Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Learnability, and Human Capability

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky
Princeton University
{vr23, suhk, ruthfong, olgarus}@cs.princeton.edu

Abstract

Concept-based interpretability methods aim to explain a deep neural network model's components and predictions using a pre-defined set of semantic concepts. These methods evaluate a trained model on a new, "probe" dataset and correlate the model's outputs with concepts labeled in that dataset. Despite their popularity, they suffer from limitations that are not well-understood and articulated in the literature. In this work, we identify and analyze three commonly overlooked factors in concept-based explanations. First, we find that the choice of the probe dataset has a profound impact on the generated explanations. Our analysis reveals that different probe datasets lead to very different explanations, suggesting that the generated explanations are not generalizable outside the probe dataset. Second, we find that concepts in the probe dataset are often harder to learn than the target classes they are used to explain, calling into question the correctness of the explanations. We argue that only easily learnable concepts should be used in concept-based explanations. Finally, while existing methods use hundreds or even thousands of concepts, our human studies reveal a much stricter upper bound of 32 concepts or less, beyond which the explanations are much less practically useful. We discuss the implications of our findings and provide suggestions for future development of concept-based interpretability methods. Code for our analysis and user interface can be found at https://github.com/princetonvisualai/OverlookedFactors

1. Introduction

Performance and opacity are often correlated in deep neural networks: the highly parameterized nature of these models that enable them to achieve high task accuracy also reduces their interpretability. However, in order to responsibly use and deploy them, especially in high-risk settings such as medical diagnoses, we need these models to be interpretable, i.e., understandable by people. With the growing recognition of the importance of interpretability, many methods have been proposed in recent years to explain some aspects of neural networks and render them more interpretable (see [4, 14, 18, 42, 44, 53] for surveys).

In this work, we dive into *concept-based* interpretability methods for image classification models, which explain model components and/or predictions using a pre-defined set of semantic concepts [5, 16, 25, 29, 56]. Given access to a trained model and a set of images labelled with semantic concepts (i.e., a "probe" dataset), these methods produce explanations with the provided concepts. See Fig. 1 for an example explanation.

Concept-based methods are a particularly promising approach for bridging the interpretability gap between complex models and human understanding, as they explain model components and predictions with human-interpretable units, i.e., semantic concepts. Recent work finds that people prefer concept-based explanations over other forms (e.g., heatmap and example-based) because they resemble human reasoning and explanations [27]. Further, concept-based methods uniquely provide a *global*, high-level understanding of a model, e.g., how it predicts a certain class [39, 56] and what the model (or some part of it) has learned [5, 16, 25]. These insights are difficult to gain from *local* explanation methods that only provide an explanation for a single model prediction, such as saliency maps that highlight relevant regions within an image.

However, existing research on concept-based interpretability methods focuses heavily on new method development, ignoring important factors such as the probe dataset used to generate explanations or the concepts composing the explanations. Outside the scope of concept-based methods, there have been several recent works that study the effect of different factors on explanations. These works, however, are either limited to saliency maps [1, 28, 31, 41] or a general call for transparency, e.g., include more information when releasing an interpretability method [47].

In this work, we conduct an in-depth study of commonly overlooked factors in concept-based interpretability methods. Concretely, we analyze four representative methods:

```
prediction: bedroom
explanation: + 4.2 bed  - 1.5 coffee-table + 1.3 sky - 1.3 sofa
             - 1.0 drinking-glass - 0.9 television - 0.9 sconce
             - 0.8 chair + 0.8 windowpane + 0.7 blind + 0.7 fan - 0.6 armchair
             - 0.6 sink - 0.6 switch + 0.5 box - 0.5 plate - 0.5 ottoman
             - 0.5 paper + 0.4 cushion - 0.4 tray + 0.4 bottle - 0.4 bowl - 0.4 bag
             + 0.3 chest-of-drawers - 0.3 curtain - 0.3 chandelier - 0.3 table + 0.3 clock
             + 0.3 pot - 0.3 wall - 0.2 telephone - 0.2 fireplace + 0.2 ceiling + 0.2 carpet
             + 0.2 book - 0.1 spotlight -  0.1 flower
```

Figure 1. **Concept-based interpretability methods** explain model components and/or predictions using a pre-defined set of semantic concepts. In this example, a scene classification model's prediction `bedroom` is explained as a complex linear combination of 37 visual concepts, with the final explanation score calculated based on the presence or absence of these concepts. The coefficients are learned by evaluating the model on a new, "probe" dataset, and correlating its predictions with visual concepts labeled in that dataset. However, concept-based explanations can (1) be noisy and heavily dependent on the probe dataset, (2) use concepts that are hard to learn (all concepts in `red` are harder to learn than the class `bedroom`) and (3) be overwhelming to people due to the complexity of the explanation.

NetDissect [5], TCAV [25], Concept Bottleneck [29] and IBD [56]. These are a representative and comprehensive set of existing concept-based interpretability methods for computer vision models. Using multiple probe datasets (ADE20k [57, 58] and Pascal [13] for NetDissect, TCAV and IBD; CUB-200-2011 [48] for Concept Bottleneck), we examine the effects of (1) the choice of probe dataset, (2) the concepts used within the explanation, and (3) the complexity of the explanation. Through our analyses, we learn a number of key insights, which we summarize below:

- **The choice of the probe dataset has a profound impact on explanations**. We repeatedly find that different probe datasets give rise to different explanations, when explaining the same model with the same interpretability method. For instance, the prediction of the `arena/hockey` class is explained with concepts {`grandstand`, `goal`, `ice-rink`, `skate-board`} with one probe dataset, and {`plaything`, `road`} with another probe dataset. We highlight that concept-based explanations are not solely determined by the model or the interpretability method. Hence, probe datasets should be chosen with caution. Specifically, we suggest using probe datasets whose data distribution is similar to that of the dataset the model-being-explained was trained on.

