This ACCV 2020 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2020
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Contextual Semantic Interpretability

Diego Marcos[1], Ruth Fong[2], Sylvain Lobry[1], Rémi Flamary[3⋆],
Nicolas Courty[4], and Devis Tuia[1,5]

[1] Wageningen University, The Netherlands
[2] Oxford University, UK
[3] CMAP, École Polytechnique, Palaiseau, France
[4] IRISA, University Bretagne Sud, CNRS, France
[5] EPFL, Switzerland

Abstract. Convolutional neural networks (CNN) are known to learn an image representation that captures concepts relevant to the task, but do so in an implicit way that hampers model interpretability. However, one could argue that such a representation is hidden in the neurons and can be made explicit by teaching the model to recognize semantically interpretable attributes that are present in the scene. We call such an intermediate layer a *semantic bottleneck*. Once the attributes are learned, they can be re-combined to reach the final decision and provide both an accurate prediction and an explicit reasoning behind the CNN decision. In this paper, we look into semantic bottlenecks that capture *context*: we want attributes to be in groups of a few meaningful elements and participate jointly to the final decision. We use a two-layer semantic bottleneck that gathers attributes into interpretable, sparse groups, allowing them contribute differently to the final output depending on the context. We test our contextual semantic interpretable bottleneck (CSIB) on the task of landscape scenicness estimation and train the semantic interpretable bottleneck using an auxiliary database (SUN Attributes). Our model yields in predictions as accurate as a non-interpretable baseline when applied to a real-world test set of Flickr images, all while providing clear and interpretable explanations for each prediction.

Keywords: interpretability; explainable AI; sparsity

1 Introduction

Deep learning, in particular convolutional neural networks (CNNs), is increasingly being applied to important yet sensitive domains, such as autonomous driving, facial recognition, and medical applications. One significant driver behind the success of CNNs is their capacity to learn to approximate complex functions from large amount of data by automatically tuning millions of parameters. However, this power comes at the expense of interpretability: because of the complexity of CNNs, their internal reasoning can not be easily assessed by humans. This has implications on scientific and societal levels.

⋆ Partially funded through the project OATMIL ANR-17-CE23-0012 and 3IA Cote d'Azur Investments ANR-19-P3IA-0002 of the French National Research Agency.

Fig. 1. We learn contextual groupings (colored coded) of semantic attributes (middle icons) in order to make predictions (right) (e.g., the meaning of "road" depends on the presence of other attributes).

The highly parameterized nature of CNNs enables them to solve a given task in a variety of ways. Some of these solutions might rely on spurious cues that would harm generalization [1]. This is well illustrated by numerous works on adversarial examples [2, 3], in which small perturbations, imperceptible to the human eye, are added to an image and subsequently cause a model to fail. Furthermore, one can easily find thousands of natural images on which CNNs fail (*e.g.*, real-world adversarial examples) [4]. Together, these findings cast doubt on the decision functions learned by CNNs and motivate the need for models that are more transparent in their decision-making process.

As deep learning (and its promise of efficient automation) increasingly affects various aspects of human life, the impenetrable complexity of this class of models also becomes a pressing societal issue. For instance, some governmental entities introduced bills for the regulation of decisions based on algorithms (*e.g.* Equal Credit Opportunity Act in the USA, General Data Protection Regulation in the EU). While research focused on understanding deep learning predictions is on the rise (see section 2), [5] highlights that there still is a gap between the understanding of explanability of the machine learning community and that of lawmakers. Explanations are one way to achieve a degree of interpretability, which can generally be defined as follows: "systems are interpretable if their operations can be understood by a human" [6]. Moreover, the majority of methods that aim to elucidate CNN decisions generate an explanation *a posteriori*; this might induce the risk of a false sense of transparency and trustworthiness [7].

In this paper, we make three main contributions. First, we introduce a novel, explicitly interpretable architecture and training paradigm. Our model first learns to predict an intermediate, semantically explicit task (e.g., predicting attributes). These intermediate attributes are then used to make the final prediction on a downstream task. In particular, we do this by learning a new, sparse, and easily interpretable grouping layer that allows attributes to interact with each other. We call our proposed layer a Contextual Semantic Interpretable Bottleneck (CSIB). Second, we demonstrate the interpretability of our model via a novel combination of visualizations. Using Sankey plots, we visualize both task-specific groups of attributes learned by our model as well as instance-specific explanations that quantify how each attribute (and group) contributed to a model's final prediction. We also highlight the image regions each group captures. Together with details on how much each group contributes to the final score, we are able to visualize with clarity, fidelity, and depth the model's decision-making process. Third, we perform a thorough, empirical analysis of our method applied to the task of scenicness estimation. Here, we demonstrate that

our model performs comparably to a baseline CNN when evaluating a real-world set of Flickr images (there is a performance gap when evaluating on a held-out set from the training distribution). Lastly, we show our paradigm uniquely allows us to identify and explain systematic errors our model makes.

2 Related Work

Post-hoc interpretability. Most interpretability research introduces post-hoc methods that aim to explain any black-box model (see [8] for an interpretability survey). Much attention has been focused on the problem of attribution, *i.e.* the identification of image regions (via heatmaps) that are responsible for a model's output [9–19]. Although attribution methods can be applied to any model, the produced heatmaps lack richness, as they only highlight which image regions are decision relevant, but are unable to characterize a specific semantic reasoning or how those image regions interact. Depending on their formulation, they can also be misleading, as [20] highlights.

Another line of research focuses on understanding the global properties of CNNs. One approach to this problem is to study CNNs in a scientific manner, *i.e.* by generating and testing hypotheses about CNN properties (e.g., sparse vs. distributed encoding [21–24], invariance vs. sensitivity to geometric transformations [15, 25]), or visualizing stimuli preferred by a network [15, 9, 26–31]. Another direction is to summarize a complex model with a simpler, more interpretable model (e.g., a sparse linear classifier or shallow decision tree) [32–36]. Our work is most related to an approach introduced by several recent works [37, 24, 38] that identify how semantic concepts are represented in a network by training linear probes on intermediate features to perform concept classification. While these techniques focus on learning post-hoc how concepts are encoded, our method explicitly learns intermediate features that correspond to concepts.

