the CW latent space can tell us the contribution of the concept to the classification. For example, it can tell us how much a concept (such as “airplane” contributes to classifying “airfield,” or how much “book” contributes to classifying “library.”*



**Figure 8.** Concept importance to different Places365 classes measured on the concept axes when CW is applied to the 16<sup>th</sup> layer. Each group in the bar plot corresponds to a target class. The bars in the same group show the concept importance scores of the learned concepts to the target class. Concepts that relate meaningfully to the target class (e.g., “airplane” and “airfield”) have larger importance scores than irrelevant concepts.

#### 4.6. Case Study: Skin Lesion Diagnosis

We provide a case study of a medical imaging dataset of skin lesions. The dataset of dermoscopic images is collected from the ISIC archive (ISIC, 2020). Because the dermoscopic images corresponding to different diagnoses vary greatly in appearance, we focus on predicting whether a skin lesion is malignant for each of the histopathology images (9058 histopathology images in total). We choose “age < 20” and “size  $\geq 10 \text{ mm}$ ” as the concepts of interest and select the images with corresponding meta information to form the concept datasets. We chose these concepts due to their availability in the ISIC dataset. The cutoff, for instance, of 10mm is used commonly for evaluation of skin lesions (Rose, 1998). Details about the experimental results including test accuracy, separability of latent representation, AUC concept purity, correlation of axes, and concept importance are shown in Supplementary Information F. The main results of the case study are:

- The conclusions of CW performance analysis on the ISIC dataset are very similar to our earlier conclusions on the Places dataset, in terms of main objective test accuracy, separability of latent representation, AUC concept purity, and correlation of axes.
- Concept importance scores measured on the CW latent space can provide practical insights on which concepts are potentially more important in skin lesion diagnosis.

## 5. Conclusion and Future Work

Concept whitening is a module placed at the bottleneck of a CNN, to force the latent space to be disentangled, and

to align the axes of the latent space with predefined concepts. By building an inherently interpretable CNN with concept whitening, we can gain intuition about how the network gradually learns the target concepts (or whether it needs them at all) over the layers without harming the main objective’s performance.

There are many avenues for possible future work. Since CW modules are useful for helping humans to define primitive abstract concepts, such as those we have seen the network use at early layers, it would be interesting to automatically detect and quantify these new concepts (see ref. (Ghorbani et al., 2019)). Also the requirement of CW to completely decorrelate the outputs of all the filters might be too strong for some tasks. This is because concepts might be highly correlated in practice such as “airplane” and “sky”. In this case, we may want to soften our definition of CW. We could define several general topics that are uncorrelated, and use multiple correlated filters to represent concepts within each general topic. In this scenario, instead of forcing the gram matrix to be the identity matrix, we could make it block diagonal. The orthogonal basis would become a set of orthogonal subspaces.

## Data Availability

All datasets that support the findings are publicly available, including Places365 at <http://places2.csail.mit.edu>, MS COCO at <https://cocodataset.org/> and ISIC at <https://www.isic-archive.com>.

## Code Availability

The code for replicating our experiments is available on <https://github.com/zhiCHEN96/ConceptWhitening> (<https://doi.org/10.5281/zenodo.4052692>).

## References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Proceedings of Conference on Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.
- Adel, T., Ghahramani, Z., and Weller, A. Discovering interpretable representations for both deep generative and discriminative models. In *Proceedings of the International Conference on Machine Learning*, pp. 50–59, 2018.
- Ba, J., Mnih, V., and Kavukcuoglu, K. Multiple object recognition with visual attention. In *Proceedings of the International Conference on Learning Representations*, 2014.
- Bouchacourt, D. and Denoyer, L. Educe: Explaining model decisions through unsupervised concepts extraction. *arXiv preprint arXiv:1905.11852*, 2019.
- Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.
- Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: deep learning for interpretable image recognition. In *Proceedings of Conference on Advances in Neural Information Processing Systems*, pp. 8930–8941, 2019.
- Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 2172–2180, 2016.
- Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. Reducing overfitting in deep networks by decorrelating representations. In *Proceedings of the International Conference on Learning Representations*, 2016.
- Desjardins, G., Simonyan, K., Pascanu, R., et al. Natural neural networks. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 2071–2079, 2015.
- Elsayed, G., Kornblith, S., and Le, Q. V. Saccader: Improving accuracy of hard attention models for vision. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 700–712, 2019.
- Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.
- Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 9273–9282, 2019.
- Granmo, O.-C., Glimsdal, S., Jiao, L., Goodwin, M., Omelin, C. W., and Berge, G. T. The convolutional tsetlin machine. *arXiv preprint arXiv:1905.09688*, 2019.
- Harandi, M. and Fernando, B. Generalized backpropagation, étude de cas: Orthogonality. *arXiv preprint arXiv:1611.05927*, 2016.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations*, 2017.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.
- Huang, L., Liu, X., Lang, B., Yu, A. W., Wang, Y., and Li, B. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018a.
- Huang, L., Yang, D., Lang, B., and Deng, J. Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 791–800, 2018b.
- Huang, L., Zhou, Y., Zhu, F., Liu, L., and Shao, L. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4874–4883, 2019.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR.
- ISIC. digital imaging in skin lesion diagnosis, 2020. data retrieved from ISIC Archive, <https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main>.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the International conference on Machine Learning*, pp. 2668–2677. PMLR, 2018.
- Lezama, J., Qiu, Q., Musé, P., and Sapiro, G. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8109–8118, 2018.
- Lezcano-Casado, M. and Martínez-Rubio, D. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *Proceedings of the International Conference on Machine Learning*, 2019.
- Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- Li, X., Song, X., and Wu, T. Aognets: Compositional grammatical architectures for deep learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755. Springer, 2014.
- Luo, P. Learning deep architectures via generalized whitened neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 2238–2246. JMLR.org, 2017.
- Mhammedi, Z., Hellicar, A., Rahman, A., and Bailey, J. Efficient orthogonal parametrisation of recurrent neural networks using householder reflections. In *Proceedings of the International Conference on Machine Learning*, pp. 2401–2409. JMLR.org, 2017.
- Mnih, V., Heess, N., Graves, A., et al. Recurrent models of visual attention. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 2204–2212, 2014.
- Patterson, G. and Hays, J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758. IEEE, 2012.
- Rose, L. Recognizing neoplastic skin lesions: A photo guide, 1998. <https://www.aafp.org/afp/1998/0915/p873.html>.
- Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, May 2019.
- Saralajew, S., Holdijk, L., Rees, M., Asan, E., and Villmann, T. Classification-by-components: Probabilistic modeling of reasoning over a set of components. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 2788–2799, 2019.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

- Sermanet, P., Frome, A., and Real, E. Attention for fine-grained categorization. In *Proceedings of the International Conference on Learning Representations Workshop*, 2015.
- Siarohin, A., Sangineto, E., and Sebe, N. Whitening and coloring batch transform for gans. In *Proceedings of the International Conference on Learning Representations*, 2018.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.
- Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations Workshop*, 2014.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. In *Proceedings of the International Conference on Machine Learning Workshop*, 2017.
- Vorontsov, E., Trabelsi, C., Kadoury, S., and Pal, C. On orthogonality and learning recurrent networks with long term dependencies. In *Proceedings of the International Conference on Machine Learning*, pp. 3570–3578. JMLR.org, 2017.
- Walter, F. M., Prevost, A. T., Vasconcelos, J., Hall, P. N., Burrows, N. P., Morris, H. C., Kinmonth, A. L., and Emery, J. D. Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study. *Br J Gen Pract*, 63(610):e345–e353, 2013.
- Wen, Z. and Yin, W. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.
- Wisdom, S., Powers, T., Hershey, J., Le Roux, J., and Atlas, L. Full-capacity unitary recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 4880–4888, 2016.
- Wu, T. and Song, X. Towards interpretable object detection by unfolding latent structures. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6033–6043, 2019.
- Yeh, C.-K., Kim, B., Arik, S. O., Li, C.-L., Ravikumar, P., and Pfister, T. On concept-based explanations in deep neural networks. *arXiv preprint arXiv:1910.07969*, 2019.
- Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pp. 818–833. Springer, 2014.
- Zhang, Q., Nian Wu, Y., and Zhu, S.-C. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, 2018a.
- Zhang, Q., Yang, Y., Liu, Y., Wu, Y. N., and Zhu, S.-C. Unsupervised learning of neural networks to explain neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence Workshop*, 2018b.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Object detectors emerge in deep scene cnns. In *Proceedings of the International Conference on Learning Representations*, 2014.
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.
- Zhou, B., Bau, D., Oliva, A., and Torralba, A. Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018a.
- Zhou, B., Sun, Y., Bau, D., and Torralba, A. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018b.