- **Concepts used in explanations are frequently harder to learn than the classes they aim to explain**. The choice of concepts used in explanations is dependent on the available concepts in the probe dataset. Surprisingly, we find that learning some of these concepts is harder than learning the target classes. For example, in one experiment we find that the target class `bathroom` is explained using concepts {`toilet`, `shower`, `countertop`, `bathtub`, `screen-door`}, all of which are harder to learn than `bathroom`. Moreover, these concepts can be hard for people to identify, limiting the usefulness of these explanations. We argue that learnability is a necessary (albeit not sufficient) condition for the correctness of the explanations, and advocate for future explanations to only use concepts that are easily learnable.[1]

- **Current explanations use hundreds or even thousands of concepts, but human studies reveal a much stricter upper bound.** We conduct human studies with 125 participants recruited from Amazon Mechanical Turk to understand how well people reason with concept-based explanations with varying number of concepts. We find that participants struggle to identify relevant concepts in images as the number of concepts increases (the percentage of concepts recognized per image decreases from $71.7\% \pm 27.7\%$ with 8 concepts to $56.8\% \pm 24.9\%$ for 32 concepts). Moreover, the majority of the participants prefer that the number of concepts be limited to 32. We also find that concept-based explanations offer little to no advantage in predicting model output compared to example-based explanations (the participants' mean accuracy at predicting the model output when given access to explanations with 8 concepts is $64.8\% \pm 23.9\%$ whereas the accuracy when given access to example-based explanations is $60.0\% \pm 30.2\%$).

These findings highlight the importance of vetting intuitions when developing and using interpretability methods. We have open-sourced our analysis code and human study user interface to aid with this process in the future: https://github.com/princetonvisualai/OverlookedFactors.

2. Related work

Interpretability methods for computer vision models range from highlighting areas within an image that contribute to a model's prediction (i.e., saliency maps) [9, 15, 37, 45, 46, 51, 52, 54] to labelling model components (e.g., neurons) [5, 16, 25, 56], highlighting concepts that contribute to the model's prediction [39, 56] and designing models that are interpretable-by-design [8, 10, 29, 34]. In this work, we

[1] Ideally, future methods would also include *causal* rather than purely correlation-based explanations.

focus on concept-based interpretability methods. These include post-hoc methods that label a trained model's components and/or predictions [5, 16, 25, 39, 56] and interpretable-by-design methods that use pre-defined concepts [29]. We focus on methods for image classification models where most interpretability research has been and is being conducted. Recently, concept-based methods are being developed and used for other types of models (e.g., image similarity models [38], language models [7, 50]), however, these are outside the scope of this paper.

Our work is similar in spirit to a growing group of works that propose checks and evaluation protocols to better understand the capabilities and limitations of interpretability methods [1–3, 21, 23, 26, 28, 32, 41, 49]. Many of these works examine how sensitive post-hoc saliency maps are to different factors such as input perturbations, model weights, or the output class being explained. On the other hand, we conduct an in-depth study of concept-based interpretability methods. Despite their popularity, little is understood about their interpretability and usefulness to human users, or their sensitivity to auxiliary inputs such as the probe dataset. We seek to fill this gap with our work and assist with future development and use of concept-based interpretability methods. To the best of our knowledge, we are the first to investigate the effect of the probe dataset and concepts used for concept-based explanations. There has been work investigating the effect of explanation complexity on human understanding [30], however, it is limited to decision sets.

We also echo the call for releasing more information when releasing datasets [17], models [12, 33] and interpretability methods [47]. More concretely, we suggest that concept-based interpretability method developers to include results from our proposed analyses in their method release, in addition to filling out the explainability fact sheet proposed by Sokol et al. [47], to aid researchers and practitioners to better understand, use, and build on these methods.

3. Dataset choice: Probe dataset has a profound impact on the explanations

Concept-based explanations are generated by running a trained model on a "probe" dataset (typically not the training dataset) which has concepts labelled within it. The choice of probe dataset has been almost entirely dictated by which datasets have concept labels. The most commonly used dataset is the Broden dataset [5]. It contains images from four datasets (ADE20k [57, 58], Pascal [13], OpenSurfaces [6], Describable Textures Dataset [11]) and labels of over 1190 concepts, comprising of object, object parts, color, scene and texture.

In this section, we investigate the effect of the probe dataset by comparing explanations generated using two different subsets of the Broden datset: ADE20k and Pascal. We experiment with three different methods for generating concept-based explanations: *Baseline*, *NetDissect* [5], and *TCAV* [25], and find that the generated explanations heavily depend on the choice of probe dataset. This finding implies that these explanations can only be used for images drawn from the same distribution as the probe dataset.

Model explained. Following prior work [5, 25, 56], we explain a ResNet18-based [20] scene classification model trained on the Places365 dataset [55], which predicts one of 365 scene classes given an input image.