Interpretability by design. In contrast to post-hoc approaches, "interpretable-by-design" paradigms focus on designing models that are explicitly interpretable. A number of works have proposed models that generate explanations alongside predictions. A few papers utilize multiple modalities in their model to produce explanations [39–41]. Another approach is to include an attention mechanism that constrains information flow [42, 43]. Then, the attended features can be used as an explanation. These are not explicitly designed to be human-interpretable, although [44] constrains attended features to match desired explanations. A shortcoming of models optimized to produce explanations is that there is often a tradeoff between their explanatory and predictive components (e.g., a generated explanation may not be both faithful to the model and easily interpretable).

Another direction focuses on encouraging a model to have interpretable intermediate features. Several works have introduced interpretable variational autoencoders [45, 46] by encouraging the latent space to be disentangled (i.e., independent factors of variation). Regarding image classifiers, [47] constrains features to be sparse, discriminative "parts" detectors, while [48] introduces BagNets, interpretable classifiers that sum up evidence from small input patches.

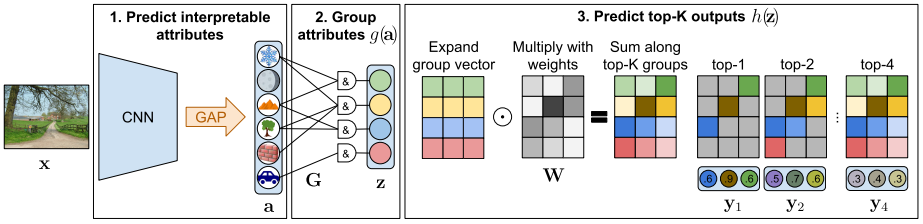

Fig. 2. Model overview. Our CSIB model learns to 1., predict human-interpretable attributes, 2., form sparse groups of attributes that describe broader, task-relevant concepts, and 3., output predictions using the top-K groups.

Our work is most similar to [49] and [50]. [49] introduce "semantic bottleneck networks," which encourage the features of the bottleneck between an encoder and decoder to align with semantic concepts. Their design incurs a negligible loss in accuracy on the final segmentation task. However, the relationship between the semantic bottleneck and output prediction is a highly non-linear decoder, making it difficult to study. To improve the interpretability of the semantic bottleneck, [50] use an semantic bottleneck based on attribute prediction and a linear layer to map onto the final task. The linear mapping makes the relation between the concepts in the bottleneck and the final task easily interpretable; however, this forces each concept to contribute independently and linearly, with no regard for the presence of other concepts.

Context. Visually similar attributes might be understood differently depending on what other elements are present, making contextualization important both for human and machine visual tasks [51, 52]. [53] learns the causal relationship between pairs of object instances (e.g., cars and wheels) as well as the relationships between objects and background context, while [54] leverages a bayesian causal model to explore the impact of counterfactuals using concepts learned from self-supervision. In order to leverage context, ScenarioNet [55] proposes to find "scenarios," groups of commonly co-occurring concepts, by learning a sparse dictionary on the co-occurrence matrix of concepts in the training set. These scenarios are then treated as classes and predicted jointly with the final task, allowing to see which scenarios are present in the image. However, this does not necessarily mean that the final decision is conditioned on the detected scenarios. We propose to connect the semantic bottleneck with the final output by using non-linear, simple, and sparse relations, so that the mapping is transparent and easy to study, while being powerful enough to solve the visual task.

3 Contextual Semantically Interpretable Bottleneck (CSIB)

Our proposed approach is illustrated in Fig. 2 and relies on two steps. First, we train a predictor of binary attributes using a standard `CNN`. The attributes

are then computed by $\mathbf{a} = \texttt{CNN}(\mathbf{x})$ and all carry the semantics encoded by the attribute dataset. They can be object or more complex concepts contained in the images and multiple attributes can be present in a given image $\mathbf{x}$. Second, we train and interpretable function f that uses the attributes to obtain a final prediction $\mathbf{y} = f(\mathbf{a}) = f(\texttt{CNN}(\mathbf{x}))$. The function f should be simple enough so that a human observer can easily understand how each attribute contributes to the result. A common choice for a simple and interpretable function is a linear mapping $\mathbf{y} = \mathbf{W}\mathbf{a}$ [50]. However, such a function is unable to capture an important process for image understanding: *contextualization*. This is because the contribution of an attribute to the output in the linear case is independent on the presence of other attributes.

Our solution consists of learning the groups of interaction by using a composition of two simple, but non-linear, functions: $\mathbf{z} = g(\mathbf{a}; \mathbf{G})$, parametrized by the sparse matrix $\mathbf{G}$, which extracts relevant groups of attributes, and $\mathbf{y} = h(\mathbf{z}; \mathbf{W})$, parametrized by $\mathbf{W}$, which captures the relations between the groups and the output. In the following we detail the three elements that compose our model: the CNN that extracts attributes, and the functions g and h.

3.1 Attribute prediction

We train a standard CNN to predict the presence probability vector $\mathbf{a} \in [0,1]^A$ of A attributes by minimizing the multi-label classification loss based on binary cross entropy, $\mathcal{L}_{\text{attr}}$, on an attribute dataset. Training the model predicting the attributes typically needs labels not available for the dataset used for the final task $\mathbf{y}$. Therefore, to minimize $\mathcal{L}_{\text{attr}}$, we resort to an auxiliary dataset providing the attributes labels via image/attributes. Note that these two datasets can be disjoint and there is no need for images with both types of annotations. By choosing the appropriate set of attributes we are able to obtain a model that makes use of the desired visual cues while being invariant to undesired attributes.

