# Estimation of Concept Explanations Should Be Uncertainty Aware

Elizabeth Chou and Amanda Boyd



October 1, 2023

# Estimation of Concept Explanations Should be Uncertainty Aware

**Elizabeth J. Chou & Aiden Boyd**


## Abstract

With increasing use of deep learning models, understanding and diagnosing their predictions is becoming increasingly important. A common approach for understanding predictions of deep nets is Concept Explanations. Concept explanations are a form of global model that aim to interpet a deep networks output using human-understandable concepts. However, prevailing concept explanations methods are not robust to concepts or datasets chosen for explanation computation. We show that this sensitivity is partly due to ignoring the effect of input noise and epistemic uncertainty in the estimation process. To address this challenge, we propose an uncertainty-aware estimation method. Through a mix of theoretical analysis and empirical evaluation, we demonstrate the stability, label efficiency, and faithfulness of the explanations computed by our approach.


## 1 Introduction

In the era of ever-larger and more powerful deep neural networks, the need for interpretability and customizability of complex deep nets has never been higher. One compelling solution to this demand is the emergence of interpretable models known as concept-based explanations. These systems attempt to explain a model's predictions by employing high-level and human-understandable concepts, a methodology championed in notable works like (Kim et al., 2018). What makes concept-based explanations particularly appealing is their alignment with semantically relevant patterns (Yeh et al., 2022). Research (Kim et al., 2018, 2023b) substantiates the preference for concept explanations over explanations derived from salient input features (Ribeiro et al., 2016; Selvaraju et al., 2017) or prominent training examples (Koh & Liang, 2017). The significance of concept explanations extends beyond interpretability alone. They also hold the potential to encode domain-specific prior knowledge effectively (Yuksekgonul et al., 2022).

This paper centers on a category of interpretable methods known as concept bottleneck models (CBM)(Koh et al., 2020). Concept explanations explain a pretrained prediction model by estimating the importance of concepts using two human-provided resources: (1) a list of potentially relevant concepts for the task, (2) a dataset of examples usually referred to as the probe-dataset. CBMs, as illustrated in Figure1, operate in a two-step process when presented with a set of concepts and a dataset of examples. In the first step, CBMs compute a score for each concept per example, reflecting the likelihood that a given example embodies a specific concept. These instance-specific concept scores are then collectively aggregated in the second step by constructing a linear model that predicts labels based on concept activations. The resulting linear model weights obtained in this second step constitute a global explanation, shedding light on which concepts bear relevance to the model predictions at hand. One remarkable feature of CBMs is their malleability. These models can be customized and tailored to achieve specific behaviors by manipulating the weights of the interpretable linear model (Yuksekgonul et al., 2022; Oikarinen et al., 2023; Yeh et al., 2022; Choi et al., 2023; Chauhan et al., 2023; Wang et al., 2023). This adaptability not only enhances model interpretability but also empowers users to fine-tune the model's decision-making process in alignment with their specific needs and preferences.

A notable drawback of traditional CBMs lies in their sensitivity to the choice of concept set and dataset (Ramaswamy et al., 2022a). Another major limitation is the need for datasets meticulously annotated with concepts. This process proves prohibitively expensive, particularly when the number of concepts runs into the thousands. However, recent advancements have significantly bolstered the data efficiency of CBMs (Oikarinen et al., 2023; Yuksekgonul et al., 2022; Moayeri et al., 2023) by harnessing pretrained multimodal models like CLIP (Radford et al., 2021) in the initial step to compute activations. While these models have been shown to be useful for common image applications, such multimodal models are not yet thoroughly evaluated for generating post-hoc concept explanations.

Our objective is to generate reliable concept explanations without requiring concept annotations. We observed that per-example concept activations, which are aggregated into a global explanation, can be noisy for irrelevant or hard-to-predict concepts. Since estimation methods do not model noise in concept activations, it cascades into the estimated concept explanation. As a further motivation for modeling uncertainty, imagine the following two scenarios, Section 4.1 presents more concrete scenarios leading to unreliable explanations. (1) When a concept is missing from the dataset, we cannot estimate its importance with confidence. Reporting uncertainty over estimated importance of a concept can thus help the user make a more informed interpretation. (2) The concept activations cannot be accurately estimated for irrelevant or hard concepts, which must be modeled using error intervals on the concept activations. Appreciating the need to model uncertainty, we present an estimator called Uncertainity-Aware Concept Explanations (U-ACE), which we show is instrumental in improving reliability of explanations.

## 2 Background and Motivation

We denote the model-to-be explained as $f : \mathbb{R}^D \to \mathbb{R}^L$ that maps D-dimensional inputs to L labels. Further, we use $f^{[l]}(\mathbf{x})$ to denote $l^{th}$ layer representation space. Given a probe-dataset of examples: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ and a list of concepts $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$, our objective is to explain the pretrained model $f$ using the specified concepts. The concepts are demonstrated using potentially small and independent datasets with concept annotations $\{\mathcal{D}_c^k : k \in [1, K]\}$ where $\mathcal{D}_c^k$ is a dataset with positive and negative examples of the $k^{th}$ concept.