Probe datasets. We use two probe datasets: ADE20k [57, 58] (19733 images, license: BSD 3-Clause) and Pascal [13] (10103 images, license: unknown).[2] They are two different subsets of the Broden dataset [5] and are labelled with objects and parts. We randomly split each dataset into training (60%), validation (20%), and test (20%) sets, using the new training set for learning explanations, validation set for tuning hyperparameters (e.g., learning rate and regularization parameters), and test set for reporting our findings.

Interpretability methods. We investigate the effect of the probe dataset on three types of concept-based explanations. First, we study a simple *Baseline* method that measures correlations between the model's prediction and concepts, and generates class-level explanations as a linear combination of concepts as in Fig. 1. Similar to Ramaswamy et al. [39], we learn a logistic regression model that matches the model-being-explained's prediction, given access to ground-truth concept labels within the image. We use an `l1` penalty to prioritize explanations with fewer concepts. Second, we study *NetDissect* [5] which identifies neurons within the model-being-explained that are highly activated by certain concepts and generates neuron-level explanations (concept labels).[3] Finally, we study *TCAV* [25] which generates explanations in the form of concept activation vectors, i.e., vectors within the model-being-explained's feature space that correspond to labelled concepts.

Results. For all three explanation types, we find that using different probe datasets result in very different explanations. To begin, we show in Tab. 1 how *Baseline* explanations differ when using ADE20k vs. Pascal as the probe dataset. For example, when explaining the `corn-field` scene prediction, the Pascal-generated explanation highlights `dog` as important, whereas the ADE20k-generated explanation does not. For the `legis-chamber` scene, ADE20k highlights `chair` as important, whereas Pascal does not.

We observe a similar difference for *NetDissect* (see Tab. 2). We label 123 neurons separately using ADE20k and Pascal, and find that 60 of them are given very different concept labels (e.g., neuron 239 is labelled `pool-table`

[2] To our best knowledge, most images used don't include personally identifiable information or offensive content. However, some feature people without their consent and might contain identifiable information.

[3] We use code provided by the authors: `https://github.com/CSAILVision/NetDissect-Lite`.

Scene class	Top concepts from ADE20k-generated explanations	Top concepts from Pascal-generated explanations
arena/hockey	**grandstand**, **goal**, **ice-rink**, **scoreboard**	**plaything**, **road**
auto-showroom	car, **light**, **trade-name**, **floor**, **wall**	car, **stage**, **grandstand**, **baby-buggy**, **ground**
bedroom	bed, **cup**, **tapestry**, **lamp**, **blind**	bed, **frame**, **wood**, **sofa**, **bedclothes**
bow-window	windowpane, **seat**, **cushion**, **wall**, **heater**	windowpane, **tree**, **shelves**, **curtain**, **cup**
conf-room	**swivel-chair**, table, **mic**, **chair**, **document**	**bench**, **napkin**, **plate**, **candle**, table
corn-field	**field**, **plant**, sky, **streetlight**	**tire**, sky, **dog**, **water**, **signboard**
garage/indoor	bicycle, **brush**, **car**, **tank**, **ladder**	bicycle, **vending-mach**, **tire**, **motorbike**, **floor**
hardware-store	shelf, **merchandise**, **pallet**, **videos**, box	**rope**, shelves, box, **bottle**, **pole**
legis-chamber	**seat**, **chair**, **pedestal**, **flag**, **witness-stand**	**mic**, **book**, **paper**
tree-farm	tree, **hedge**, **land**, **path**, **pole**	tree, **tent**, **sheep**, **mountain**, **rock**

Table 1. **Impact of probe dataset on *Baseline* (Sec. 3).** We compare *Baseline* explanations generated using ADE20k vs. Pascal. For 10 randomly selected scene classes, we show concepts with the largest coefficients in each explanation. In **bold** are concepts in one explanation but not the other, e.g., the concept **grandstand** is important for explaining the arena/hockey scene prediction when using ADE20k, but not when using Pascal. These results show that the probe dataset has a huge impact on the explanations.

Neuron	ADE20k label & score		Pascal label & score	
9	plant	0.082	potted-plant	0.194
181	plant	0.068	potted-plant	0.140
318	computer	0.079	tv	0.251
386	autobus	0.067	bus	0.200
435	runway	0.071	airplane	0.189
185	chair	0.077	horse	0.153
239	pool-table	0.069	horse	0.171
257	tent	0.042	bus	0.279
384	washer	0.043	bicycle	0.201
446	pool-table	0.193	tv	0.086

Table 2. **Impact of probe dataset on *NetDissect* [5] (Sec. 3).** We compare NetDissect explanations (concept labels) for 10 neurons of the model-being-explained generated using ADE20k vs. Pascal. We find that while some neurons correspond to the same or similar concepts (top half), others correspond to wildly different concepts (bottom half), highlighting the impact of the probe dataset.

by ADE20k and horse by Pascal).[4] Again, this result highlights the impact of the probe dataset on explanations.