3.2 Attribute grouping function $g(\cdot)$

Given the attributes $\mathbf{a}$ predicted in the images, we now want to group them into semantically meaningful groups. To do so, we use a grouping function $\mathbf{z} = f(\mathbf{a})$ that groups attributes together into a vector of group presence probabilities $\mathbf{z} \in [0,1]^Z$, with Z the number of groups. This function is parametrized by the sparse non-negative matrix $\mathbf{G} \in [0,1]^{Z \times A}$. Each row $\mathbf{G}_{i,:}$ represents one group and is constrained on the probability simplex ($G_{i,j} \geq 0, \sum_j G_{i,j} = 1, \forall i$) by orthogonal projection after each SGD step [56]. The output for group z_i is computed as:

$$z_i = \prod_{j=1\ldots A} a_i^{\mathbf{G}_{i,j}}. \tag{1}$$

This corresponds to a weighted geometric mean and acts as a soft-AND logical function, which means that a group i will only be fully active ($z_i = 1$) if, for every attribute j required by the group ($G_{i,j} > 0$), the attribute is fully present

($a_j = 1$). Also, it suffices that one of these attributes is absent ($a_j = 0$) to result in $z_i = 0$. Since all the attributes for which $G_{i,j} > 0$ must be present for the group to become active, $\mathbf{G}$ tends to become sparse during the learning process. This is a direct consequence of the projection onto the simplex which is naturally sparse. To increase numerical stability, the operation is implemented as a standard linear mapping over the log probabilities of the attributes:

$$\mathbf{z} = e^{\mathbf{G}\log(\mathbf{a})}, \tag{2}$$

which allows us to use a numerically stable implementation of log-sum-exp.

Unsupervised group pretraining The soft-AND function in Eq. (1) will output values close to zero if one or more of the attributes that correspond to a high weight $G_{i,j}$ are not present. Therefore, initializing $\mathbf{G}$ with random weights, encoding for random groups that are thus not very likely to exist, results in mostly inactive groups and a very low learning signal. To make sure that $\mathbf{G}$ is initialized with groups that are present in the dataset, we first minimize the following loss on $\mathbf{Z} \in [0,1]^{B\times Z}$ corresponding to the concatenation for a batch of images with batch size B :

$$\mathcal{L}_{\text{groups}}(\mathbf{Z}) = \mathcal{L}_{\text{on}}(\mathbf{Z}) + \mathcal{L}_{\text{off}}(\mathbf{Z}) + \mathcal{L}_{H}(\mathbf{Z}). \tag{3}$$

The first two terms are designed to encourage the groups to become diverse. For a group to be of any use, it needs to be active in at least a few images. In addition, we want to make sure that at least one group is active per image. For this reason, we encourage the highest values along each row and each column of $\mathbf{Z}$ to be close to one, the highest possible value:

$$\mathcal{L}_{\text{on}}(\mathbf{Z}) = -\sum_{i=1}^{Z} \max_{u}(Z_{ui}) - \sum_{u=1}^{B} \max_{i}(Z_{ui}). \tag{4}$$

At the same time, we want to make sure that no particular group is active in all the samples of the batch, because such group would not be a discriminative one. We therefore minimize the maximum of the lowest per-group values:

$$\mathcal{L}_{\text{off}}(\mathbf{Z}) = \max_{u}(\min_{i}(Z_{u,i})). \tag{5}$$

However, this is not enough to guarantee the diversity in the groups. Ideally, we would want the batch-wise vector of group activations $\mathbf{Z}_{:,i}$ and $\mathbf{Z}_{:,j}$ of any two groups to be as different as possible. Simultaneously, we would like the groups to help discriminate between images, and thus the sample-wise vectors of group activations $\mathbf{Z}_{u,:}$ and $\mathbf{Z}_{v,:}$ of any pair of images should also be as different as possible. We encourage this by maximizing the cross-entropy $H(\mathbf{u},\mathbf{v}) = -\sum_i u_i \log(v_i)$ between all pairs of per-group activation vectors and all pairs of per-sample activation vectors:

$$\mathcal{L}_{H}(\mathbf{Z}) = -\sum_{i,j\neq i} H\left(\frac{\mathbf{Z}_{:,i}}{\sum_k Z_{k,i}}, \frac{\mathbf{Z}_{:,j}}{\sum_k Z_{k,j}}\right) - \sum_{i,j\neq i} H\left(\frac{\mathbf{Z}_{i,:}}{\sum_k Z_{i,k}}, \frac{\mathbf{Z}_{j,:}}{\sum_k Z_{j,k}}\right) \tag{6}$$

Note that the regularization term above will have the effect of promoting the activations across groups and images to the maximally independent, tending towards source separation. Minimizing $\mathcal{L}_{\text{groups}}$ provides $\mathbf{G}$ with a set of initial groups that do occur in some images ($\mathcal{L}_{\text{on}}$) but not in all ($\mathcal{L}_{\text{off}}$) and that are discriminative and different from each other ($\mathcal{L}_H$).

3.3 Output contribution function $h(\cdot)$

Given the group activations $\mathbf{z}$, we want a function $\mathbf{y} = h(\mathbf{z})$ that produces the desired final output $\mathbf{y} \in \mathbb{R}^Y$. Function h is parametrized by matrix $\mathbf{W} \in \mathbb{R}^{Y \times Z}$.

We want as few groups as possible to contribute to the output $\mathbf{y}$. This can be enforced by taking only the top-K most contributing groups to compute $\mathbf{y}$, as proposed in [57]. The following steps have to be taken:

- A matrix element-wise multiplication $\mathbf{Y} = \mathbf{W} \circ \mathbf{z}$, where $\mathbf{z}$ is broadcasted to the shape of $\mathbf{W} \in \mathbb{R}^{Y \times Z}$.
- A sparsification of $\mathbf{Y}$ by keeping only the top-K values in each row and setting the rest to zero.
- Row-wise sum to obtain $\mathbf{y}$.