Concept-Based Explanations (CBE) estimate explanations in two steps. In the first step, they learn concept activation vectors that predict the concept from $l^{th}$ layer representation of an example. More formally, we learn the concept activation vector $v_k$ for $k^{th}$ concept by optimizing $v_k = \arg\max_v \mathbb{E}_{(x,y)\sim\mathcal{D}_k^{(k)}}[\ell(v^T f^{[l]}(\mathbf{x}), y)]$ where $\ell$ is the usual cross-entropy loss. The inner product of representation with the concept activation vector: $v_k^T f^{[l]}(\mathbf{x})$ is what we refer to as concept activations. Various approaches exist on how the concept activations are used to compute global explanations for the second step. Kim et al. [14] computes sensitivity of logits to interventions on concept activations to compute what is known as TCAV score per example per concept and reports fraction of examples in the probe-dataset with a positive TCAV score. Zhou et al. [35] proposed to decompose the classification layer weights with $[v_1, v_2, \dots, v_k]$ and use coefficients as the importance score. We refer the reader to Yeh et al. [31] for an in-depth survey.

**Data-efficient concept explanations.** A major limitation of CBEs is their need for datasets with concept annotations: $\{\mathcal{D}_c^1, \mathcal{D}_c^2, \dots\}$. In practical applications, we may wish to find important concepts among thousands of potentially relevant concepts, which is not possible without expensive data collection. Recent proposals [32, 21, 20] suggested using pretrained multimodal models like CLIP to evade the data annotation cost for a related problem called Concept Bottleneck Models (CBM) [18]. CBMs aim to train inherently interpretable model with concept bottleneck. Although CBMs cannot generate explanations for a model-to-be-explained, a class of algorithms propose to train what are known as Posthoc-CBMs using the representation layer of a pretrained task model for data efficiency. Given that Posthoc-CBMs base on the representation of a pretrained task model, we may use them to generate concept explanations. We describe briefly two such CBM proposals below.

Oikarinen et al. [21] (O-CBM) estimates the concept activation vectors by learning to linearly project from the embedding space of CLIP where the concept is encoded using its text description to the embedding space of the model-to-be-explained: $f$. It then learns a linear classification model on concept activations and returns the weight matrix as the concept importance score. Based on the proposal of Yuksekgonul et al. [32], we can also generate explanations by training a linear model to match the predictions of model-to-be-explained using the concept activations of CLIP, which we denote by (Y-CBM).

**Limitation: Unreliable Explanations.** We noted critical reliability concerns with existing CBEs in the same spirit as the challenges raised in Ramaswamy et al. [23]. As we demonstrate in Section 4.1, concept explanations for the same model-to-be-explained vary with the choice of probe-dataset and the concept set bringing into question the reliability of explanations.

## 3 Uncertainity-Aware Concept Explanations

As summarized in the previous section, CBEs rely on concept activations for generating explanations. It is not hard to see that the activation score of a concept cannot be predicted confidently if the concept is hard or if it is not used by the model-to-be-explained. The noise in concept activations if not modeled cascades into the next step leading to high variance or poor explanations. Moreover, importance of a concept cannot be confidently estimated if it is missing from the dataset, which must be informed to the user through confidence interval on the concept's estimated

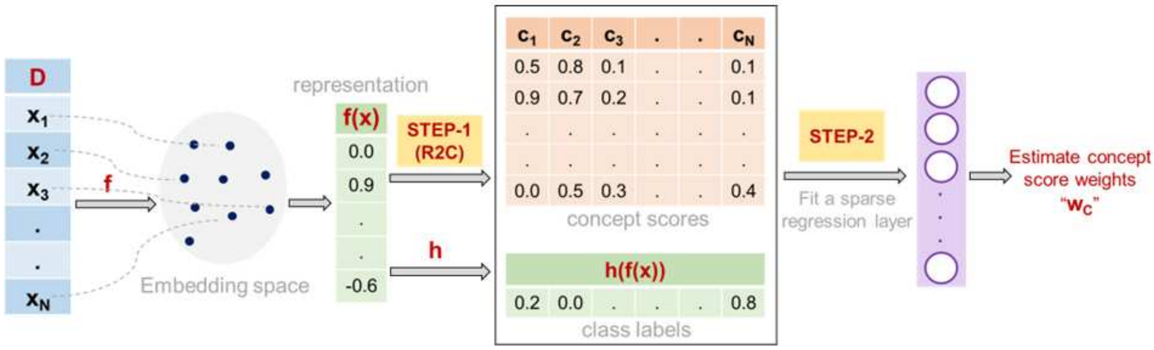


Figure 1: Overview of estimation of Concept Explanations. Existing estimation methods do not account for (1) noise in concept score estimation, (2) model/knowledge uncertainty due to lack of spread in the dataset.

importance score. Motivated by the role of uncertainty in estimation and for explanations, we design our estimator described below.

Our approach has the following steps. (1) Estimate concept activations along with their error interval, (2) Compute and return a linear predictor model that is robust to input noise. We describe the estimation of concept activations and their error given an instance $\mathbf{x}$ denoted as $\vec{m}(\mathbf{x}), \vec{s}(\mathbf{x})$ respectively in Section 3.1. Once concept activations are computed, we proceed with the linear estimator as follows.

