Appendix

A Maximum Likelihood Estimation of U-ACE parameters: λ, β

The posterior on weights shown in Equation 1 has two parameters: λ, β as shown below with C_X and Y are array of concept activations and logit scores (see Algorithm 1).

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

We obtain the best values of λ and β that maximize the log-likelihood objective shown below.

$$\lambda^*, \beta^* = \arg\max_{\lambda,\beta} \quad \mathbb{E}_Z[-\frac{\beta^2 \|Y - (C_X + Z)^T \vec{w}(\lambda, \beta)\|^2}{2} + \log(\beta)]$$

where Z is uniformly distributed in the range given by error intervals

$$Z \sim Unif([-\vec{s}(\mathbf{x}_1), -\vec{s}(\mathbf{x}_2), \ldots,], [\vec{s}(\mathbf{x}_1), \vec{s}(\mathbf{x}_2), \ldots,])$$

We implement the objective using Pyro software library (Bingham et al., 2019) and Adam optimizer.

B Proof of Proposition 1

We restate the result for clarity.
For a concept k and $cos(\alpha_k)$ defined as cos-sim$(e(v_k, f, \mathcal{D}), e(w_k, g, \mathcal{D}))$, 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(θ_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.

Proof. Corresponding to v_k in f, there must be an equivalent vector w in the embedding space of g.

$$cos(\alpha_k) = \text{cos-sim}(e(v_k, f, \mathcal{D}), e(w_k, g, \mathcal{D})) = \text{cos-sim}(e(w, g, \mathcal{D}), e(w_k, g, \mathcal{D}))$$

Denote the matrix of vectors embedded using g by $G = [g(\mathbf{x}_1), g(\mathbf{x}_2), \ldots, G(\mathbf{x}_N)]^T$ a $N \times D$ matrix (D is the dimension of g embeddings). Let U be a matrix with S basis vectors of size $S \times D$. We can express each vector as a combination of basis vectors and therefore $G = AU$ for a $N \times S$ matrix A.

Substituting the terms in the cos-sim expression, we have:

$$\begin{aligned} cos(\alpha_k) &= \text{cos-sim}(Gw, Gw_k) = \text{cos-sim}(AUw, AUw_k) \\ &= \frac{w^T U^T A^T AU w_k}{\sqrt{(w^T U^T A^T AUw)(w_k^T U^T A^T AU w_k)}}. \end{aligned}$$

If the examples in $\mathcal{D}$ are diversely distributed without any systematic bias, $A^T A$ is proportional to the identity matrix, meaning the basis of G and W are effectively the same. We therefore have $cos(\alpha_k) = \text{cos-sim}(Gw, Gw_k) = \text{cos-sim}(Uw, Uw_k)$, i.e. the projection of w, w_k on the subspace spanned by the embeddings have $cos(\alpha_k)$ cosine similarity. Since w, w_k are two vectors that are α_k apart, an arbitrary new example $\mathbf{x}$ that is at an angle of θ from w_k is at an angle of $\theta \pm \alpha_k$ from w. The cosine similarity follows as below.

$$\begin{aligned} cos(\theta) = \text{cos-sim}(w_k, g(\mathbf{x})) \implies \text{cos-sim}(w, g(\mathbf{x})) &= cos(\theta \pm \alpha_k) \\ &= cos(\theta)cos(\alpha_k) \pm sin(\theta)sin(\alpha_k) \end{aligned}$$

Because w is a vector in g corresponding to v_k in f, cos-sim$(w, g(\mathbf{x}))$ = cos-sim$(v_k, f(\mathbf{x}))$. □

C PROOF OF 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. We use the known result that inner product of two random vectors is close to 0 when the number of dimensions is large, i.e. $u_i^Tu_j \approx 0, i \neq j$.

Result with vanilla estimator. We first show the solution using vanilla estimator is distributed as given by the result above. We wish to estimate $v_1, v_2, \dots$ such that we approximate the prediction of model-to-be-explained: $y = w^T\mathbf{x}$. We denote by w_k sampled from the normal distributin of concept vectors. We require $w^T\mathbf{x} \approx \sum_k v_k w_k^T\mathbf{x}$. In effect, we are optimising for vs such that $\|w - \sum_k v_k w_k\|^2$ is minimized. We multiply the objective by u_k and use the result that random vectors are almost orthogonal in high-dimensions to arrive at objective $\arg\min_{v_k} \|w_k^Tw - v_k(w_k^Tw_k)\|$. Which is minimized trivially when $v_k = \frac{w_k^Tw}{\|w_k\|^2}$. Since w_k is normally distributed with $\mathcal{N}(u_k, \sigma_k^2I)$, $w_k^Tw = (u_k+\epsilon)^Tw, \quad \epsilon \sim \mathcal{N}(0, I)$ is also normally distributed with $\mathcal{N}(u_k^Tw, \sigma_k^2\|w\|^2)$. We approximate the denominator with its average and ignoring its variance, i.e. $\|w_k\|^2 = \mathcal{N}(\|u_k\|^2, \sigma_k^2) \approx \|u_k\|^2$ which is when $\|u_k\|^2 >> \sigma^2$. We therefore have the result on distribution of v_k.