# Concept Whitening for Interpretable Image Recognition

## Supplementary Information

### A. Concept Activation Calculation and Concept Activation Comparison Experiments

#### A.1. Calculations of Concept Activation Based on Feature Maps

The output of a single filter is a  $h \times w$  feature map. However, a scalar is needed to quantify how much a sample is activated on a concept, which is used in both optimization and evaluation. Based on a feature map, multiple reasonable ways exists to calculate the concept activation.

Specifically, we try the following calculations to produce an activation value:

- Mean of all feature map values
- Max of all feature map values
- Mean of all positive feature map values
- Mean of down-sampled feature map obtained by max pooling.

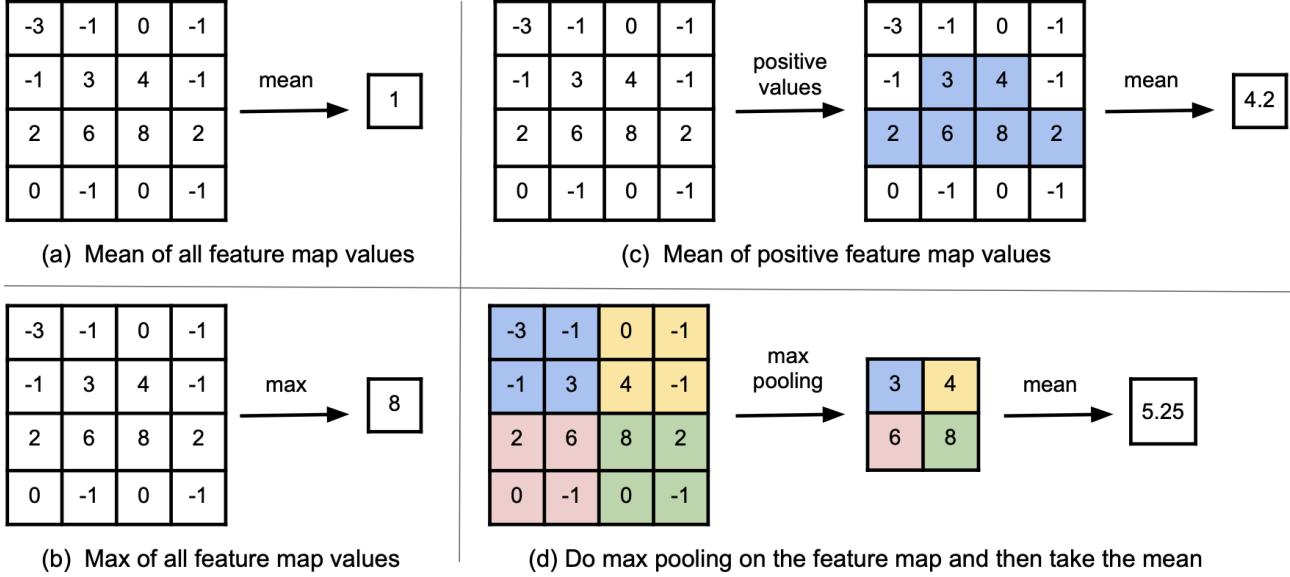
Supplementary Figure 9 shows these four methods of calculating the activation through demonstration. Among them, the mean of values is more suitable for capturing low-level concepts since they are distributed throughout the feature map. For high-level concepts, the max value and mean of positive values are more powerful: they can capture high-level concepts such as objects, since objects usually occur just in one location, not repeatedly throughout an image. The mean of max-pooled values is a combination of the previous types and is capable of representing both high-level and low-level concepts. Intuitively, the mean of max pooled values is more similar to the max function when applied to higher layers and more similar to the mean function when applied to lower layers. This is because, for higher layers, the mean is taken of only a few values, simply because higher layers are smaller in size. Thus, the max is the dominant calculation. In contrast, for lower layers, which are much larger, the max's are taken over a relatively small number of elements (local regions), and then the mean is taken over all of the local regions. Hence the mean is the dominant calculation for lower layers.

#### A.2. Top-10 Activated Images Based on Different Calculations

Supplementary Figure 10 shows the top-10 activated images under the four different calculations for concept activation. The CNN architecture, dataset and the depth of the CW module are the same as before. The figures show that when concept activation is calculated in different ways, the most activated images may look different and the network even may discover completely different lower-level characteristics. For example, when CW is applied to the 2<sup>nd</sup> layer, the network discovered the lower-level characteristics of the concept “bed” to be warm colors when the activation was the mean of feature map values, while the lower-level characteristics seems to involve boundaries of colors if activation is calculated as the max value. Also if the activation is calculated as the mean of all values, the “person” concept gives rise to dense texture, while under the mean of max-pooled values, the “person” concept is characterized as a dark background with vertical lights. This difference in the discovered lower-level characteristics could be explained by the fact that these calculation methods focus on different locations within the image: the mean value focuses on the whole image while the max value only looks at one place within the image.

#### A.3. Concept AUC Based on Different Activation Calculations

Supplementary Table 2 shows concept AUC when different concept activation definitions are used. The definition and calculation of concept AUC is the same as in the main paper. The dataset and CNN architecture are also the same. To compare these concept activations’ capability to capture both high-level concepts and low-level concepts, we apply CW to



Supplementary Figure 9. Four methods of calculating concept activation based on the feature map.

the 2<sup>nd</sup> and 16<sup>th</sup> layers of ResNet18. Supplementary Table 2 indicates that in the 2<sup>nd</sup> layer, the max value of the feature map performs worse on AUC than the other calculation methods for two out of our three concepts. In contrast, in the 16<sup>th</sup> layer, the mean performs poorly compared to the other methods. The max-pool-mean method performs well on both layers, for all concepts. This result matches our intuitive reasoning that the max-pool-mean combines the advantages of mean and max. It is suitable for capturing both low-level concepts and high-level concepts.

Supplementary Table 2. Concept AUC obtained by different calculations of concept activation. Max-pool-mean performs well when CW is applied both to low and high layers.

	AUC-“airplane”		AUC-“bed”		AUC-“person”	
	2 <sup>nd</sup> layer	16 <sup>th</sup> layer	2 <sup>nd</sup> layer	16 <sup>th</sup> layer	2 <sup>nd</sup> layer	16 <sup>th</sup> layer
Mean	0.820	0.981	0.687	0.853	0.714	0.918
Max	0.716	0.992	0.589	0.904	0.759	0.969
Positive-mean	0.798	0.992	0.614	0.924	0.757	0.968
Max-pool-mean	0.818	0.993	0.692	0.906	0.757	0.966

## B. Sensitivity Analysis of Main Objective Accuracy

### B.1. Main Objective Accuracy when CW is Applied to Different Layers

As mentioned in the main paper, we measure the main objective accuracy when CW applied to different layers. Tables 3 through 6 show the layer-wise test accuracy of different CNN architectures. The dataset and CNN architectures are the same as in the main paper. Results in Tables 3 through 6 indicate that no matter which layer we apply CW, accuracy is not substantially impacted.

### B.2. Main Objective Accuracy versus Number of Concepts Trained in CW

We measure the main objective accuracy on Places365 when different numbers of concepts are trained within the CW module. The CNN architecture we evaluate is ResNet18. For each number of concepts, we average the result over three groups of randomly selected concepts. Also, for each group of simultaneous concepts, the result is averaged over different

layers that CW is applied to. As shown in Supplementary Figure 11, both top-1 (Supplementary Figure 11(a)) and top-5 (Supplementary Figure 11(b)) accuracy are not significantly affected by the number of concepts. The drop of accuracy is less than 0.5% when the number of concepts increases from 3 to 9.