Similarly, *TCAV* concept activation vectors learned using ADE20k vs. Pascal are different, i.e., they have low cosine similarity (see Fig. 2). We compute concept activation vectors for 32 concepts which have a base rate of over 1% in both datasets combined, then calculate the cosine similarity of each concept vector. We also compute the ROC AUC for each concept vector to measure how well the concept vector corresponds to the concept. We find that the similarity is low (0.078 on average), even though the selected concepts were those that can be learned reasonably well (mean ROC AUC for these concepts is over 85%). We suspect that the explanations are radically different due to differences in the probe dataset distribution. For instance, some concepts have very different base rates in the two datasets: dog has a base rate of 12.0% in Pascal but 0.5% in ADE20k; chair has a base rate of 16.7% in ADE20k but 13.5% in Pascal.

Concept	ADE20k AUC	Pascal AUC	Cosine sim
ceiling	96.6	93.0	0.267
box	83.0	80.1	0.086
pole	89.0	79.3	0.059
bag	79.4	75.4	0.006
rock	92.6	82.8	-0.024
mean	92.0	88.1	0.087

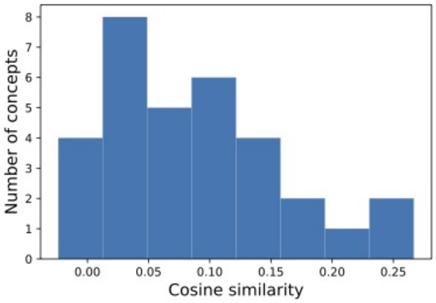


Figure 2. **Impact of probe dataset on *TCAV* [25] (Sec. 3).** We compare TCAV concept activation vectors learned using ADE20k vs. Pascal. *(Top)* For 5 concepts randomly selected out of 32, we show their learnability in each dataset (AUC) and cosine similarity between the two vectors. While these concepts can be learned reasonably well (AUCs are high), their learned activation vectors have low similarity (Cosine sim is low). *(Bottom)* The histogram of cosine similarity scores for all 32 concepts again shows that the two activation vectors for the same concept are not very similar.

We present more analyses in the supp. mat.

4. Concept learnability: Concepts used are less learnable than target classes

In Sec. 3, we investigated how the choice of the probe dataset influences the generated explanations. In this section, we investigate the individual concepts used within explanations. An implicit assumption made in concept-based interpretability methods is that the concepts used in explanations are easier to learn than the target classes being explained. For instance, when explaining the class bedroom with the concept bed, we are assuming (and hoping) that

[4] It is possible that these neurons are poly-semantic, i.e., neurons that reference multiple concepts, as noted in [16, 35]. However, as we explore in the supp. mat., the score for the concept from the other dataset is usually below 0.04, the threshold used in [5] to identify "highly activated neurons."

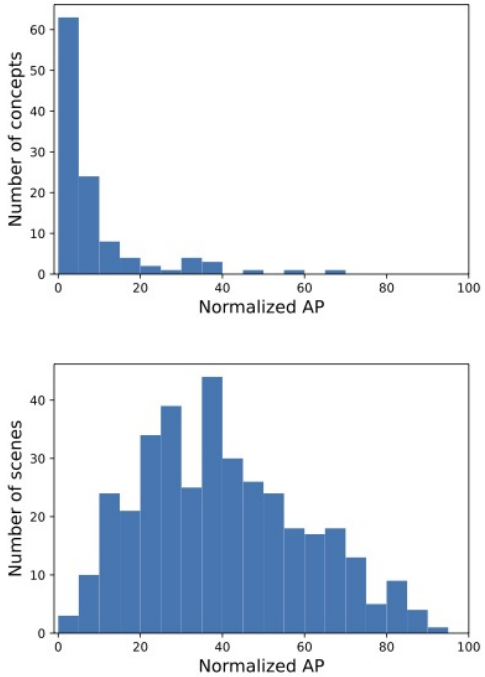


Figure 3. **Overall comparison of concept vs. class learnability (Sec. 4).** We compare the learnability, quantified as normalized AP of concept/class predictors, of Broden concepts (*top*) vs. Places365 scene classes (*below*). Overall, the concepts have much lower normalized AP (i.e., are harder to learn) than the classes.

the model first learns the concept bed, then uses this concept and others to predict the class bedroom. However, if bed is harder to learn than bedroom, this would not be the case. This assumption also aligns with works that argue that “simpler” concepts (i.e., edges and textures) are learned in early layers and “complex” concepts (i.e., parts and objects) are learned in later layers [5, 16].

We thus investigate the learnability of concepts used by different explanation methods. Somewhat surprisingly, we find that the concepts used are frequently harder to learn than the target classes, raising concerns about the correctness of concept-based explanations.

Setup. To compare the learnability of concepts vs. classes, we learn models for the concepts (the learnability of the classes is already known from the model-being-explained). Concretely, we extract features for the probe dataset using an ImageNet [43]-pretrained ResNet18 [20] model and train a linear model using sklearn’s [36] LogisticRegression to predict concepts from the ResNet18 features.[5] We do so for the two most commonly used probe datasets: Broden [5] and CUB-200-2011 [48]. Broden concepts are frequently used to explain Places365 classes (as done in NetDissect [5], Net2Vec [16], IBD [56], and ELUDE [39]), while CUB concepts are used to explain the CUB target classes (as done in Concept Bottleneck [29] and ELUDE [39]).

Evaluation. We evaluate learnability with normalized average precision (AP) [22]. We choose normalized AP for two reasons: first, to avoid having to set a threshold and second, to fairly compare concepts and scenes that have very different base rates. In our experiments, we set the base rate to be that of the classes: $\frac{1}{365}$ when comparing Broden concepts vs. Places365 classes and $\frac{1}{200}$ when comparing CUB concepts vs. CUB classes.