Our objective is to learn linear model weights $W_c$ of size $L \times K$ (recall that K is number of concepts and L the number of labels) that map the concept activations to their logit scores, i.e. $f(\mathbf{x}) \approx W_c\vec{m}(\mathbf{x})$. Since the concept activations contain noise, we require that $W_c$ is such that predictions do not change under noise, that is $W_c[\vec{m}(\mathbf{x}) + \vec{s}(\mathbf{x})] \approx W_c\vec{m}(\mathbf{x}) \implies W_c\vec{s}(\mathbf{x}) \approx 0$. I.e. the inner product of each row ($\vec{w}$) of $W_c$ with $\vec{s}(\mathbf{x})$ must be negligible. The constraint translates to a neat distributional prior over weights when we approximate the heteroskedastic input noise with its average: $\epsilon = \frac{\sum_{x\in\mathcal{D}} s(\vec{\mathbf{x}})}{N}$, which is shown below.

$$\begin{aligned} &|\vec{w}^T\epsilon| \leq \delta, \text{ for some small } \delta > 0 \text{ with high probability} \\ &\implies \vec{w}^T\text{diag}(\epsilon\epsilon^T)\vec{w} \leq \delta^2 \implies \vec{w} \sim \mathcal{N}(0, \lambda\text{diag}(\epsilon\epsilon^T)), \lambda > 0 \end{aligned}$$

We observe therefore that the weight vectors drawn from $\mathcal{N}(0, \lambda\text{diag}(\epsilon\epsilon^T))$ satisfy the invariance to input noise constraint with high probability (w.h.p.) for a sufficiently large $\lambda$. We now estimate the posterior on the weights after having observed the data with the prior on weights set to $\mathcal{N}(0, \lambda\text{diag}(\epsilon\epsilon^T))$. The posterior over weights has the following closed form[26] where $C_X = [\vec{m}(\mathbf{x}_1), \vec{m}(\mathbf{x}_2), \ldots, \vec{m}(\mathbf{x}_N)]$ and $Y = [f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)]^T$.

$$\vec{w} \sim \mathcal{N}(\mu, \Sigma) \qquad \text{where } \mu = \Sigma^{-1}C_XY, \quad \Sigma^{-1} = \beta C_X C_X^T + (\lambda\text{diag}(\epsilon\epsilon^T))^{-1} \tag{1}$$

$\beta$ is the inverse variance of noise in observations. We optimise $\beta$ and $\lambda$ using MLE on $\mathcal{D}$ (Appendix A).

**Sparsifying weights for interpretability.** Because a dense weight matrix can be hard to interpret, we induce sparsity in $W_c$ by setting all the values below a threshold to zero. The threshold is picked such that the accuracy on train split does not fall by more than $\kappa$, which is a positive hyperparameter.

The estimator shown in Equation 1 and details on how we estimate the noise in concept activations presented in the next section completes the description of our estimator. We call our estimator Uncertainity-Aware Concept Explanations (U-ACE) because it models also the uncertainty in concept activations. Algorithm 1 summarizes our proposed system.

### 3.1 Estimation of concept activations and their noise

Pretrained image-text multimodal systems can embed both images and text in a shared representation space, which enables one to estimate the similarity of an image to a sentence. This presents us an interesting solution approach of specifying a concept using its text description ($T_k$ for the $k^{th}$ concept) thereby avoiding the need for concept datasets. We denote by $g(\mathbf{x})$ the image embedding of $\mathbf{x}$ by CLIP and $g_{text}(T_k)$ the text embedding. We may compute a concept activation score of an instance $\mathbf{x}$ for a concept k by simply computing the inner product of CLIP embeddings $g(\mathbf{x})^T g_{text}(T_k)$. We require, however, to estimate concept activations using the model-to-be-explained. We can do so if we can find a vector in the embedding space of $f$ corresponding to $g_{text}(T_k)$. We turn to the method proposed in Oikarinen et al. [21] to register representation spaces. Their procedure is summarised below, where we wish to optimise for a weight vector $v_k$ in the representation space of $f$ corresponding to $w_k = g_{text}(T_k)$ in $g$.

Embed $v$ in the representation space of $f$: $e(v, f, \mathcal{D}) = [v^T f(\mathbf{x}_1), v^T f(\mathbf{x}_2), \dots, v^T f(\mathbf{x}_N)]^T$
Embed $w_k = g_{text}(c_k)$ in the representation space of $g$: $e(w_k, g, \mathcal{D}) = [w_k^T g(\mathbf{x}_1), \dots, w_k^T g(\mathbf{x}_N)]^T$
optimize for $v$ that is closest to $w_k$: $v_k = \arg\max_v[\text{cos-sim}(e(v, f, X), e(w_k, g, \hat{\mathcal{D}}))]$
$cos(\alpha_k) \triangleq$ cos-sim$(e(v_k, f, \mathcal{D}), e(w_k, g, \mathcal{D}))$, which loosely informs how well $v_k$ approximates $w_k$.

We may repeat the estimation procedure and set $\alpha_k$ to sample mean for a better estimate. The mean concept activations and their confidence interval can now be estimated using $cos(\alpha_k)$ as given by the following result, proof in Appendix B.