## C. Correlation Matrix when CW is Applied to Different Layers

As is shown in the experiments, we calculate the correlations of axes in the latent space to quantitatively compare CW and other concept-based methods. Here, in Supplementary Figure 12, we present the absolute correlation coefficient matrices as heatmaps when these methods are applied to different layers ( $2^{nd}$ ,  $4^{th}$ ,  $6^{th}$ ,  $8^{th}$ ,  $12^{th}$ ,  $14^{th}$  and  $16^{th}$  layer) in ResNet-18. The correlation coefficient is calculated on the test set. The darker the off-diagonal elements are, the more decorrelated the latent space is. Heatmaps of CW in Supplementary Figure 12 are all near pure black. *This demonstrates that CW can consistently decorrelate the latent space – whichever layer it is applied on – while the neural networks trained without any constraints can have strong correlations between different axes. Such a strongly decorrelated latent space enables the signal of one concept to be concentrated on one axis rather than throughout the latent space.*

## D. Results on More Concepts

To show the capability of dealing with many concepts as well as the usefulness of the proposed method, we conduct experiments on more concepts, including concepts defined as objects (D.1) and concepts defined as general characteristics of objects and scenes (D.2).

### D.1. Object Concepts

The object concept bank contains 80 concepts obtained from MS COCO (Lin et al., 2014) by cropping out objects in bounding boxes. We chose 7 concepts randomly selected from the concept bank each time, where the CW module was trained on all of these concepts at the same time.

#### D.1.1. TOP-10 ACTIVATED IMAGES

Supplementary Figure 13 shows the two groups of top-10 activated images along the seven different concepts' axes. On the left of Supplementary Figure 13, we can see that when the CW module is applied to a lower layer (the  $2^{nd}$  layer), it captures some low-level information such as color and texture about the concept. Top activated images on the right of Supplementary Figure 13 demonstrate the concepts' high-level meaning when CW is located at a higher layer (the  $16^{th}$  layer). An interesting finding is that when two similar concepts are given, for example “bus” and “car,” the network can learn their difference and distinguish them successfully in both low and high layers. Moreover, if images with the concept do not exist in the main dataset, the concept axes can be activated by images that are very similar to the concept, like slats and wood textures for the “bench” concept and tents for the “umbrella” concept.

#### D.1.2. AUC CONCEPT PURITY

Supplementary Figure 14 compares the AUC concept purity of 14 concepts learned by TCAV (Kim et al., 2018), IBD (Zhou et al., 2018a), filters in standard CNNs (Zhou et al., 2014), and the CW module in eight different layers ( $2^{nd}$ ,  $4^{th}$ ,  $6^{th}$ ,  $8^{th}$ ,  $10^{th}$ ,  $12^{th}$ ,  $14^{th}$ , and  $16^{th}$  layers). In the figure, the blue line with error bars frequently dominates the AUC across the layers. Therefore, the concepts learned by CW module are generally purer than those learned by other methods.

### D.2. General Characteristics Concepts

The concept bank describing general characteristics of objects and scenes is obtained from the SUN Attribute Database (Patterson & Hays, 2012). The attributes in the dataset are used as concepts, and images given three (out of three) MTurk votes on having such attributes are selected to form the concept datasets. Here we train the CW module on two groups of concepts describing weather of the scene (“cold,” “moist/damp,” “warm”), and materials of objects in the scene (“metal,” “rubber/plastic,” “wood”). Supplementary Figure 15 shows the top-10 activated images along these concepts' axes. This figure demonstrates that the CW module can also decently capture high-level meaning (right column) and low-level aspects (left column) of both types of general concepts.

## E. Top Activated Images Visualized with Empirical Receptive Fields

To show what local feature could be detected along each concept axis, we visualize the top activated images with the empirical receptive field (Zhou et al., 2014). Empirical receptive fields, in our case, are locations in the image, such that when we black them out, they lead to the greatest reduction in activation values on the different axes of the CW output. We have used  $32 \times 32$  random covering patches and a stride of 5 for the sliding window. Supplementary Figure 16 shows the visualization results when CW is applied to the 2<sup>nd</sup>, 12<sup>th</sup> and 16<sup>th</sup> layer. Generally, the top activated images for a concept tend to have a larger receptive field on that concept’s axis. For an early layer (the 2<sup>nd</sup> layer), the features captured by the concept axes appear to be color and textures. As we proceed to deeper layers, concepts learned by CW become closer to the concepts they aim to represent. For example, in the 12<sup>th</sup> layer, the “horse” axis looks at image segments similar to horse legs and the “person” axis looks mainly at the hands and faces of people. In the 16<sup>th</sup> layer, the “horse” axis is looking at the body of the horse and “person” axis is looking at the person’s face; both are more representative features. Interestingly, when two concepts occur in the same image, the two concept axes can detect the correct local regions corresponding to these concepts (e.g., the image containing both “book” and “person” on the 5<sup>th</sup> row of the bottom right subfigure).

## F. Case Study: Skin Lesion Diagnosis

In this section, we provide a case study of a medical imaging dataset of skin lesions. The dataset of dermoscopic images is collected from the ISIC archive (ISIC, 2020). Because the dermoscopic images corresponding to different diagnoses vary greatly in appearance, we focus on predicting whether a skin lesion is malignant for each of the histopathology images (9058 histopathology images in total). We choose “age < 20” and “size  $\geq 10$  mm” as the concepts of interest and select the images with corresponding meta information to form the concept datasets. We chose these concepts due to their availability in the ISIC dataset. The cutoff, for instance, of 10mm is used commonly for evaluation of skin lesions (Rose, 1998). Details about the experimental results are shown in the following order: test accuracy, separability of latent representation, AUC concept purity, correlation of axes, and concept importance.

### F.1. Test Accuracy

We trained both a standard ResNet18 and a ResNet18 with CW on 80% of the dataset and tested it on the other 20%. Since the two classes are imbalanced, we measured the balanced accuracy to compare their performances. The test balanced accuracy of standard ResNet18 is 71.65% while ResNet18 with CW achieves 72.26% test balanced accuracy (this is the average over different layers CW was applied to). Thus, adding CW improved performance over the black box; this may have resulted from whitening, which acts as a regularizer. The latent representation in the standard neural network may be elongated due to the inter similarity of the dermoscopic images (empirically shown in Section F.4), potentially leading to worse performance. This is why whitening could have provided better numerical conditioning for the gradient, as discussed also by (Huang et al., 2018b).

### F.2. Separability of Latent Representation

Similar to experiments on the Places dataset, we measured the separability of concepts in the latent space of CW and a standard ResNet (see Figure 17). When including the CW module, the separability of concepts is also significantly improved.

### F.3. AUC Concept Purity

Similar to experiments on the Places dataset, we quantitatively compare the purity of learned concepts with concept-based posthoc methods. As is shown in Figure 18, the concept “age < 20” is purer using the CW module. All methods were approximately tied in the purity of the concept “size  $\geq 10$  mm.”