Results. In both settings, we find that the concepts are much harder to learn than the target classes. The median normalized AP for Broden concepts is 7.6%, much lower than 37.5% of Places365 classes. Similarly, the median normalized AP for CUB concepts is 2.3%, much lower than 65.9% of CUB classes. Histograms of normalized APs are shown in Fig. 3 (Broden/Places365) and the supp. mat. (CUB).

However, is it possible that each class is explained by concepts that are more learnable than the class? Our investigation with IBD [56] explanations suggests this is not the case. IBD greedily learns a basis of concept vectors, as well as a residual vector, and decomposes each model prediction into a linear combination of the basis and residual vectors.[6] For 10 randomly chosen scene classes, we compare the normalized AP of the scene class vs. 5 concepts with the highest coefficients (i.e., 5 concepts that are the most important for explaining the prediction). See Tab. 3 for the results. We find that all 10 scene classes are explained with at least one concept that is harder to learn than the class. For some classes (e.g., bathroom, kitchen), all concepts used in the explanation are harder to learn than the class.

Our experiments show that a significant fraction of the concepts used by existing concept-based interpretability methods are harder to learn than the target classes, issuing a wake-up call to the field. In the following section, we show that these concepts can also be hard for people to identify.

5. Human capability: Human studies reveal an upper bound of 32 concepts

Existing concept-based explanations use a large number of concepts: NetDissect [5] and Net2Vec [16] use all 1197 concepts labelled within the Broden [5] dataset; IBD [56] uses Broden object and art concepts with at least 10 examples (660 concepts); and Concept Bottleneck [29] uses all concepts that are predominantly present for at least 10 classes from CUB [48] (112 concepts). However, can people actually reason with these many concepts?

In this section, we study this important yet overlooked aspect of concept-based explanations: *explanation com-*

[5] We also tried using features from a Places365 pretrained model and did not find a significant difference.

[6] We use code provided by the authors: https://github.com/CSAILVision/IBD.

Scene class	Concepts				
arena/perform 38.8	**tennis court** **74.0**	**grandstand** **44.4**	**ice rink** **40.7**	*valley* *19.0*	*stage* *11.9*
art-gallery 27.4	**binder** **42.6**	*drawing* *10.8*	*painting* *10.5*	*frame* *2.5*	*sculpture* *0.7*
bathroom 43.3	*toilet* *39.9*	*shower* *18.8*	*countertop* *12.6*	*bathtub* *11.1*	*screen door* *9.6*
kasbah 50.2	**ruins** **64.3**	*desert* *17.3*	*arch* *16.2*	*dirt track* *8.9*	*bottle rack* *4.2*
kitchen 33.9	*work surface* *24.8*	*stove* *18.2*	*cabinet* *10.3*	*refrigerator* *8.8*	*doorframe* *2.8*
lock-chamber 36.5	**water wheel** **47.4**	**dam** **43.7**	*boat* *16.1*	*embankment* *4.8*	*footbridge* *4.1*
pasture 19.2	**cow** **63.7**	**leaf** **21.1**	*valley* *19.0*	*field* *6.8*	*slope* *4.1*
physics-lab 17.1	**computer** **25.4**	*machine* *4.5*	*monitor-device* *3.3*	*bicycle* *1.7*	*sewing-machine* *1.5*
store/indoor 20.4	**shanties** **72.5**	*patty* *18.5*	*bookcase* *13.5*	*shelf* *4.2*	*cup* *1.3*
water-park 38.3	**roller coaster** **73.0**	**hot tub** **59.1**	**playground** **44.9**	*ride* *38.0*	*swimming pool* *36.7*

Table 3. **Class-level comparison of concept vs. class learnability (Sec. 4).** We report normalized AP scores (↑ indicates high learnability) for 10 randomly chosen scene classes, along with 5 concepts with the highest IBD explanation coefficients for each. Concepts whose normalized AP scores are lower than the scene class are shown in *red*, whereas concepts with higher scores are shown in **blue**. All scenes are explained by at least one concept with a lower normalized AP. Some scenes are only explained by concepts with lower normalized AP.

plexity and how it relates to human capability and preference. Specifically, we investigate: (1) How well do people recognize concepts in images? (2) How do the (concept recognition) task performance and time change as the number of concepts vary? (3) How well do people predict the model output for a new image using explanations? (4) How do people trade off simplicity and correctness of concept-based explanations? To answer these questions, we design and conduct a human study. We describe the study design in Sec. 5.1 and report findings in Sec. 5.2.