**Proposition 1.** *For a concept k and $cos(\alpha_k)$ defined as above, we have the following result when concept activations in f for an instance $\mathbf{x}$ are computed as cos-sim$(f(\mathbf{x}), v_k)$ instead of $v_k^T f(\mathbf{x})$.*

$$\vec{m}(\mathbf{x})_k = cos(\theta_k)cos(\alpha_k), \quad \vec{s}(\mathbf{x})_k = sin(\theta_k)sin(\alpha_k)$$

*where $cos(\theta_k)$=cos-sim$(g_{text}(T_k), g(\mathbf{x}))$ and $\vec{m}(\mathbf{x})_k, \vec{s}(\mathbf{x})_k$ denote the $k^{th}$ element of the vector.*

The mean and scale values above have a clean interpretation. If model-to-be-explained $(f)$ uses the $k^{th}$ concept for label prediction, the information about the concept is encoded in $f$ and we get a good fit, i.e. $cos(\alpha_k) \approx 1$, and a small error on concept activations. On the other hand, error bounds are large and concept activations are suppressed when the fit is poor, i.e. $cos(\alpha_k) \approx 0$.

---

Algorithm 1: Uncertainity-Aware Concept Explanations (U-ACE)

---

**Require:** $\mathcal{D}$={$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$}, f (model-to-be-explained), g (CLIP), $\kappa$ (tolerance hparam)
  **for** $y = 1, \dots, L$ **do**
    Y = $[f(\mathbf{x})$ for $\mathbf{x} \in \hat{\mathcal{D}}]^T$ ▷ Gather logits
    $C_X = [\vec{m}(\mathbf{x}_1), \dots, \vec{m}(\mathbf{x}_N)]$, $\epsilon = \mathbb{E}_{\mathcal{D}}[\vec{s}(\mathbf{x})]$ ▷ Estimate $\vec{m}(\mathbf{x}), \vec{s}(\mathbf{x})$ (Section 3.1)
    $\vec{w}_y \sim \mathcal{N}(\mu_y, \Sigma_y)$ where $\mu_y, \Sigma_y$ from Equation 1 ▷ Estimate $\lambda, \beta$ using MLL
  **end for**
  $W_c$ = sparsify$([\vec{\mu}_1, \vec{\mu}_2, \dots \vec{\mu}_L], \kappa)$ ▷ Suppress less useful weights, Section 3
  **return** $W_c$, $[\text{diag}(\Sigma_1), \text{diag}(\Sigma_2), \dots \text{diag}(\Sigma_L)]$

---

### 3.2 Theoretical motivation

The motivation of this section is to demonstrate unreliability of concept explanations estimated using standard methods that do not model uncertainty during estimation. We particularly focus on unreliability due to misspecified concept set for the ease of analysis. In our study, we compared explanations generated using a standard linear estimator and U-ACE. Recall that posthoc-CBMs (O-CBM, Y-CBM), which are our primary focus for comparison, estimate explanations by fitting a linear model on concept activations.

We present two scenarios with noisy concept activations. In the first scenario (over-complete concept set), we analyzed the estimation when the concept set contains many irrelevant concepts. We show that the likelihood of marking an irrelevant concept as more important than a relevant concept increases rapidly with the number of concepts when the explanations are estimated using a standard linear estimator that is ignorant of the noise. We also show that U-ACE do not suffer the same problem. In the second scenario (under-complete concept set), we analyzed the explanations when the concept set only includes irrelevant concepts, which should both be assigned a zero score ideally. We again show that standard linear model attributes a significantly non-zero score while U-ACE mitigates the issue well. In Section 4.1, we confirm our theoretical findings with an empirical evaluation.

**Unreliable explanations due to over-complete concept set**. We analyze a simple setting where the output is linearly predicted from the input $(\mathbf{x})$ as $y = \mathbf{w}^T\mathbf{x}$. We wish to estimate the importance of K concepts fitted using a linear estimator on concept activations. The concept activations are computed using concept activation vectors $(\mathbf{w}_k)$ that are distributed as $\mathbf{w}_k \sim \mathcal{N}(\mathbf{u}_k, \sigma_k^2 I), k \in [1, K]$.

**Proposition 2.** *The concept importance estimated by U-ACE when the input dimension is sufficiently large and for some $\lambda > 0$ is approximately given by $v_k = \frac{\mathbf{u}_k^T\mathbf{w}}{\mathbf{u}_i^T\mathbf{u}_k + \lambda\sigma_k^2}$. On the other hand, the importance scores estimated using vanilla linear estimator under the same conditions is distributed as $v_k \sim \mathcal{N}(\frac{\mathbf{u}_k^T\mathbf{w}}{\mathbf{u}_k^T\mathbf{u}_k}, \sigma_k^2\frac{\|w\|^2}{\|u_k\|^2})$.*

Proof of the result can be found in Appendix C. If we consider a setting where only the first of the K random concepts is relevant and the rest random, i.e. $\mathbf{u}_1 = \mathbf{w}, \sigma_1 \approx 0$ and $\mathbf{u}_k$ such that $\mathbf{u}_k^T\mathbf{w} \approx 0 \quad \forall k \in [2, K]$. In this setting, U-ACE estimated importance scores is 1 for the relevant concept and 0 for the rest, while the importance scores estimated by the vanilla linear regression model are normally distributed with means at 1 for the relevant concept and 0 for the irrelevant concepts. However, due to variance of importance scores estimated by the vanilla model, the probability that at least of the K-1 random concepts is estimated to be more important than the relevant concept is $1-\prod_{k=2}^{K}\Phi(\frac{\|u_k\|}{\sigma_k\|w\|})$, where $\Phi$ is the CDF of standard normal. We observe that the probability of a random concept being estimated as more important than the relevant concept quickly converges to 1 with the number of random concepts: K-1.