In order to avoid choosing a value of K a priori, we compute $\mathbf{y}$ using multiple K values and apply a loss to each output. The specific loss used at this stage is problem-dependent (*e.g.* MSE for regression, cross entropy for classification, *etc.*). The average of such losses is the final output loss, $\mathcal{L}_y$.

3.4 CSIB training strategy

The training procedure to minimize the described losses consists of three steps:

1. Train the CNN to predict the concepts in the semantic bottleneck by minimizing $\mathcal{L}_{\text{attr}}$. Note that this provides no learning signal to $\mathbf{G}$ nor $\mathbf{W}$.
2. Keeping the weights of the CNN frozen, minimize $\mathcal{L}_{\text{groups}}$ to initialize $\mathbf{G}$ with relevant and discriminative groups.
3. Finetune the whole model end-to-end on the final task by minimizing $\mathcal{L}_{\text{attr}} + \lambda\mathcal{L}_y$, with $\lambda << 1$ to ensure that the performance on attribute prediction is not degraded.

4 Experiments in landscape scenicness prediction

4.1 Experimental set-up

In the experiments below, we aim at predicting the scenicness (*i.e.* landscape beauty) score of a collection of images. The training images come from the ScenicOrNot [58] dataset, collected across Great Britain, and where each image has an average scenicness score (between 1 and 10), obtained by crowdsourcing. Out of the 212,104 available images, we used the first 180,000, ordered by image

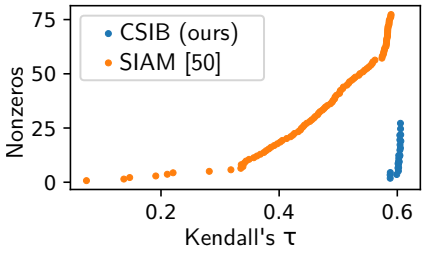

Fig. 3. Average number of attributes with a nonzero contribution to the final output against the resulting Kendall's τ score. The number of nonzeros was varied by increasingly pruning the smaller contributions. Our model allows for much more concise explanations.

ID, for training, the following 5,000 for validation and the rest we held out for testing. Given the subjectivity of the final task, we want to make the reasoning of the model explicit by using a semantic bottleneck that detects attributes occurring in the image as an intermediate task: therefore, the semantic bottleneck is trained to predict the presence of the 102 classes of the SUN Attributes [59] dataset. We use the same train-test splits as in [59]. Previous works have already established that there is a correlation between some of these attributes and scenicness [60, 61], and with SCIB we aim at constraining this further and explain scenicness using exclusively this pre-defined set of attributes.

For attribute prediction, we finetune a ResNet-50 [62], pre-trained on ImageNet [63]. We remove the last layer of the pre-trained model and add a 1×1 convolutional layer to map down the 2048 activation maps to the 102 SUN attributes, followed by a global average pooling. The model is trained using Stochastic Gradient Descent (SGD) with 0.9 momentum for 20,000 iterations with a batch size of 10. The learning rate is initially 0.002 and is decayed by a factor of 4 after 10,000 and 15,000 iterations. In the second step (group initialization), we initialize the sparse grouping matrix $\mathbf{G}$ with 150 groups in an unsupervised way by minimizing $\mathcal{L}_{\text{groups}}$. The learning rate was fixed at 0.002 for 4,000 iterations with a batch size of 100. Note that, as mentioned in section 3.4 above, the ResNet-50 base model was left frozen, allowing for a larger batch size, which is important to capture the diversity between groups in $\mathcal{L}_{\text{groups}}$. As a last step (finetuning), $\lambda\mathcal{L}_y$ is minimized along with $\mathcal{L}_{\text{attr}}$ for 50,000 iterations and $\lambda = 0.1$. This time the whole model is trained end-to-end and two batches of size 10 are used in every iteration, one from SenicOrNot and one from SUN Attributes. The learning rate is initially 0.002 and is decayed by a factor of 4 after 10,000 and 20,000 and 30,000 iterations. We train simultaneously with nine levels of top-K sparsity: $K = 1, 2, ..., 8, 150$. Since $\mathbf{W}$ is initialized will all zeros, the dense branch is important to make sure that all groups receive a learning signal. The bias of the last layer ($\mathbf{W}$) is fixed to the average scenicness value on the training set, a score of 4.43, and is kept constant.

4.2 Numerical comparisons within ScenicOrNot

On the ScenicOrNot test set, both CSIB and the baseline are able to generalize well, with CSIB showing a small drop in performance in terms of Kendall's τ [64] and root mean square error (RMSE), comparable to the one observed

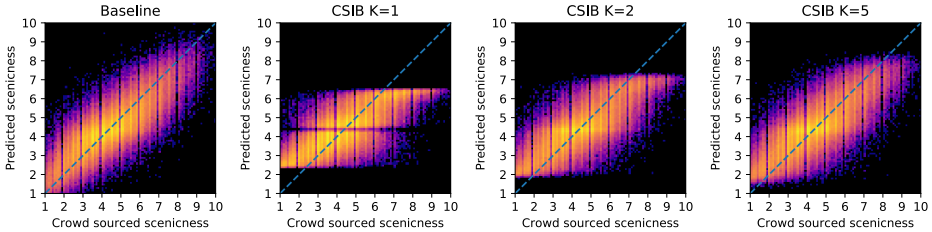

Fig. 4. Scatter plots showing our CSIB model's predictions for different K, which controls the number of groups used to make the final prediction (see fig. 5 for groups).

in [50] (see Tab. 1). However, [50] requires many more attributes to contribute to the result compared to CSIB, as shown in Fig. 3, making the explanations provided by CSIB more desirable in terms of the number of required *cognitive chunks* [65]. This highlights the effectiveness in terms of sparsification of the proposed constrained optimization. In addition, we observe only a minor degradation in the performance on the task of attribute prediction with respect to a baseline trained exclusively on that task. CSIB enables us to choose at test time the number of groups that can contribute to the final result by setting the K parameter in the top-K activation layer. Fig. 4 depicts the results for $K \in \{1, 2, 5\}$ in the form of scatter plots. When using $K = 1$, CSIB is not able to predict very high or very low values, since the maximum deviation from the average it can predict is $[-2.14, 2.29]$, which corresponds to the contribution of the most contributing groups (see red and yellow groups in Fig. 5). At the same time, values close to the average are also missed, since the top-1 layer is required to choose the single most contributing group among the groups present, forcing the output away from the average. This undesirable behaviour is already corrected by setting $K = 2$. The accuracy saturates when using $K = 5$, where the model is capable of predicting more extreme values of scenicness in a comparable way to the baseline, although it maintains a bias towards the average in the extreme cases. Such a small value of K, together with the sparsity of $\mathbf{G}$, allows to easily understand the relations encoded in CSIB (see Section 4.3). We observed that finetuning the whole model end-to-end (step 3 in Section 3.4) was important to obtain the mentioned results, with a Kendall's τ of 0.468 before finetuning.