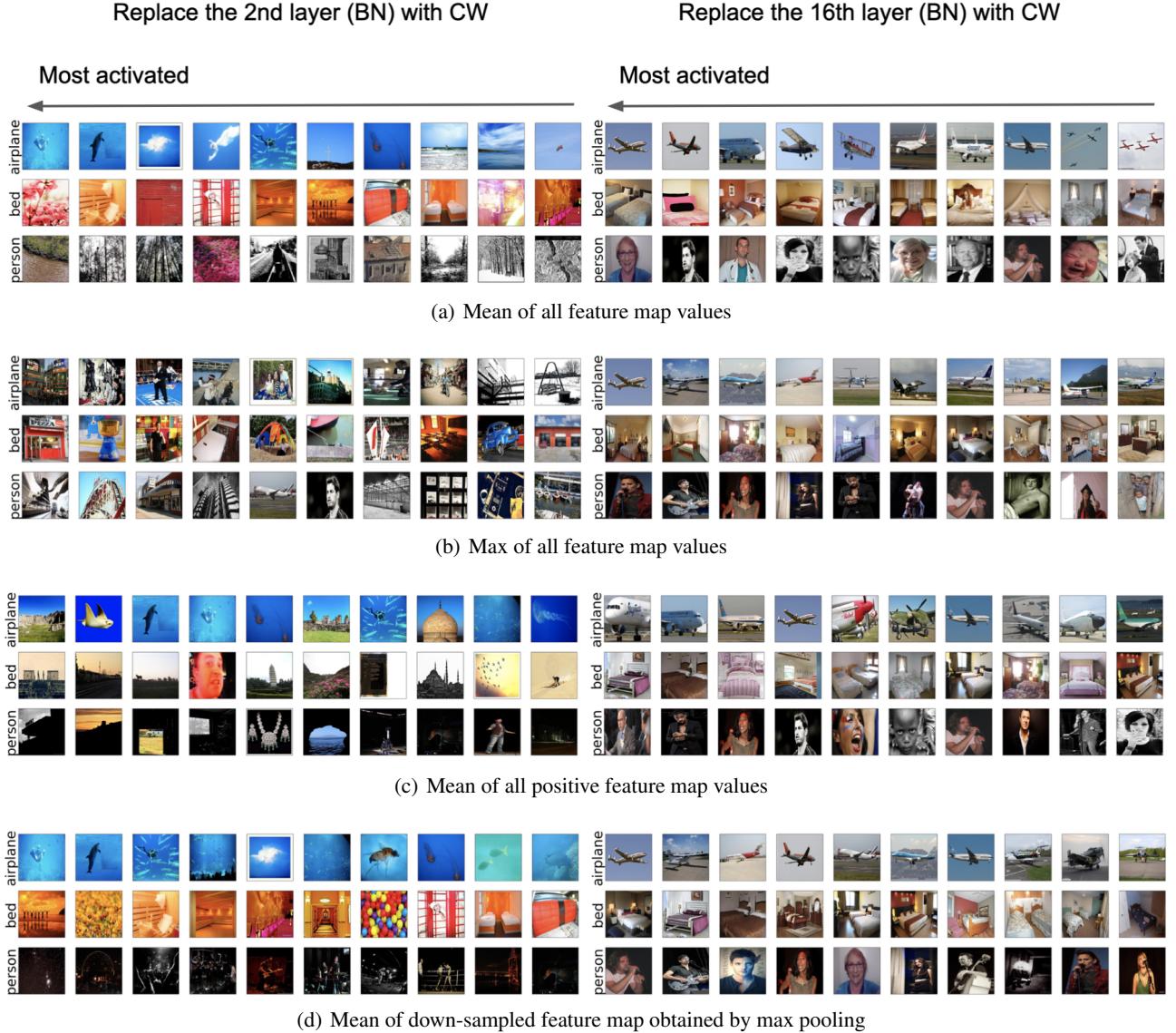
### F.4. Correlation of Axes

Figure 19 shows the correlation of axes in the 16<sup>th</sup> layer of ResNet18 with and without the CW module. Shown in Figure 19(a), the correlations of different axes in standard neural networks are very strong (near 1 in many cases). Such highly correlated data distributions in the latent space may negatively influence both the concept separation and stochastic gradient descent, consistent with results in Section F.1 and F.2. On the contrary, CW can decorrelate the latent space successfully (shown in Figure 19(b)).

## F.5. Concept Importance

Similar to experiments on the Places dataset, we measure the concept importance scores of concepts in the ISIC dataset. Since the dataset only has two classes, we can measure the contribution to the entire classification problem, using balanced binary cross entropy loss for  $e_{\text{switch}}^{(j)}$  and  $e_{\text{orig}}$ . Figure 20 shows the concept importance of different axes of the latent space when CW is applied to the 16<sup>th</sup> layer. We choose the 16<sup>th</sup> layer to investigate because the concepts are purer in the layer as shown in Figure 18. We measure the concept importance of the two concepts we selected and the max and mean concept importance of the 512 axes in the latent space (left subplot of Figure 20). To compare them with the concept importance of other axes, we also visualize the rough distribution of the concept importance with a box plot (right subplot of Figure 20). We observe that the concept “age > 20” is not important at all ( $\approx 1.0$ ). The concept “size  $\geq 10\text{mm}$ ” is more important than most axes (approximately the third quartile among the 512 axes). This concept is known to be important for the way physicians interpret skin lesions (Walter et al., 2013).

It is interesting to contemplate what concept the most important axis (the 76<sup>th</sup> axis) might represent. This axis was not trained to represent a concept, but insight from examining it might lead to possible ideas for concepts we would consider in the future. In Figure 21, we visualize the top-10 activated images along this interesting axis, as well as other axes (axes 0, 1, 100, 150, 200, 250) for comparison. We highlight the empirical receptive fields (Zhou et al., 2014) on the images. Compared to other axes, the empirical receptive fields of 76<sup>th</sup> axis seems to more consistently focus on the borders of the lesions. The lesion border is well known to be important for early detection of melanoma; an irregular border is a major factor, and is even more important than the overall size of the lesion (Walter et al., 2013). This observation naturally leads to a direction for future research: create a concept axis for irregular lesion borders. Since the ISIC dataset does not have each image labeled as to whether the lesion’s borders are irregular, this would need to be labeled by a physician in future work. Doing this would allow us to measure the importance of irregular borders for predicting malignancy of skin lesions by a neural network model.



*Supplementary Figure 10.* Top-10 activated images obtained by different calculations of concept activation. Depending on which choice (max, mean, mean of positives, mean of max pool values), different abstract concepts are generated in the second layer. For instance, for the person class, using the mean calculation on the second layer (top left), the abstract concept is a dense texture. For the person class using the mean of max-pooled values (bottom left), the abstract concept is a dark background with vertical lights. For the bed concept with the mean of max-pooled values (bottom left), the abstract concept is warm colors, whereas for the max calculation (second row left) the abstract concept seems to be related to boundaries of different colors. These concepts could later be formalized, if desired, to create better or more interpretable classifiers in the future.

*Supplementary Table 3.* Top-1 and top-5 accuracy of VGG16-CW on Places365 dataset. Our results indicate that the choice of layer to apply CW does not have a practical impact on accuracy.

CW layer	Top-1 acc.	Top-5 acc.
1 <sup>nd</sup>	53.2	83.8
2 <sup>th</sup>	53.3	83.8
3 <sup>th</sup>	53.4	83.8
4 <sup>th</sup>	53.4	83.9
5 <sup>th</sup>	53.2	83.9
6 <sup>th</sup>	53.3	83.8
7 <sup>th</sup>	53.5	83.8
8 <sup>rd</sup>	53.3	83.9
9 <sup>nd</sup>	53.4	83.8
10 <sup>th</sup>	53.2	83.8
11 <sup>nd</sup>	53.2	83.9
12 <sup>th</sup>	53.3	83.7

*Supplementary Table 5.* Top-1 and top-5 accuracy of DenseNet161-CW on Places365 dataset. Our results indicate that the choice of layer to apply CW does not have a practical impact on accuracy.

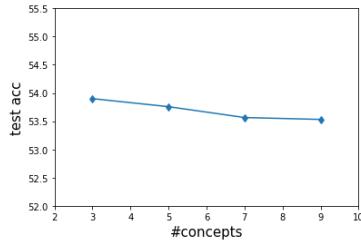
CW layer	Top-1 acc.	Top-5 acc.
14 <sup>th</sup>	55.6	85.7
39 <sup>th</sup>	55.5	85.5
88 <sup>nd</sup>	55.5	85.6
161 <sup>th</sup>	55.5	85.6

*Supplementary Table 4.* Top-1 and top-5 accuracy of ResNet50-CW on Places365 dataset. Our results indicate that the choice of layer to apply CW does not have a practical impact on accuracy.

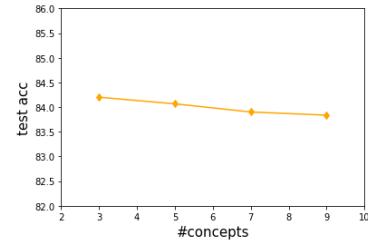
CW layer	Top-1 acc.	Top-5 acc.
2 <sup>nd</sup>	55.2	85.4
5 <sup>th</sup>	55.3	85.5
8 <sup>th</sup>	55.3	85.5
11 <sup>th</sup>	55.2	85.5
14 <sup>th</sup>	55.3	85.5
17 <sup>th</sup>	54.8	85.2
20 <sup>th</sup>	54.7	85.0
23 <sup>rd</sup>	54.8	85.0
26 <sup>nd</sup>	54.7	85.0
29 <sup>th</sup>	54.8	85.0
32 <sup>nd</sup>	54.8	85.1
35 <sup>th</sup>	54.7	85.0
38 <sup>th</sup>	54.8	85.1
41 <sup>st</sup>	54.6	85.0
44 <sup>th</sup>	54.7	84.9
47 <sup>th</sup>	54.6	85.0

*Supplementary Table 6.* Top-1 and top-5 accuracy of ResNet18-CW on Places365 dataset. Our results indicate that the choice of layer to apply CW does not have a practical impact on accuracy.