5.1. Human study design

We build on the study design and user interface (UI) of HIVE [26], and design a two-part study to understand how understandable and useful concept-based explanations are to human users with potentially limited knowledge about machine learning . To the best of our knowledge, we are the first to investigate such properties of concept-based explanations for computer vision models.[7]

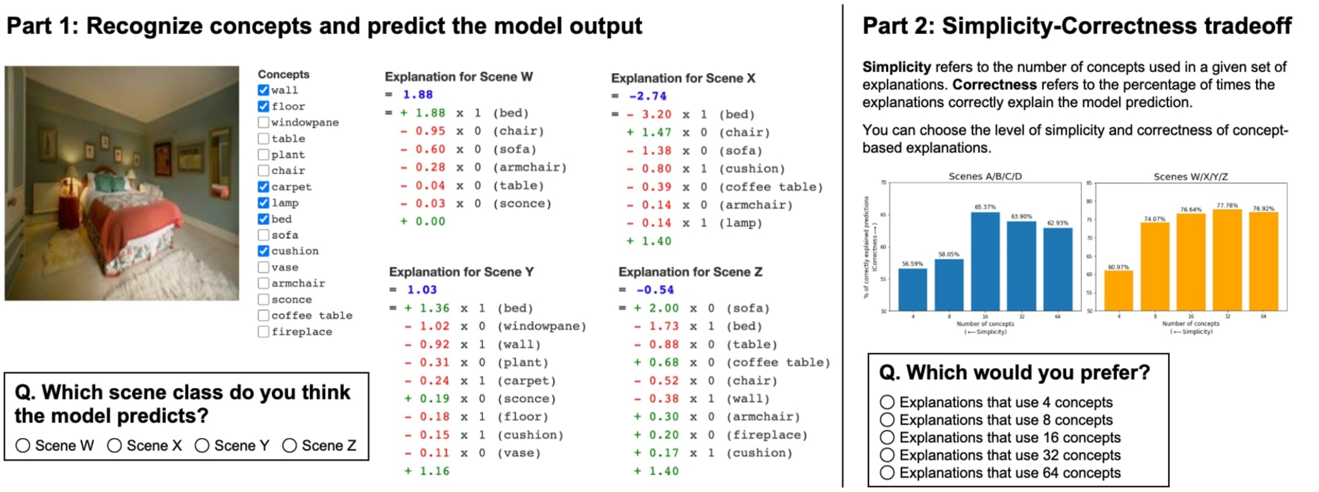
Part 1: Recognize concepts and predict the model output. First, we present participants with an image and a set of concepts and ask them to identify whether each concept is present or absent in the image. We also show explanations for 4 classes whose scores are calculated real-time based on the concepts selected. As a final question, we ask participants to select the class they think the model predicts for the given image. See Fig. 4 (*left*) for the study UI.

To ensure that the task is doable and is only affected by explanation complexity (number of concepts used) and not the complexity of the model and its original prediction task (e.g., 365 scenes classification), we generate explanations for only 4 classes and ask participants to identify which of the 4 classes corresponds to the model's prediction. We only show images where the model output matches the explanation output (i.e., the model predicts the class with the highest explanation score, calculated with ground-truth concept labels), since our goal is to understand how people reason with concept-based explanations with varying complexity.

Part 2: Choose the ideal tradeoff between simplicity and correctness. Next, we ask participants to reason about two properties of concept-based explanations: *simplicity*, i.e., the number of concepts used in a given set of explanations, and *correctness*, i.e., the percentage of model predictions correctly explained by explanations, which is the percentage of times the model output class has the highest explanation score. See Fig. 4 *(right)* for the study UI. We convey the notion of a simplicity-correctness tradeoff through bar plots that show the correctness of explanations of varying simplicity/complexity (4, 8, 16, 32, 64 concepts). We then ask participants to choose the explanation they prefer the most and provide a short justification for their choice.

Full study design and experimental details. In summary,

[7] We note that there are works examining complexity of explanations for other types of models, for example, Lage et al. [30] investigate complexity of explanations over decision sets, Bolubasi et al. [7] investigate this for concept-based explanations for language models.



Part 1: Recognize concepts and predict the model output

Concepts
☑ wall
☑ floor
□ windowpane
□ table
□ plant
□ chair
☑ carpet
☑ lamp
☑ bed
□ sofa
☑ cushion
□ vase
□ armchair
□ sconce
□ coffee table
□ fireplace

Explanation for Scene W
```
=  1.88
= + 1.88  x  1  (bed)
  - 0.95  x  0  (chair)
  - 0.60  x  0  (sofa)
  - 0.28  x  0  (armchair)
  - 0.04  x  0  (table)
  - 0.03  x  0  (sconce)
  + 0.00
```

Explanation for Scene X
```
=  -2.74
= - 3.20  x  1  (bed)
  + 1.47  x  0  (chair)
  - 1.38  x  0  (sofa)
  - 0.80  x  1  (cushion)
  - 0.39  x  0  (coffee table)
  - 0.14  x  0  (armchair)
  - 0.14  x  1  (lamp)
  + 1.40
```

Explanation for Scene Y
```
=  1.03
= + 1.36  x  1  (bed)
  - 1.02  x  0  (windowpane)
  - 0.92  x  1  (wall)
  - 0.31  x  0  (plant)
  - 0.24  x  1  (carpet)
  + 0.19  x  0  (sconce)
  - 0.18  x  1  (floor)
  - 0.15  x  1  (cushion)
  - 0.11  x  0  (vase)
  + 1.16
```

Explanation for Scene Z
```
=  -0.54
= + 2.00  x  0  (sofa)
  - 1.73  x  1  (bed)
  - 0.88  x  0  (table)
  + 0.68  x  0  (coffee table)
  - 0.52  x  0  (chair)
  - 0.38  x  1  (wall)
  + 0.30  x  0  (armchair)
  + 0.20  x  0  (fireplace)
  + 0.17  x  1  (cushion)
  + 1.40
```

Q. Which scene class do you think the model predicts?
○ Scene W ○ Scene X ○ Scene Y ○ Scene Z

Part 2: Simplicity-Correctness tradeoff

Simplicity refers to the number of concepts used in a given set of explanations. **Correctness** refers to the percentage of times the explanations correctly explain the model prediction.