		CSIB			
	baseline	$K = 1$	$K = 2$	$K = 5$	$K = 7$
SoN Kendall's τ	0.645	0.580	0.603	0.609	0.609
SoN RMSE	0.940	1.111	1.037	1.018	1.019
SUN AP	0.610	0.601			

Table 1. Task performance. ScenicOrNot (SoN) results are reported using Kendall's τ ranking metric and root mean square error (RMSE); average precision (AP) is reported for SUN (higher is better for τ and AP; lower is better for RMSE). Performance plateaus at $K = 5$ for our CSIB model; our model underperforms the baseline.

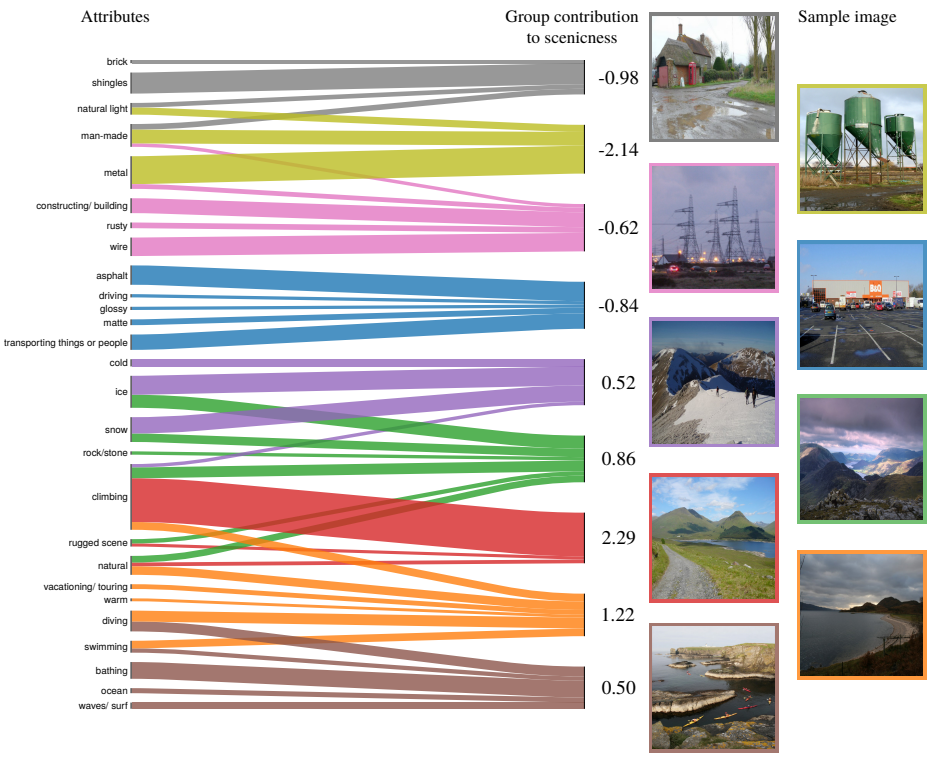

Fig. 5. Learned groups for scenicness estimation. Line thickness denotes the contribution of a SUN Attribute (left) to a group. For each group, we show its scenicness score (middle) and an example image (right). The groups and their scores appear coherent and consistent with the task. They are also sparse and interpretable: only 27 of 102 attributes are chosen and only 9 groups (out of a possible 150) become relevant.

4.3 Visualization of the model

Entire model. The semantics captured by SCIB, together with the high sparsity, allow us to comprehend the reasoning used to compute the output from the attributes in the semantic bottleneck. Fig. 5 depicts the model by showing the contribution of each attribute to the groups (the weights in $\mathbf{G}$) that contribute more than 0.5 score points towards the scenicness values. This already provides a good understanding of the relations learned and encoded in the model. For instance, the last two groups (orange and brown) show that "diving" and "swimming" are assigned higher scores if they co-occur with "climbing" and "natural." We also see that it typically assigns high scores to wilderness-related attributes and low scores for those related to man-made elements. In the same figure, we also provide an image that scores strongly for that group.

Individual results. Individual decisions for specific images are also easily interpretable. Using the activations in CSIB, we can now visualize which paths are