CW layer	Top-1 acc.	Top-5 acc.
2 <sup>nd</sup>	53.9	84.2
4 <sup>th</sup>	54.0	84.5
6 <sup>th</sup>	54.0	84.3
8 <sup>th</sup>	54.0	84.2
10 <sup>th</sup>	54.0	84.3
12 <sup>th</sup>	53.9	84.1
14 <sup>th</sup>	53.7	83.9
16 <sup>th</sup>	53.5	83.8

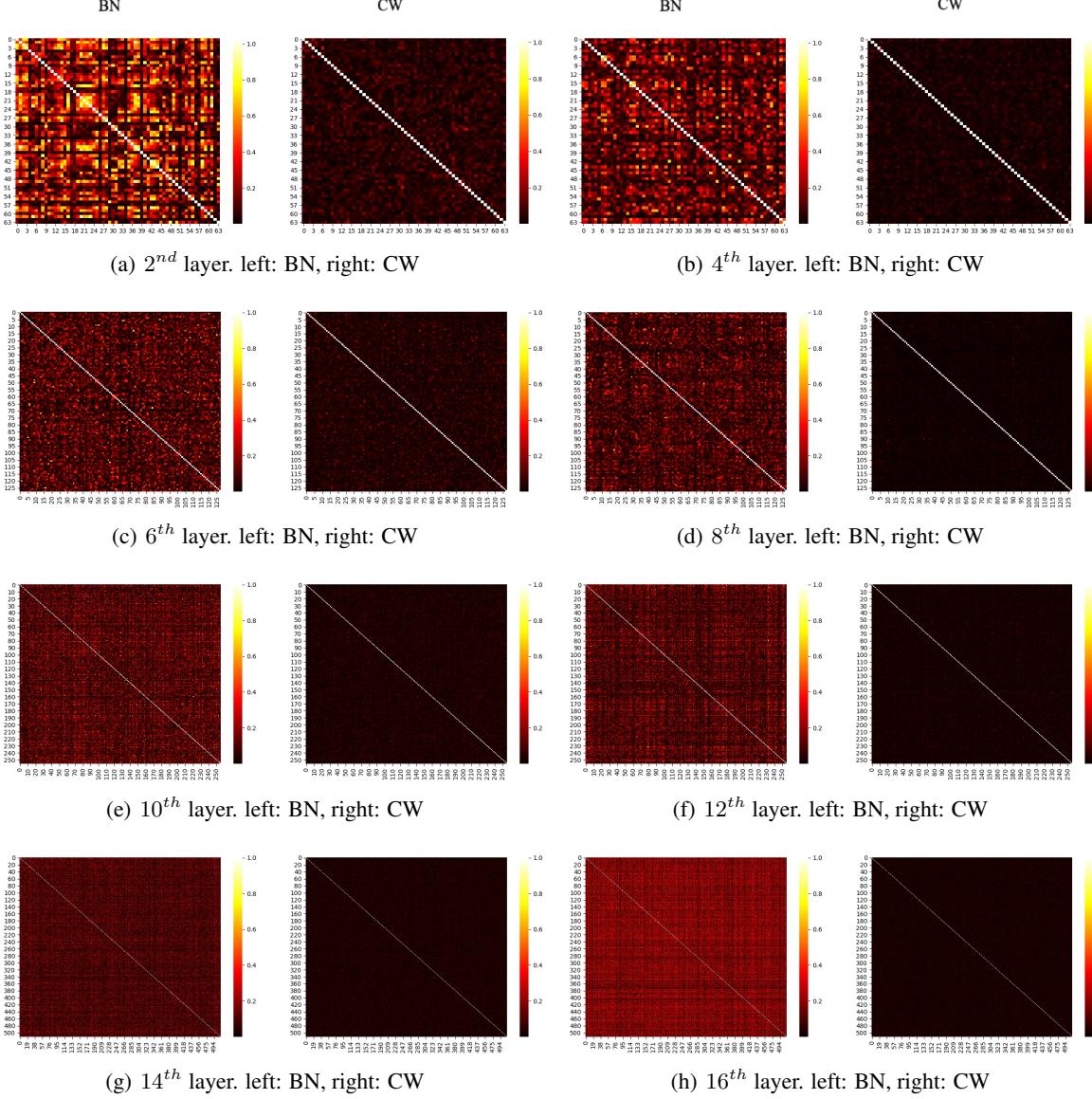


(a) Top-1 test accuracy



(b) Top-5 test accuracy

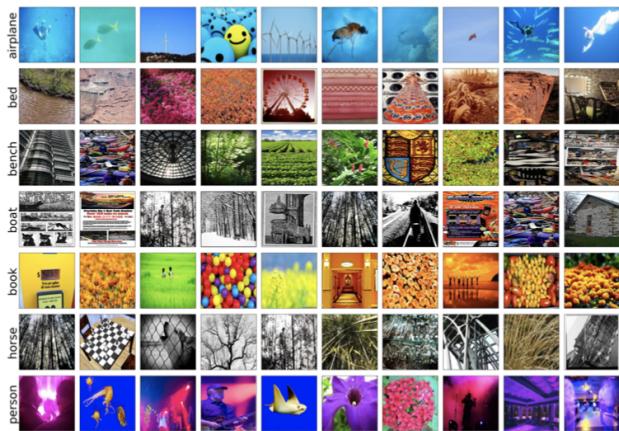
*Supplementary Figure 11.* Test accuracy on Places365 when different number of concepts are learned in CW. (a) Top-1 accuracy; (b) Top-5 accuracy.



*Supplementary Figure 12.* Absolute correlation coefficient of every latent feature pair in the 2<sup>nd</sup>, 4<sup>th</sup>, 6<sup>th</sup>, 8<sup>th</sup>, 12<sup>th</sup>, 14<sup>th</sup> and 16<sup>th</sup> layer, calculated on the test set. For each pair of figures, the left figure is when the layer is a BN module; the right figure is when the layer is a CW module. For the CW module, the first several features represent the concepts. The correlations of CW are much lower off the diagonal, as desired.