**Unreliable explanations due to under-complete concept set**. We now analyze explanations when the concept set only includes two irrelevant concepts. Consider normally distributed inputs: $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$, and define two orthogonal unit vectors: $u, v$. The concept activations: $c_1^{(i)}, c_2^{(i)}$ and label $y^{(i)}$ for the $i^{th}$ instance $\mathbf{x}^{(i)}$ are as defined below.

$$y^{(i)} = u^T\mathbf{x}^{(i)}, \quad c_1^{(1)} = (\beta_1 u + (1-\beta_1)v)^T\mathbf{x}^{(i)}, \quad c_2^{(i)} = (\beta_2 u + (1-\beta_2)v)^T\mathbf{x}^{(i)}$$

If $\beta_1, \beta_2$ are very small, then both the concepts are expected to be unimportant for label prediction. However, we can see with simple working (Appendix D) that the importance scores computed by a standard estimator are $\frac{1-\beta_2}{\beta_1-\beta_2}, \frac{1-\beta_1}{\beta_1-\beta_2}$, which are large because $\beta_1 \approx 0, \beta_2 \approx 0 \therefore \beta_1-\beta_2 \approx 0$. We will now show that U-ACE estimates near-zero importance scores as expected.

**Proposition 3.** *The importance score, denoted $v_1, v_2$, estimated by U-ACE are bounded from above by $\frac{1}{N\lambda}$, i.e. $v_1, v_2 = \mathcal{O}(1/N\lambda)$ where $\lambda > 0$ is a regularizing hyperparameter and N the number of examples.*

Proof can be found in Appendix D. It follows from the result that the importance scores computed by U-ACE are near-zero for sufficiently large value of $\lambda$ or N.

## 4 Experiments

We evaluate U-ACE on two synthetic and two real-world datasets. We demonstrate how reliability of explanations is improved by U-ACE in Section 4.1. We make a quantitative assessment with known ground-truth on a controlled dataset in Section 4.2. Finally, we evaluate on two challenging real-world datasets with more than 700 concepts in Section 4.3.

**Baselines.** *Simple:* $W_c$ is estimated using lasso regression of ground-truth concept annotations to estimate logit values of $f$. This baseline is used in the past [24, 23] for estimating completeness of concepts. Other baselines are introduced in Section 2: *TCAV* [14], *O-CBM* [21], *Y-CBM* based on [32].

**Standardized comparison between importance scores.** The interpretation of the importance score varies between different estimation methods. For instance, the importance scores in TCAV correspond to fraction of examples that meet certain criteria while other methods the importance scores are the weights from linear model that predicts logits. Further, *Simple* operates on binary attributes and *O-CBM* operates on cosine-similarities as the input. For this reason, we cannot directly compare importance scores or their normalized variants. We instead use negative scores to obtain a ranked list of concepts and assign to each concept an importance score given by its rank in the list normalized by number of concepts. Our sorting algorithm ranks any two concepts with same score by alphabetical order of their text description. In all our comparisons we use the rank score if not mentioned otherwise.

**Other experiment details.** For all our experiments, we used a Visual Transformer (with 32 patch size called "ViT-B/32") based pretrained CLIP model that is publicly available for download. We use $l = -1$, i.e. last layer just before computation of logits for all the explanation methods. U-ACE returns the mean and variance of the importance scores as shown in Algorithm 1, we use mean divided by standard deviation as the importance score estimated by U-ACE everywhere for comparison with other methods.

### 4.1 Simulated Study

In this section, we consider explaining a two-layer CNN model trained to classify between solid color images with pixel noise as shown in Figure 2. The colors on the left: red, green are defined as label 0 and the ones on the right are defined as label 1: blue, white. The model-to-be-explained is trained on a dataset with equal proportion of all colors, so we expect that all constituent colors of a label are equally important for the label. We specify a concept set with the four colors encoded by their literal name: *red, green, blue, white*. U-ACE (along with others) attribute positive importance for *red, green* and negative or zero importance for *blue, white* when explaining label

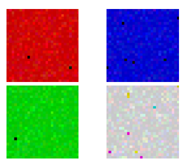

Figure 2: Toy

0 using a concept set with only the four task-relevant concepts and when the probe-dataset is the same distribution as the the training dataset. However, quality of explanations quickly degrade when the probe-dataset is shifted or if the concept set is misspecified.