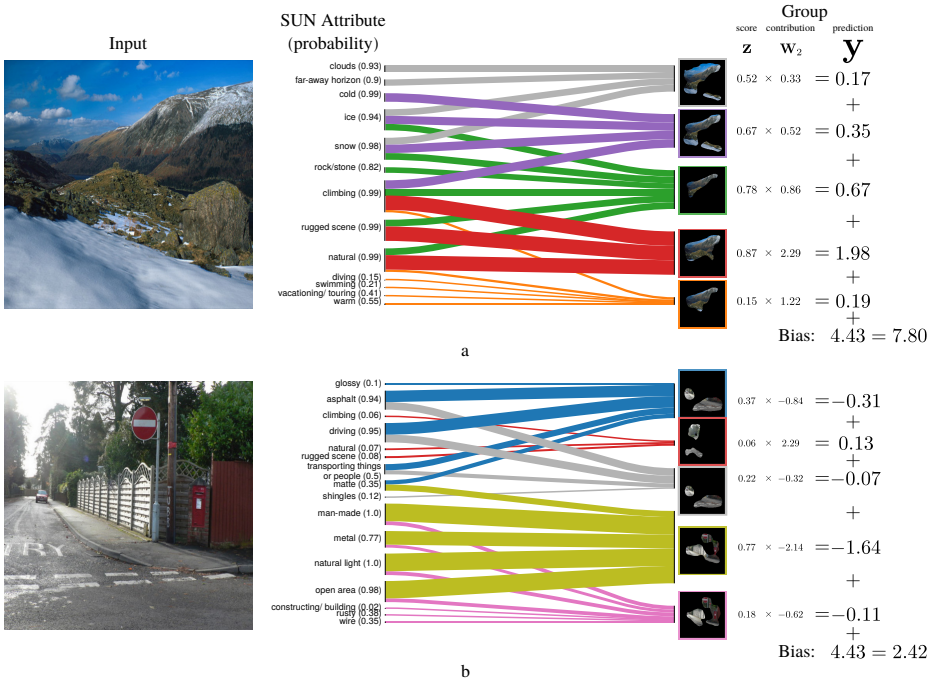

Fig. 6. Diagram of attribute contribution for a scenic image (top: GT = 9.43, baseline = 8.39) and an unscenic one (bottom: GT = 1.6, baseline = 2.18). These explanations point to a sensible decision-making process.

followed to reach the final decision. Fig. 6 shows two examples with $K = 5$. In this figure, the thickness of the lines is proportional to the contribution of the attribute to the group, which depends on the presence of the other attributes required by the group due to the multiplicative nature of Eq. (1). The part of the images contributing the most to the groups is depicted using thresholded activation maps. In these two cases, we can see how the explanations suit the images and our preconceptions of landscape beauty, with the first one rated with a 7.8/10 due to the rugged snowy mountain scene and the second one a mere 2.4/10 because of its man-made nature. On the other hand, Fig. 7 depicts the same visualization for two images in which there is a strong disagreement with the crowdsourced value. In the first case, the man-made look of the image and the transport-related aspect of the boat trigger the model to predict a low score, while in the second case the ruggedness and climbing-related aspects dominate, while the graffiti and the narrow view of the image are ignored, since these cannot be captured by the attributes used.

4.4 Validation of the group predictions by geographical distribution

Being the attributes learned from a dataset that is disjoint from the one used for the final task (*i.e.* we have no test set of the 102 SUN attributes on the

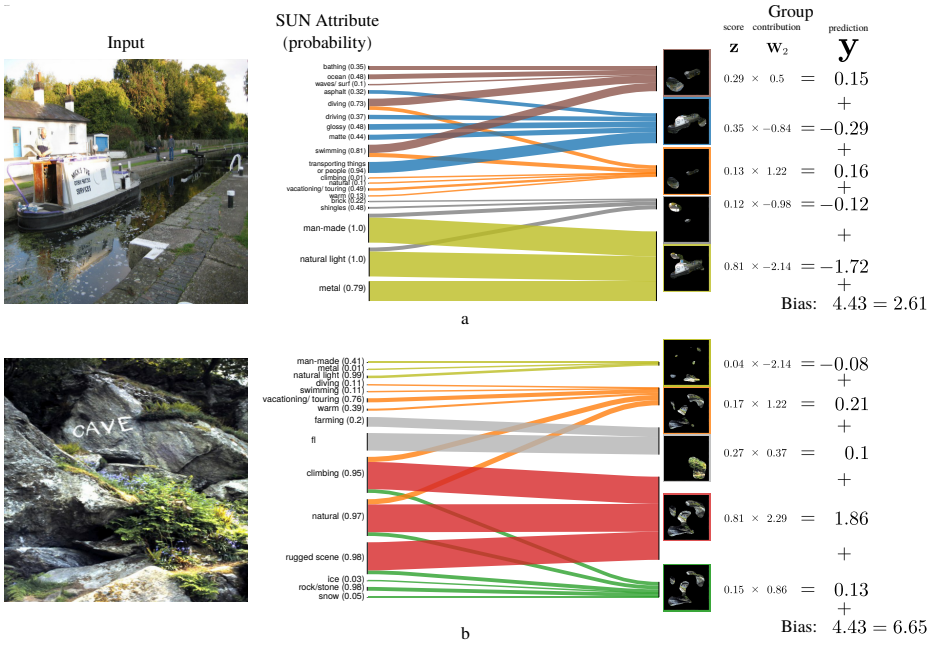

Fig. 7. Diagram of attribute contribution for an underpredicted image (top: GT = 7.17, baseline = 3.66) and an overpredicted one (bottom: GT = 2.11, baseline = 6.65). Our CSIB model allows us to understand why it disagrees with ground truth annotations.

ScenicOrNot images), we evaluate the performance of the attribute prediction on the SoN images qualitatively, by mapping the geographical distribution of the average activation of the groups (Fig. 8a) and comparing them to the 2012 CORINE land-cover map [66] of Great Britain (Fig. 8b). The learned groups with mountain-related attributes show a good overlap with the bare soil and rock surfaces in the landcover map, and the group that also includes snow and ice is more present in mountainous regions of Scotland. The groups with man-made attributes overlap with urban areas, and the ones with water activities are most active along the coast and in the lake filled northwest. These results suggest a good performance of the attribute detector on the SoN image dataset.