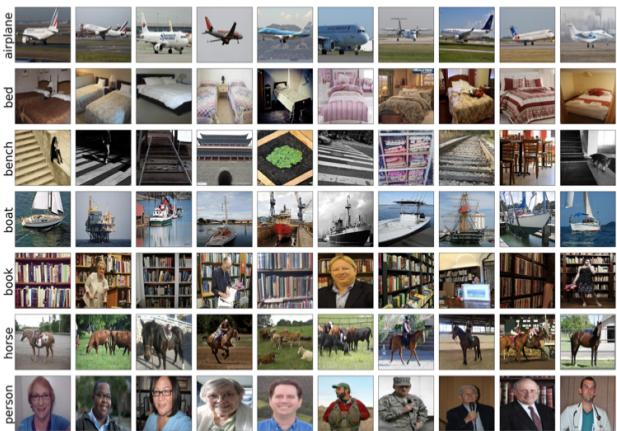
Replace the 2nd layer (BN) with CW

Most activated

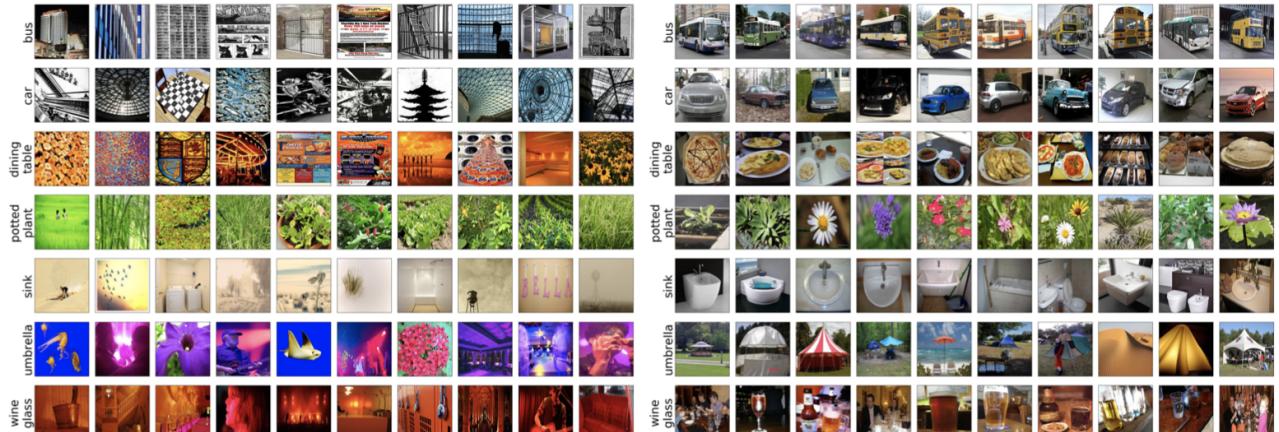


Replace the 16th layer (BN) with CW

Most activated

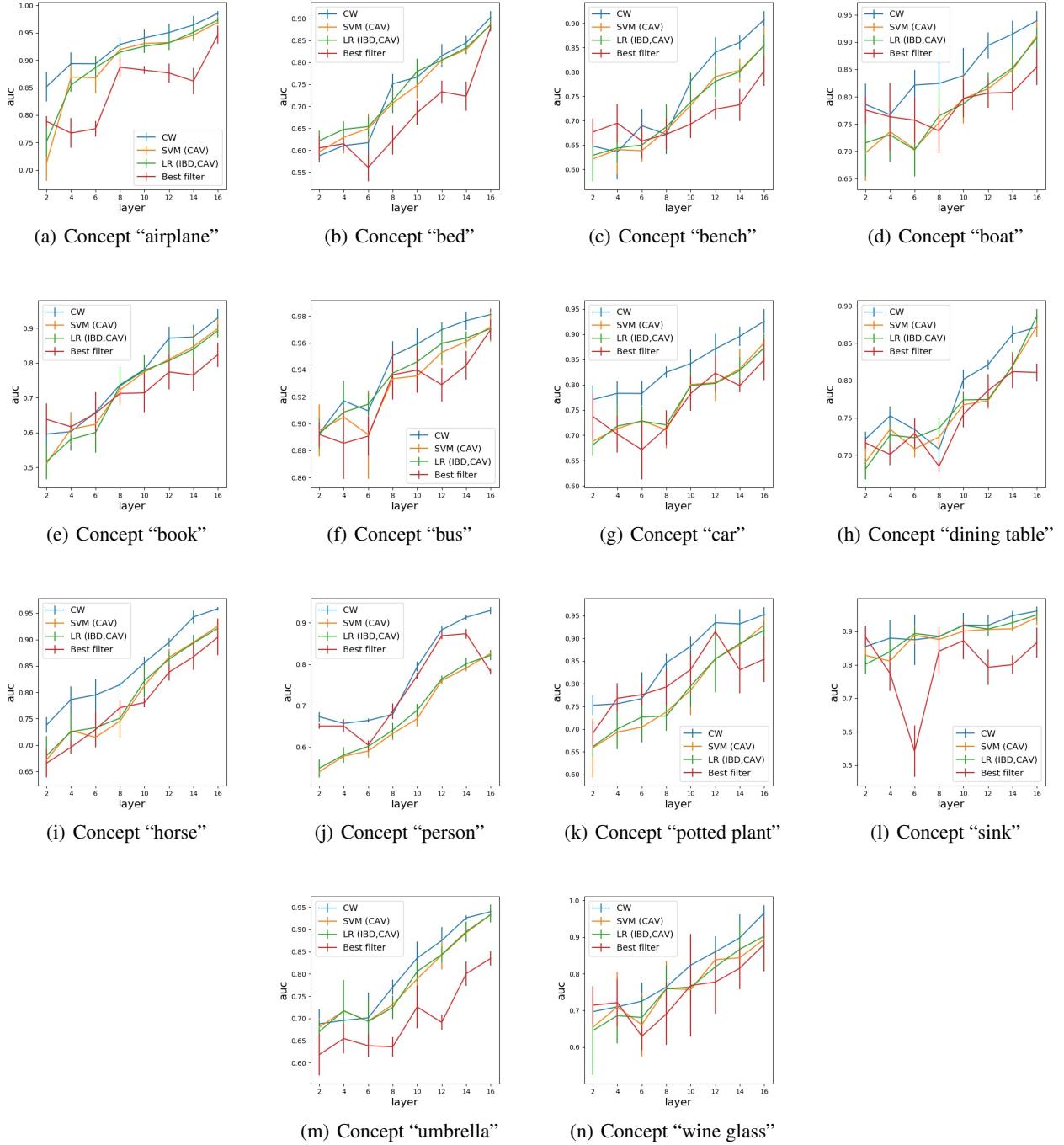


(a) Concepts trained together: airplane, bed, bench, boat, book, person

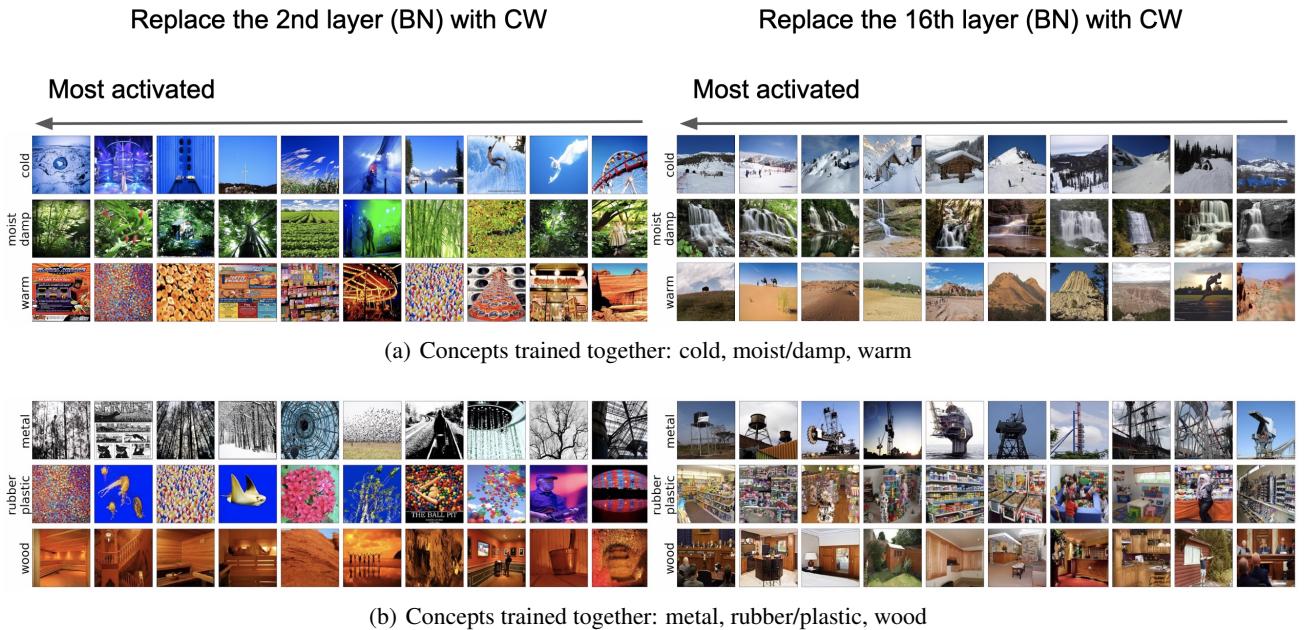


(b) Concepts trained together: bus, car, dining table, potted plant, sink, umbrella, wine glass