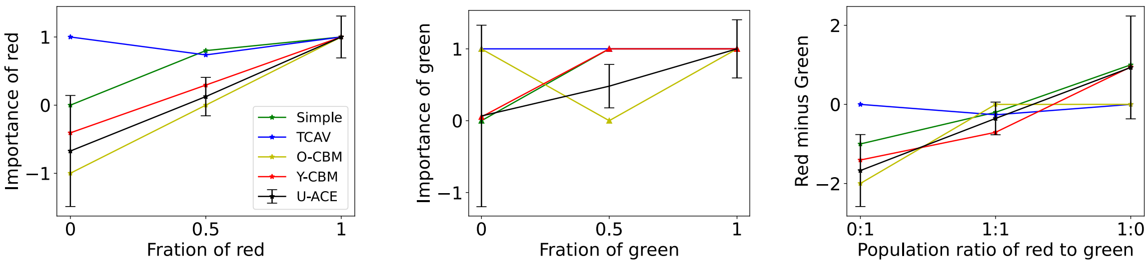


Figure 3: Left, middle plots show the importance of red and green concepts while the rightmost plot shows their importance score difference. U-ACE estimated large uncertainty in importance score when red or green concept is missing from the dataset as seen in the left of the left and middle plots.

**Unreliability due to dataset shift.** We varied the probe-dataset to include varying population of different colors while keeping the concept set and model-to-be-explained fixed. We observed that importance of a concept estimated with standard CBEs varied with the choice of probe-dataset for the same underlying model-to-be-explained as shown in left and middle plots of Figure 3. Most methods attributed incorrect importance to the *red* concept when it is missing (left extreme of left plot), and similarly for the *green* concept (left extreme of middle plot). The explanations have led the user to believe that *green* is more important than *red* or *red* is more important than *green* depending on the probe-dataset used as shown in the right most plot. Because U-ACE also informs the user of uncertainty in the estimated importance, we see that the difference in importance scores between the two colors at either extremes is not statistically significant, also shown in the rightmost plot.

**Unreliability due to misspecified concept set.** We simulate a over-complete concept set scenario analogous to the settings analyzed in Section 3.2 and empirically confirm the merits of U-ACE. Appendix F presents and evaluates on an under-complete concept setting.

**Over-complete concept set**. We gradually expanded the concept set to also include common fruit names as concepts along with the four initial color concepts (Appendix E contains the full list) while using an in-distribution probe-dataset. Figure 4 shows the most salient fruit concept with increasing number of fruit (nuisance) concepts and note that U-ACE is far more robust to the presence of nuisance concepts. Robustness to irrelevant concepts is important because it allows the user to begin with a superfluous set of concepts and find their relevance to model-to-be-explained instead of requiring to guess relevant concepts, which is ironically the very purpose of using concept explanations.

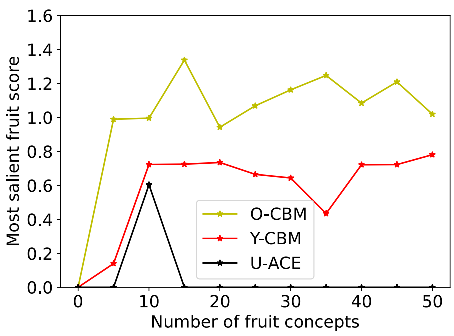


Figure 4

## 4.2 Assessment with known ground-truth

Our objective in this section is to establish that U-ACE generates faithful and reliable concept explanations. Subscribing to the common evaluation practice [14], we generate explanations for a model that is trained on a dataset with controlled correlation of a spurious pattern. We make a dataset using two labels from STL-10 dataset [9]: *car, plane* and paste a tag: *U* or *Z* in the top-left corner as shown in the left panel of Figure 5. The probability that the examples of *car* are added the *Z* tag is p and 1-p for the *U* tag. Similarly for the examples of *plane*, the probability of *U* is p and *Z* is 1-p. We generate three training datasets with p=0, p=0.5 and p=1, and train three classification models using 2-layer convolutional network. Therefore, the three models are expected to have a varying and known correlation with the tag, which we hope to recover from its concept explanation.

We generate concept explanations for the three model-to-be-explained using a concept set that includes seven car-related concepts and three plane-related concepts along with the two tags: *U, Z*. We obtain the importance score of

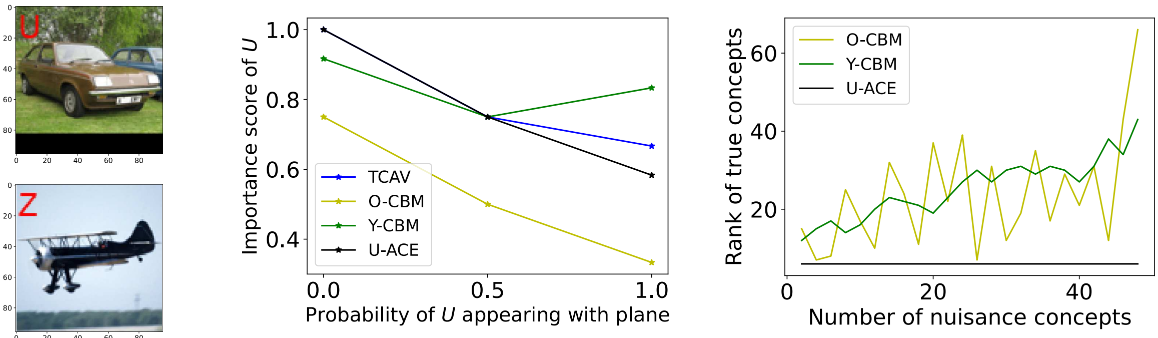


Figure 5: Left: STL dataset with a spurious tag. Middle: Importance of a tag concept for three model-to-be-explained. X-axis shows the probability of tag in the training dataset of model-to-be-explained. Right: Average rank of true concepts with irrelevant concepts (lower is better).