You can choose the level of simplicity and correctness of concept-based explanations.

Q. Which would you prefer?
○ Explanations that use 4 concepts
○ Explanations that use 8 concepts
○ Explanations that use 16 concepts
○ Explanations that use 32 concepts
○ Explanations that use 64 concepts

Figure 4. **Human study UI (Sec. 5).** We show a simplified version of the UI we developed for our human studies. In Part 1, we ask participants to guess the model's prediction for a given image by recognizing concepts and using the provided explanations. In Part 2, we show participants explanations with different levels of simplicity and correctness, then ask which one they prefer the most.

our study consists of the following steps. For each participant, we introduce the study, receive informed consent for participation in the study, and collect information about their demographic (optional) and machine learning experience. We then introduce concept-based explanations in simple terms, and show a preview of the concept recognition and model output prediction task in Part 1. The participant then completes the task for 10 images. In Part 2, the participant indicates their preference for explanation complexity, given simplicity and correctness information. There are no foreseeable risks in participation in the study, and our study design was approved by our institution's IRB.

Using this study design, we investigate explanations that take the form of a linear combination of concepts (e.g., Baseline, IBD [56], Concept Bottleneck [29]). Explanations are generated using the *Baseline* method, which is a logistic regression model trained to predict the model's output using concepts (see Sec. 3 for details). Note that we are evaluating the form of explanation (linear combination of concepts) rather than a specific explanation method. The choice of the method does not impact the task.

Specifically, we compare four types of explanations: concept-based explanations that use (1) 8 concepts, (2) 16 concepts, (3) 32 concepts, and (4) example-based explanations that consist of 10 example images for which the model predicts a certain class. We include (4) as a method that doesn't use concepts. In Jeyakumar et al. [24], this type of explanation is shown to be preferred over saliency-type explanations for image classification; here, we compare this to concept-based explanations.

For a fair comparison, all four are evaluated on the same set of images. In short, we conduct a between-group study with 125 participants recruited through Amazon Mechanical Turk. Participants were compensated based on the state-level minimum wage of $12/hr. In total, ∼$800 was spent on running human studies. See supp. mat. for more details.

5.2. Key findings from the human studies

When presented with more concepts, participants spend more time but are worse at recognizing concepts. The median time participants spend on each image is 17.4 sec. for 8 concept-, 27.5 sec. for 16 concept-, and 46.2 sec. for 32 concept-explanations. This is expected, since participants are asked to make a judgment for each and every concept. When given example-based explanations with no such task, participants spend only 11.6 seconds on each image. Interestingly, the concept recognition performance, reported in terms of mean recall (i.e., the percentage of concepts in the image that are recognized) and standard deviation, decreases from 71.7% ± 27.7% (8 concepts) to 61.0% ± 28.5% (16 concepts) to 56.8% ± 24.9% (32 concepts). While these numbers are far from perfect recall (100%), participants are better at judging whether concepts are present when shown fewer number of concepts.

Concept-based explanations offer little to no advantage in model output prediction over example-based explanations. Indeed, we see that the participants' errors in concept recognition result in an incorrect class having the highest explanation score. When predicting the model output as the class with the highest explanation score, calculated based on the participants' concept selections, the mean accuracy and standard deviation of model output prediction are 64.8% ± 23.9% (8 concepts), 63.2% ± 26.9% (16 concepts), 63.6% ± 22.2% (32 concepts). These are barely higher than 60.0% ± 30.2% of example-based explanations, which are simpler and require less time to complete the task.

The majority of participants prefer explanations with 8, 16, or 32 concepts. When given options of explanations that use 4, 8, 16, 32, or 64 concepts, 82% of participants prefer explanations with 8, 16, or 32 concepts (28%, 33%, 21% respectively). Only 6% prefer those with 64 concepts, suggesting that existing explanations that use hundreds or even thousands of concepts do not cater to human preferences. In the written responses, many favored having fewer concepts (e.g., "the lesser, the better") and expressed concerns against having too many (e.g., "I think 32 is a lot, but 16 is an adequate enough number that it could still predict well..."). In making the tradeoff, some valued correctness above all else (e.g., "Out of all the options, 32 is the most correct"), while others reasoned about marginal benefits (e.g., "I would prefer explanations that use 16 concepts because it seems that the difference in percentage of correctness is much closer and less, than other levels of concepts"). Overall, we find that participants actively reason about both simplicity and correctness of explanations.

6. Discussion

Our analyses yield immediate suggestions for improving the quality and usability of concept-based explanations. First, we suggest choosing a probe dataset whose distribution is similar to that of the dataset the model was trained on. Second, we suggest only using concepts that are more learnable than the target classes. Third, we suggest limiting the number of concepts used within an explanation to under 32, so that explanations are not overwhelming to people.

The final suggestion is easy to implement. However, the first two are easier said than done, since the number of available probe datasets (i.e., large-scale datasets with concept labels) is minimal, forcing researchers to use the Broden dataset [5] or the CUB dataset [48]. Hence, we argue creating diverse and high-quality probe datasets is of upmost importance in researching concept-based explanations.