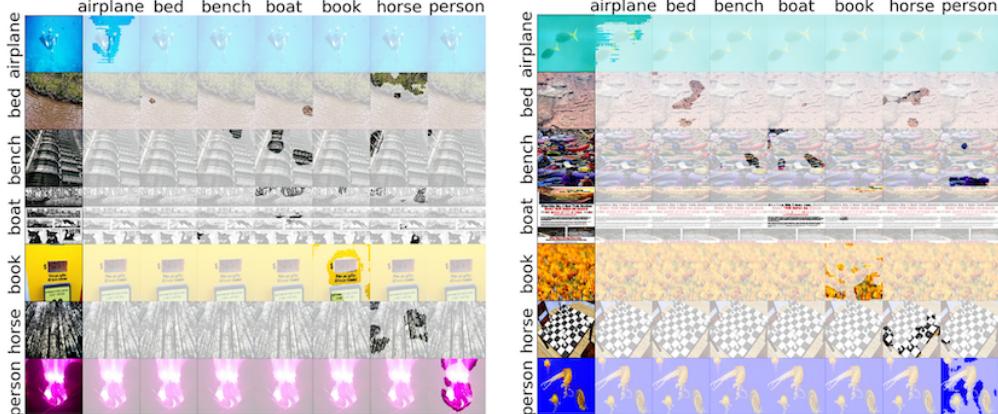
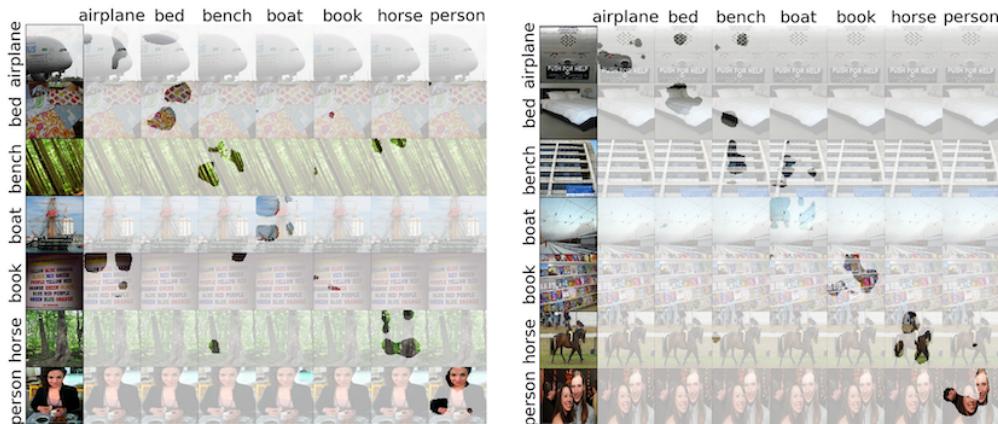
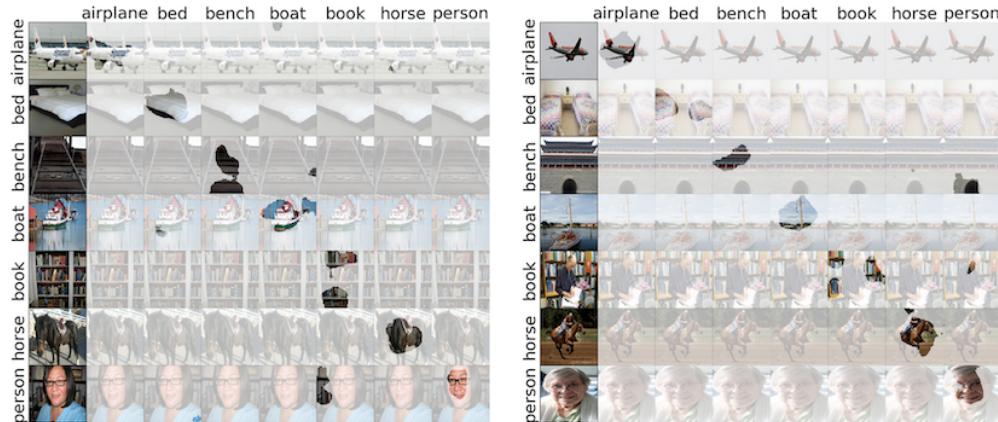
*Supplementary Figure 13.* Top-10 activated images when CW is trained on more object concepts.



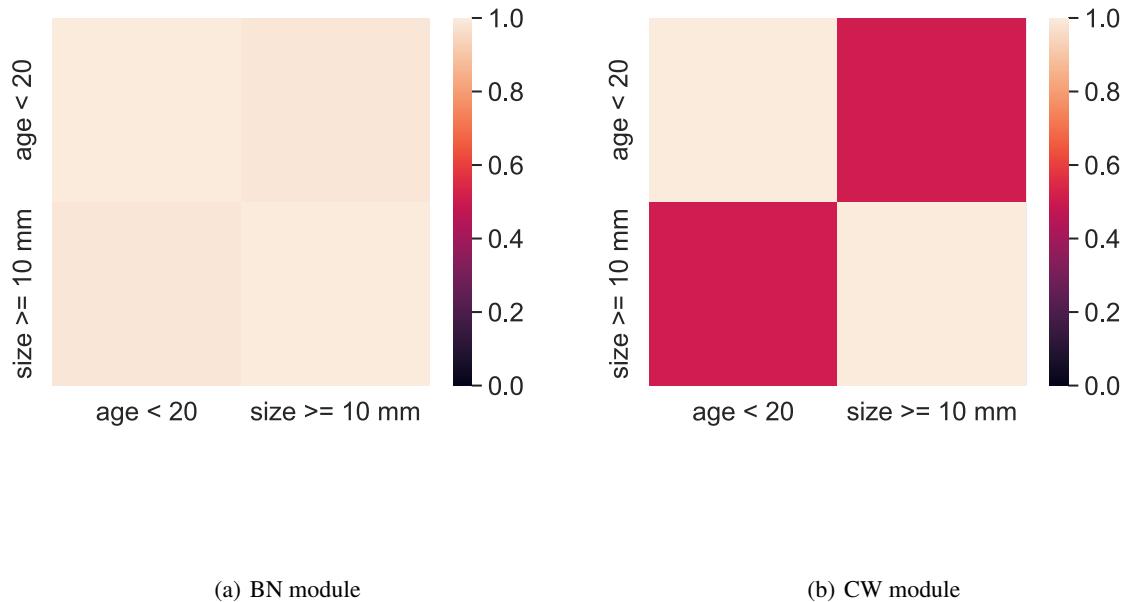
*Supplementary Figure 14.* Concept purity measured by AUC score on 14 different concepts. Concept purity of CW module is compared to several posthoc methods on different layers. For CW, these figures summarize the results of 16 trained neural networks, each of which had a CW layer containing 7 simultaneous concepts at a different location within the network. For the black box baseline, we use the PlacesCNN neural network (Zhou et al., 2017). The error bar is calculated by the standard deviation over 5 different test sets, and each one is 20% of the entire test set.



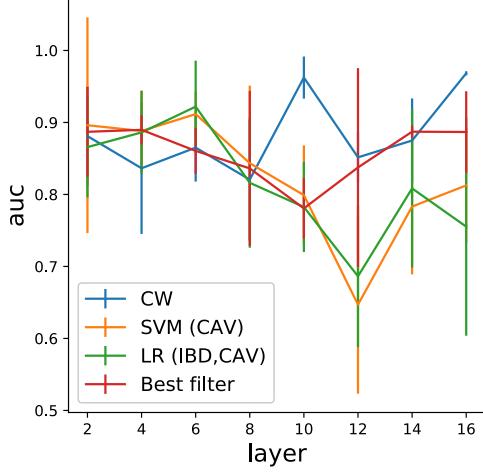
*Supplementary Figure 15.* Top-10 activated images when CW is trained on concepts that describe general properties of objects and scenes. In the second layer (left subfigures), the concepts seem to be represented by simpler primitive concepts, such as color and texture, whereas in the 16<sup>th</sup> layer (right subfigures) the most activated images seem to correctly capture the high-level meaning of the concept.