**Tree Farm**
*Simple*: `tree, field, bush`
O-CBM: `forest, pot, sweater`
Y-CBM: `field, forest, elevator`
*U-ACE*: `foliage, forest, grass`

**Pasture**
*Simple*: `horse, sheep, grass`
*O-CBM*: `shaft, hoof, exhibitor`
*Y-CBM*: `field, grass, ear`
*U-ACE*: `grass, cow, banded`

**Coast**
*Simple*: `sea, water, river`
*O-CBM*: `sea, island, pitted`
*Y-CBM*: `sea, sand, towel rack`
*U-ACE*: `sea, lake, island`

**Runway**
*Simple*: `plane, field, sky`
*O-CBM*: `plane, fuselage, apron`
*Y-CBM*: `plane, clouds, candlestick`
U-ACE: `plane, windscreen, sky`

Figure 6: Top-2 salient concepts plus any mistake (marked in red) from top-10 salient concepts for a scene-classification model estimated with PASCAL (left) or ADE20K (right) probe-dataset.

the concept $U$ with *car* class using a probe-dataset that is held-out from the corresponding training dataset (i.e. probe-dataset has the same input distribution as the training dataset). The results are shown in the middle plot of Figure 5. Since the co-occurrence probability of $U$ with *car* class goes from 1, 0.5 to 0, we expect the importance score of $U$ should change from positive to negative as we move right. We note that U-ACE, along with others, show the expected decreasing importance of the tag concept. The result corroborates that U-ACE estimates a faithful explanation of model-to-be-explained while also being more reliable as elaborated below.

**Unreliability due to misspecified concept set.** In the same spirit as the previous section, we repeat the over-complete experiment of Section 4.1 and generated explanations as animal (irrelevent) concepts are added. Right panel of Figure 5 shows the average rank of true concepts (lower the better). We note that U-ACE generates expected explanations even with 50 nuisance concepts.

### 4.3 Real-world evaluation

We expect that our reliable estimator to also generate higher quality concept explanations in practice. To verify the same, we generated explanations for a scene classification model with ResNet-18 architecture pretrained on Places365 [33], which was publicly available. Following the experimental setting of Ramaswamy et al. [23], we generate explanations using PASCAL [7] or ADE20K [34] that are part of the Broden dataset collection [4]. The dataset contains images with dense annotations with more than 1000 attributes. We ignored around 300 attributes describing the scene since model-to-be-explained is itself a scene classifier. For the remaining 730 attributes, we defined a concept per attribute using literal name of the attribute. We picked 50 scene labels (Appendix E contains the full list) that have support of at least 20 in both ADE20K and PASCAL datasets.

We evaluate quality of explanations by their closeness to the explanations generated using the *Simple* baseline. *Simple* estimates explanation using concept annotations and therefore its explanation must be the closest to the ground-truth. For the top-20 concepts identified by *Simple*, we compute the average absolute difference in importance scores estimated using any estimation method and *Simple*. Table 1 presents the deviation in explanations averaged over all the 50 scene labels. Figure 6 shows the most salient concepts for four scene labels. We note that U-ACE generated explanations are more convincing over O-CBM or Y-CBM. We also evaluated the explanation quality using a standard measure for comparing ranked lists, which is presented in Appendix E, and further confirms the dominance of U-ACE.

**Dataset shift.** Ramaswamy et al. [23] demonstrated with results the drastic shift in concept explanations for the same model-to-be-explained when using ADE20K or PASCAL as the probe-dataset. Explanations diverge partly because (a) population of concepts may vary between datasets thereby influencing their perceived importance when using standard methods, (b) variance in explanations. We have demonstrated that U-ACE estimated importance scores have low variance (shown in Section 3.2, 4.1) and attributes high uncertainty and thereby near-zero importance to concepts that are rare or missing from the probe-dataset (Section 4.1). For these reasons, we expect U-ACE to mitigate the data-shift problem. We confirm the same by estimating the average difference in importance scores estimated using ADE20K and PASCAL for different estimation techniques (where the average is only over salient concepts with non-zero importance). The results are shown in Table 2 and are inline with our prediction.

| Dataset↓ | TCAV | O-CBM | Y-CBM | U-ACE |
|---|---|---|---|---|
| ADE20K | 0.13 | 0.19 | 0.16 | **0.09** |
| PASCAL | 0.41 | 0.20 | 0.18 | **0.11** |

Table 1: *Evaluation of explanation quality.* Each cell shows the average absolute difference of importance scores for top-20 concepts estimated using *Simple*.

| Simple | TCAV | O-CBM | Y-CBM | U-ACE |
|---|---|---|---|---|
| 0.41 | 0.41 | 0.32 | 0.33 | **0.19** |

Table 2: *Effect of data shift.* Average absolute difference between concept importance scores estimated using ADE20K and PASCAL datasets for the same model-to-be-explained using different estimation methods.