4.5 Generalization: numerical results on 1.7M Flickr images

In order to test its generalization capabilities, we applied both CSIB and the baseline over a large set of 1.7 million geo-located outdoor images obtained from Flickr. Although no scenicness ground truth is available for these images, we can create a map of scenicness based on the values predicted on the Flickr dataset (depicted in Fig. 9b, c, e and f), and compare it to the map obtained using the ScenicOrNot ground truth (Fig. 9a and d). In Tab. 2 we show the results of comparing these values averaged over grids of different size (5000 × 5000, 500 × 500 and 50 × 50 bins across the region), and we consistently see

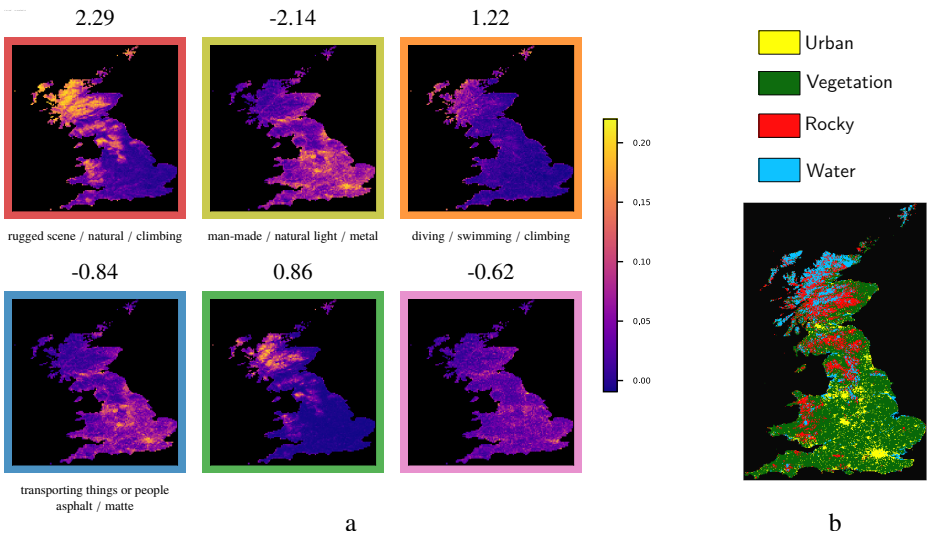

Fig. 8. a. Geographical distribution of group activations. For each image, the number on top is the group's scenicness score, and the three most contributing attributes for the group are shown below. Group colors are taken from fig. 5. **b. Land-cover of Great Britain.** Data from [66].

that the results of the baseline and CSIB are numerically equivalent, with the CSIB being slightly better in terms of RMSE and behind in terms Kendall's τ. This suggests that the better performance of the baseline on ScenicOrNot might be partially attributable to over-fitting to the dataset, and that restricting the model to be inherently interpretable with CSIB reduces its capacity to overfit, but not the capability to generalize. Fig. 9 shows maps of scenicness at two different scales (100 and 5000 bins) using the ScenicOrNot ground truth and CSIB results on the Flickr images. Fig. 9a/b at the country level, showcases the agreement of both maps. Fig. 9c/d show the maps for London. At this scale the sparsity of ScenicOrNot becomes apparent. On the map produced by CSIB, scenicness seems to be predicted highest in green areas (such as the Hyde and Regent's parks) and lowest in areas of transport infrastructure, such as railroads and airport (see Fig. 9d). These conclusions regarding the notion of scenicness validate the relations captured by CSIB from the dataset and that are clearly visible and interpretable from the model itself (as seen in Fig. 5).

5 Conclusion

We presented a paradigm to make the decision making process of a CNN inherently interpretable, which we call a Contextual Semantic Interpretable Bottleneck (CSIB). A standard CNN is trained to detect human-interpretable attributes. These attributes are then used to determine the final decision in a contextual and sparse manner (*i.e.* the meaning an attribute takes on is dependent on what other attributes are present, and only a select subset of attributes

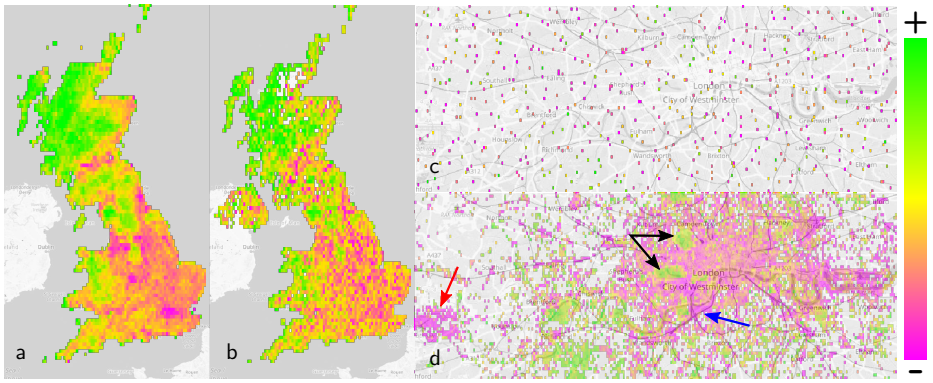

Fig. 9. Geographical distribution of scenicness at the national level (a and b) and in London (c and d). Panels a and c show the ground truth scenicness interpolated from ScenicOrNot; panels b and d show our CSIB model's predictions. In panel d, a few London regions are quite salient, such as Hyde and Regent's parks (black arrows), Heathrow airport (red arrow) and a railway intersection (blue arrow).

	5000 bins		500 bins		50 bins	
	baseline	CSIB	baseline	CSIB	baseline	CSIB
Kendall's τ	0.399	0.391	0.384	0.382	0.624	0.621
RMSE	1.528	1.497	1.213	1.166	0.749	0.679

Table 2. Performance on Flickr images. We evaluate models on 1.7M Flickr images in Great Britain and bin the predictions spatially at different scales; we then compare the spatial predictions against ScenicOrNot ground truth averages. Our CSIB model performs comparably to the baseline (higher is better for τ; lower is better for RMSE).

are used to form a small number of groups). This makes it possible to understand what relationships the model has learned by simply inspecting its weights as well as which of these relationships have been used for an individual image. Note that CSIB requires an auxiliary dataset containing attributes relevant to the final task. Nevertheless, the same attribute predictor can be reused for multiple downstream tasks; this would also enable model comparisons *across* tasks and reveal what attribute groupings are relevant to which tasks.