 (a)  $2^{nd}$  layer

 (b)  $12^{th}$  layer

 (c)  $16^{th}$  layer

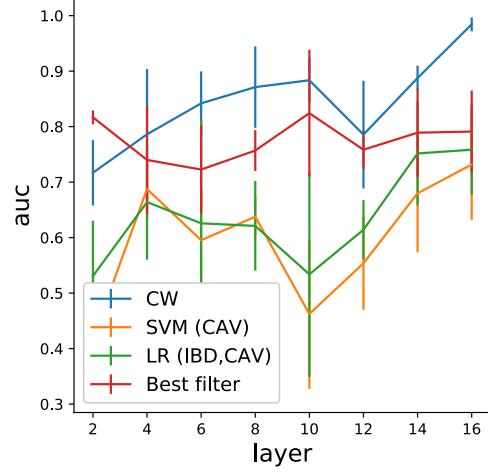
*Supplementary Figure 16.* Some top activated images visualized with empirical receptive fields (highlighted regions). (a) When CW is applied to the  $2^{nd}$  layer; (b) when CW is applied to the  $12^{th}$  layer; (c) when CW is applied to the  $16^{th}$  layer. In every subfigure, the leftmost column contains the most activated image for each concept axis. For each image, we calculate its empirical receptive fields on different axes, shown as the 7 images on the right. The empirical receptive field tends to be larger on the portions of the image that are important for recognizing the correct concept.



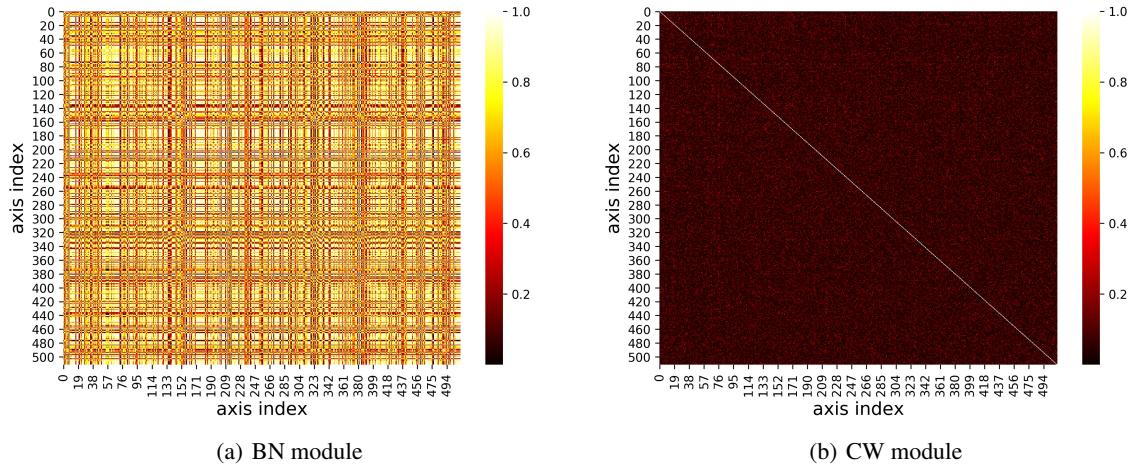
*Supplementary Figure 17.* Normalized intra-concept and inter-concept similarities (ISIC dataset). Diagonal values are normalized average similarities between latent representations of images of the same concept; off-diagonal values are normalized average similarities between latent representations of images of different concepts. (a) when the 16<sup>th</sup> layer is a BN module; (b) when 16<sup>th</sup> layer is a CW module.



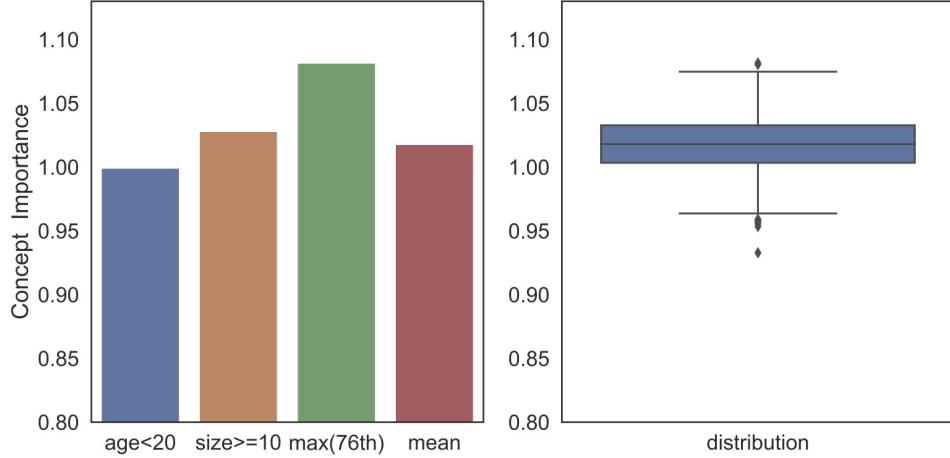
(a) Concept “age &lt; 20”


 (b) Concept “size  $\geq 10 \text{ mm}$ ”

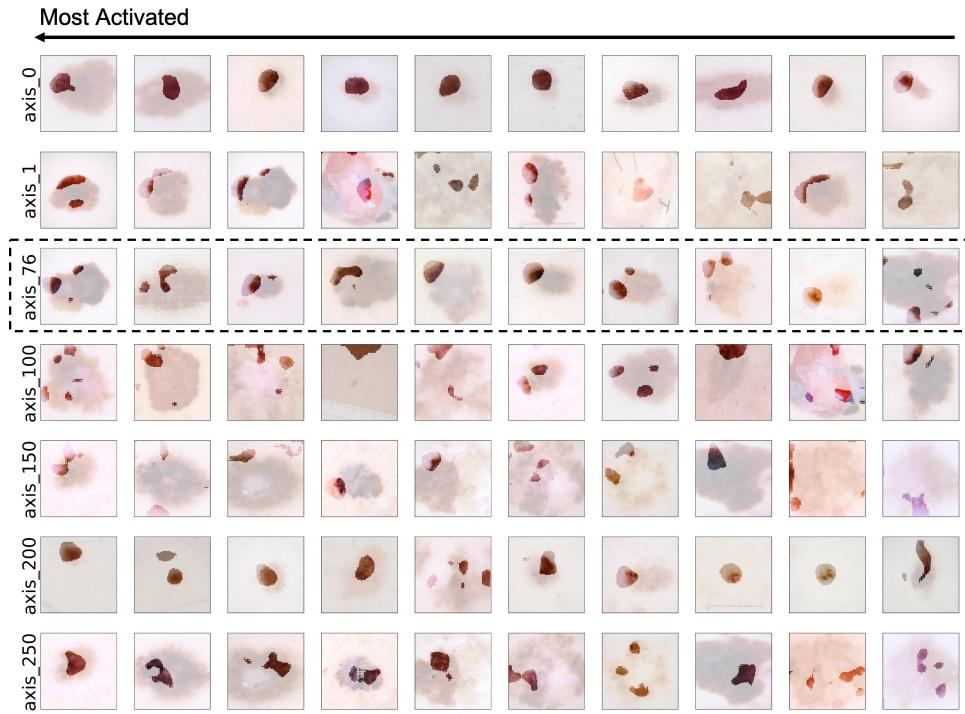
Supplementary Figure 18. Concept purity measured by AUC score (ISIC dataset). Concept purity of CW module is compared to other posthoc methods on different layers. The error bar is the standard deviation over 5 different test sets, and each one is 20% of the entire test set.



Supplementary Figure 19. Absolute correlation coefficient of every feature pair in the 16<sup>th</sup> layer (ISIC dataset). (a) when the 16<sup>th</sup> layer is a BN module; (b) when 16<sup>th</sup> layer is a CW module.



*Supplementary Figure 20.* Concept importance measured on each axis when CW is applied to the 16<sup>th</sup> layer (ISIC dataset). The figure on the left shows the concept importance of the axes representing the concepts “age < 20” and “size  $\geq 10$  mm,” as well as the max and mean concept importance over the set of axes. In order to calculate the latter two quantities, we compute the concept importance of each axis. Then we find the axis with the maximum concept importance (which is the 76th axis) and, for comparison, we calculate the mean of the concept importance values over all the axes. The box plot on the right roughly shows the distribution of concept importance among the 512 axes in the latent space.



*Supplementary Figure 21.* Top 10 activated images on different axes plotted with empirical receptive fields (highlighted region). Axis 76 (most important axis) is highlighted by a dashed box and plotted with other axes (Axis 0, 1, 100, 150, 200 and 250). Axis zero is age, axis one is size, whereas the other axes are not trained as concept axes. Axis 76 seems to more consistently focus on the borders of the lesion, indicating that in future work one might add a concept axis for irregular border.