Another concern is that these methods do highlight hard-to-learn concepts when given access to them, suggesting that they sometimes learn correlations rather than causations. Methods by Goyal et al. [19], which output patches within the image that need to be changed for the model's prediction to change, or Fong et al. [15], which find regions within the image that maximally contribute to the model's prediction, are more in line with capturing causal relationships. However, these only produce *local* explanations, i.e., explanations of a single model prediction, and not class-level *global* explanations. One approach to capturing causal relationships is to generate counterfactual images with or without certain concepts using generative models [40] and observe changes in model predictions.

7. Limitations and future work

Our findings come with a few caveats. First, due to the lack of available probe datasets, we tested each concept-based interpretability method in a single setting. That is, we tested NetDissect [5], TCAV [25] and IBD [56] on a scene classifier trained on the Places365 dataset [55], and Concept Bottleneck [29] on the CUB dataset [48]. We plan to expand our analyses as more probe datasets become available. Second, all participants in our human studies were recruited from Amazon Mechanical Turk. This means that our participants represent a population with limited ML background: the self-reported ML experience was 2.5 ± 1.0 (on a scale of 1 to 5), which is between "2: have heard about..." and "3: know the basics..." We believe Part 1 results of our human studies (described in Sec. 5.1) will not vary with participants' ML expertise or role in the ML pipeline, as we are only asking participants to identify concepts in images. However, Part 2 results may vary (e.g., developers debugging a ML model may be more willing to trade off explanation simplicity for correctness than lay end-users). Investigating differences in perceptions and uses of concept-based explanations, is an important direction for future research.

8. Conclusion

In this work, we examined implicit assumptions made in concept-based interpretability methods along three axes: the choice of the probe datasets, the learnability of the used concepts, and the complexity of explanations. We found that the choice of the probe dataset profoundly influences the generated explanations, implying that these explanations can only be used for images from the probe dataset distribution. We also found that a significant fraction of the concepts used within explanations are harder for a model to learn than the target classes they aim to explain. Finally, we found that people struggle to identify concepts in images when given too many concepts, and that explanations with less than 32 concepts are preferred. We hope our proposed analyses and findings lead to more careful use and development of concept-based explanations.

Acknowledgements. We foremost thank our participants for taking the time to participate in our study. We also thank the authors of [5, 26, 29, 56] for open-sourcing their code and/or models. Finally, we thank the anonymous reviewers and the Princeton Visual AI Lab members (especially Nicole Meister) who provided helpful feedback on our work. This material is based upon work partially supported by the National Science Foundation under Grants No. 1763642, 2145198 and 2112562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We also acknowledge support from the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award to OR, Princeton SEAS Project X Fund to RF and OR, Open Philanthropy Grant to RF, and NSF Graduate Research Fellowship to SK.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018. 1, 3

[2] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *ICLR*, 2022. 3

[3] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *NeurIPS*, 2020. 3

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020. 1

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 1, 2, 3, 4, 5, 8

[6] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. Graph.*, 2014. 3

[7] Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert, 2021. 3, 6

[8] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2018. 2

[9] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018. 2

[10] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, 2019. 2

[11] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 3

[12] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. In *FAccT*, 2022. 3

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 2010. 2, 3

[14] Ruth Fong. *Understanding convolutional neural networks*. PhD thesis, University of Oxford, 2020. 1

[15] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019. 2, 8

[16] Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *CVPR*, 2018. 1, 2, 3, 4, 5

[17] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021. 3

[18] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *DSAA*, 2018. 1

[19] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *ICML*, 2019. 8

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[21] Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. This looks like that... does it? Shortcomings of latent space prototype interpretability in deep networks. In *ICML Workshops*, 2021. 3

[22] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 5

[23] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, 2019. 3

[24] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. In *NeurIPS*, 2020. 7

[25] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018. 1, 2, 3, 4, 8

[26] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In *ECCV*, 2022. 3, 6, 8

[27] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "Help me help the AI": Understanding how explainability can support human-AI interaction. In *CHI*, 2023. 1

[28] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Cham, 2019. 1, 3

[29] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020. 1, 2, 3, 5, 7, 8

[30] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *HCOMP*, 2019. 3, 6

[31] Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *ECCV*, 2016. 1

[32] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? In *ICLR Workshops*, 2021. 3

[33] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAccT*, 2019. 3

[34] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *CVPR*, 2021. 2

[35] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. 4

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 2011. 5

[37] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMCV*, 2018. 2

[38] Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, and David Forsyth. Why do these match? explaining the behavior of image similarity models. In *ECCV*, 2020. 3

[39] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, and Olga Russakovsky. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features. *arXiv:2206.07690*, 2022. 1, 2, 3, 5

[40] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[41] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *CVPR*, 2020. 1, 3

[42] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. In *Statistics Surveys*, 2021. 1

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In *IJCV*, 2015. 5

[44] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer, 2019. 1

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 2

[46] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshops*, 2014. 2

[47] Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *FAccT*, 2020. 1, 3

[48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 5, 8

[49] Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. *arXiv:1907.09701*, 2019. 3

[50] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *NeurIPS*, 2020. 3

[51] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2

[52] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 2018. 2

[53] Quanshi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, 2018. 1

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2

[55] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 3, 8

[56] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *ECCV*, 2018. 1, 2, 3, 5, 7, 8

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20k dataset. In *CVPR*, 2017. 2, 3

[58] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20k dataset. *IJCV*, 2019. 2, 3