We demonstrate the validity of our method on a scenicness estimation task; we use the ScenicOrNot dataset and also evaluate on a large set (1.7M images) of real-world images from Flickr. CSIB is able to generate a map of scenicness to the same level of accuracy as that of a non-interpretable baseline. Lastly, we show how visualization techniques can be combined in order to explain what (and how) visual information has been leveraged in our model's decision making process. This allows us to understand the instances in which our model disagrees with the labelled annotation, among other things. In conclusion, we introduce a novel architecture that is both inherently interpretable and powerful enough so as to not sacrifice in real-world performance. This suggests that the assumed tradeoff between interpretability and performance may not always be necessary.

References

1. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature communications **10** (2019) 1096
2. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
3. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
4. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. arXiv preprint arXiv:1907.07174 (2019)
5. Edwards, L., Veale, M.: Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. Duke L. & Tech. Rev. **16** (2017) 18
6. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 workshop on explainable AI (XAI). Volume 8. (2017) 1
7. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1** (2019) 206–215
8. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE (2018) 80–89
9. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proc. ICLR workshop. (2014)
10. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. (2015)
11. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proc. CVPR. (2016)
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proc. ICCV. (2017)
13. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. IJCV **126** (2018) 1084–1102
14. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10** (2015) e0130140
15. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proc. ECCV. (2014)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: SIGKDD. (2016)
17. Fong, R., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proc. ICCV. (2017)
18. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proc. BMVC. (2018)
19. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: Proc. ICCV. (2019)
20. Adebayo, J., Gilmer, J., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proc. NeurIPS. (2018)

21. Morcos, A.S., Barrett, D.G., Rabinowitz, N.C., Botvinick, M.: On the importance of single directions for generalization. arXiv preprint arXiv:1803.06959 (2018)
22. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Revisiting the importance of individual units in CNNs via ablation. arXiv preprint arXiv:1806.02891 (2018)
23. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proc. CVPR. (2017)
24. Fong, R., Vedaldi, A.: Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proc. CVPR. (2018)
25. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. IJCV (2018)
26. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proc. CVPR. (2015)
27. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Proc. NeurIPS. (2016)
28. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 6541–6549
29. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill **2** (2017) e7
30. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proc. CVPR. (2018)
31. Mordvintsev, A., Pezzotti, N., Schubert, L., Olah, C.: Differentiable image parameterizations. Distill **3** (2018) e12
32. Bastani, O., Kim, C., Bastani, H.: Interpretability via model extraction. arXiv (2017)
33. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. arXiv (2017)
34. Tan, S., Caruana, R., Hooker, G., Koch, P., Gordo, A.: Learning global additive explanations for neural nets using model distillation. In: Proc. NeurIPS Workshop. (2018)
35. Zhang, Q., Cao, R., Shi, F., Wu, Y.N., Zhu, S.C.: Interpreting cnn knowledge via an explanatory graph. In: Proc. AAAI. (2018)
36. Zhang, Q., Yang, Y., Ma, H., Wu, Y.N.: Interpreting cnns via decision trees. In: Proc. CVPR. (2019)
37. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: Proc. ECCV. (2018)
38. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In: Proc. ICML. (2018)
39. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: Proc. ECCV, Springer (2016)
40. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. In: Proc. CVPR. (2017)
41. Huk Park, D., Anne Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: Proc. CVPR. (2018)
42. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proc. CVPR. (2015)
43. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Proc. NIPS. (2016)

44. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017)
45. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Proc. NIPS. (2016)
46. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: Proc. ICLR. (2017)
47. Zhang, Q., Nian Wu, Y., Zhu, S.C.: Interpretable convolutional neural networks. In: Proc. CVPR. (2018)
48. Brendel, W., Bethge, M.: Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In: Proc. ICLR. (2019)
49. Losch, M., Fritz, M., Schiele, B.: Interpretability beyond classification output: Semantic bottleneck networks. arXiv preprint arXiv:1907.10882 (2019)
50. Marcos, D., Lobry, S., Tuia, D.: Semantically interpretable activation maps: what-where-how explanations within CNNs. arXiv preprint arXiv:1909.08442 (2019)
51. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in cognitive sciences **11** (2007) 520–527
52. Barenholtz, E.: Quantifying the role of context in visual object recognition. Visual Cognition **22** (2014) 30–56
53. Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., Bottou, L.: Discovering causal signals in images. In: Proc. CVPR. (2017)
54. Harradon, M., Druce, J., Ruttenberg, B.: Causal learning and explanation of deep neural networks via autoencoded activations. arXiv (2018)
55. Daniels, Z.A., Metaxas, D.: Scenarionet: An interpretable data-driven model for scene understanding. In: IJCAI Workshop on XAI. (2018) 33
56. Shalev-Shwartz, S., Singer, Y.: Efficient learning of label ranking by soft projections onto polyhedra. Journal of Machine Learning Research **7** (2006) 1567–1599
57. Sun, Y., Ravi, S., Singh, V.: Adaptive activation thresholding: Dynamic routing type behavior for interpretability in convolutional neural networks. In: Proc. ICCV. (2019)
58. : ScenicOrNot. http://scenicornot.datasciencelab.co.uk (2020) Accessed: 2020-03-03.
59. Patterson, G., Xu, C., Su, H., Hays, J.: The sun attribute database: Beyond categories for deeper scene understanding. IJCV **108** (2014) 59–81
60. Seresinhe, C.I., Preis, T., Moat, H.S.: Using deep learning to quantify the beauty of outdoor places. Royal Society open science **4** (2017) 170170
61. Workman, S., Souvenir, R., Jacobs, N.: Understanding and mapping natural beauty. In: Proc. ICCV. (2017) 5589–5598
62. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR. (2016)
63. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)
64. Kendall, M.G.: A new measure of rank correlation. Biometrika **30** (1938) 81–93
65. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
66. : CORINE Land Cover – Copernicus Land Monitoring Service. https://land.copernicus.eu/pan-european/corine-land-cover (2020) Accessed: 2020-03-03.