## 5 Related Work

**Concept Bottleneck Models** use a set of predefined human-interpretable concepts as an intermediate feature representation to make the predictions [18, 3, 14, 35]. CBM allows human test-time intervention which has been shown to improve overall accuracy [2]. Traditionally, they require labelled data with concept annotations and typically the accuracy is worse than the standard models without concept bottleneck. To address the limitation of concept annotation, recent works have leveraged large pretrained multimodal models like CLIP [21, 32]. There have also been efforts to enhance the reliability of CBMs by focusing on the information leakage problem [13, 19], where the linear model weights estimated from concept activations utilize the unintended information, affecting the interpretability. Concept Embedding Models (CEM) [11] overcome the trade-off between accuracy and interpretability by learning high-dimensional concept embeddings. However, addressing the noise in the concept prediction remains underexplored. Collins et al. [10] have studied human uncertainty in concept-based models and have shown the importance of considering uncertainty over concepts in improving the reliability of the model. Kim et al. [15] proposed the Probabilistic Concept Bottleneck Models (ProbCBM) and is closely related to our work. They too argue for the need to model uncertainty in concept prediction for reliable explanations. However, their method of noise estimation in concept activations requires retraining the model and cannot be applied directly when concept activations are estimated using CLIP. Moreover, they use simple MC sampling to account for noise in concept activations.

**Concept based explanations** use a separate probe dataset to first learn the concept and then explain through decomposition either the individual predictions or overall label features. Yeh et al. [31] contains a brief summary of existing concept based explanation methods. Our proposed method is very similar to concept based explanations (CBE) [14, 3, 35, 12]. Ramaswamy et al. [23] emphasized that the concepts learned are sensitive to the probe dataset used and therefore pose problems when transferring to applications that have distribution shift from the probe dataset. Moreover, they also highlight other drawbacks of existing CBE methods in that concepts can sometimes be harder to learn than the label itself (meaning the explanations may not be causal) and that the typical number of concepts used for explanations far exceed what a typical human can parse easily. Achtibat et al. [1] championed an explanation method that provides explanation highlighting important feature (answering “where”) and what concepts are used for prediction thereby combining the strengths of global and local explanation methods. Choi et al. [8] have built upon the current developments in CBE methods for providing explanations for out-of-distribution detectors. Wu et al. [30] introduced the causal concept based explanation method (Causal Proxy Model), that provides explanations for NLP models using counterfactual texts. Moayeri et al. [20] also used CLIP to interpret the representations of a different model trained on uni-modal data.

## References


[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From" where" to" what": Towards human-understandable explanations through concept relevance propagation. *arXiv preprint arXiv:2206.03208*, 2022.

[2] Matthew Barker, Katherine M Collins, Krishnamurthy Dvijotham, Adrian Weller, and Umang Bhatt. Selective concept models: Permitting stakeholder customisation at test-time. *arXiv preprint arXiv:2306.08424*, 2023.

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

[5] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019. URL http://jmlr.org/papers/v20/18-403.html.

[6] Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5948–5955, 2023.

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978, 2014.

[8] Jihye Choi, Jayaram Raghuram, Ryan Feng, Jiefeng Chen, Somesh Jha, and Atul Prakash. Concept-based explanations for out-of-distribution detectors. In *International Conference on Machine Learning*, pp. 5817–5837. PMLR, 2023.

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

[10] Katherine Maeve Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. Human uncertainty in concept-based ai systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 869–889, 2023.

[11] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35: 21400–21413, 2022.

[12] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[13] Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

[14] 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 *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

[15] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*, 2023.

[16] Sunnie SY 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 *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2023.

[17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

[18] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

[19] Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: Interpretable, leak-proof concept-based models. *Advances in Neural Information Processing Systems*, 35:21212–21227, 2022.

[20] Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. *arXiv preprint arXiv:2305.06386*, 2023.

[21] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[23] Vikram V Ramaswamy, Sunnie SY Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. *arXiv preprint arXiv:2207.09615*, 2022.

[24] Vikram V Ramaswamy, Sunnie SY Kim, Nicole Meister, Ruth Fong, and Olga Russakovsky. Elude: Generating interpretable explanations via a decomposition into labelled and unlabelled features. *arXiv preprint arXiv:2206.07690*, 2022.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[26] Russ Salakhutdinov. Statistical machine learning, 2011. URL https://www.utstat.toronto.edu/~rsalakhu/sta4273/notes/Lecture2.pdf#page=15.

[27] 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 *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[28] Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Learning bottleneck concepts in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10962–10971, 2023.

[29] Wikipedia. Kendall tau distance — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Kendall%20tau%20distance&oldid=1163706720, 2023. [Online; accessed 25-September-2023].

[30] Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. Causal proxy models for concept-based model explanations. In *International Conference on Machine Learning*, pp. 37313–37334. PMLR, 2023.

[31] Chih-Kuan Yeh, Been Kim, and Pradeep Ravikumar. Human-centered concept explanations for neural networks. *arXiv preprint arXiv:2202.12451*, 2022.

[32] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

[33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

[35] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018.