Using U-ACE. Similar to vanilla estimator, U-ACE optimizes v_k using the following objective.

$$\ell = \arg\min_v\{\|w - \sum_k v_k u_k\|^2 + \lambda\sum_k \sigma_k^2 v_k^2\}$$

setting $\frac{\partial \ell}{\partial v_k} = 0$ and using almost zero inner product result above, we have

$$-u_k^T(w - \sum_j v_j u_j) + \lambda\sigma_k^2 v_k = 0$$

$$\implies v_k = \frac{u_k^Tw}{\|u_k\|^2 + \lambda\sigma_k^2}$$

□

D PROOF OF 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. We first show that the values of v_1, v_2 in closed form are as below before we derive the final result.

$$v_1 = \frac{\frac{S_1}{S_2}(1-\beta_2)^2}{\frac{S_1}{S_2}(\beta_2^2(1-\beta_1)^2 + \beta_1^2(1-\beta_2)^2) + \lambda(1-\beta_1)(1-\beta_2)}$$

$$v_2 = \frac{\frac{S_1}{S_2}(1-\beta_1)^2}{\frac{S_1}{S_2}(\beta_1^2(1-\beta_2)^2 + \beta_2^2(1-\beta_1)^2) + \lambda(1-\beta_1)(1-\beta_2)}$$

where $S_1 = \sum_i y_1$, $S_2 = \sum_i y_i^2$ and $\lambda > 0$ is a regularizing hyperparameter.

We then observe that if $\mathbf{x}$ is normally distributed then $y = w^T\mathbf{x}$ is also normally distributed with the value of $\frac{S_1}{S_2}$ is of the order $\mathcal{O}(1/N)$. Since β_1, β_2 are very close to 0, we can approximate the expression for v_1 as below.

$$v_1 \approx \frac{S_1}{S_2}(1-\beta_2)^2\frac{1}{\lambda(1-\beta_1)(1-\beta_2)} = \mathcal{O}(1/N\lambda)$$

□

Importance scores from a standard estimator.

When $c_1^{(1)} = (\beta_1 u + (1-\beta_1)v)^T z^{(i)}, \quad c_2^{(i)} = (\beta_2 u + (1-\beta_2)v)^T z^{(i)}$
we can estimate

$$\frac{(1-\beta_2)c_1 - (1-\beta_1)c_2}{(1-\beta_2)\beta_1 - (1-\beta_1)\beta_2} = \frac{(1-\beta_2)c_1 - (1-\beta_1)c_2}{\beta_1 - \beta_2} = u^T z_i = y_i$$

$\frac{1-\beta_2}{\beta_1-\beta_2}, \frac{1-\beta_1}{\beta_1-\beta_2}$

E Additional experiment details

List of fruit concepts from Section 4.1.

List of animal concepts from Section 4.2.

Scene labels considered in Section 4.3.

```
/a/arena/hockey, /a/auto_showroom, /b/bedroom, /c/conference_room, /c/corn_field
 /h/hardware_store, /l/legislative_chamber, /t/tree_farm, /c/coast,
 /p/parking_lot, /p/pasture, /p/patio, /f/farm, /p/playground, /f/field/wild
 /p/playroom, /f/forest_path, /g/garage/indoor
 /g/garage/outdoor, /r/runway, /h/harbor, /h/highway
 /b/beach, /h/home_office, /h/home_theater, /s/slum,
 /b/berth, /s/stable, /b/boat_deck, /b/bow_window/indoor,
/s/street, /s/subway_station/platform, /b/bus_station/indoor, /t/television_room,
 /k/kennel/outdoor, /c/campsite, /l/lawn, /t/tundra, /l/living_room,
 /l/loading_dock, /m/marsh, /w/waiting_room, /c/computer_room,
/w/watering_hole, /y/yard, /n/nursery, /o/office, /d/dining_room, /d/dorm_room,
 /d/driveway
```

E.1 Addition results for Section 4.3

We report also the tau (Wikipedia, 2023) distance from concept explanations computed by *Simple* as a measure of explanation quality. Kendall Tau is a standard measure for measuring distance between two ranked lists. It does so my computing number of pairs with reversed order between any two lists. Since *Simple* can only estimate the importance of concepts that are correctly annotated in the dataset, we restrict the comparison to only over concepts that are attributed non-zero importance by *Simple*.

Dataset↓	TCAV	O-CBM	Y-CBM	U-ACE
ADE20K	0.36	0.48	0.48	**0.34**
PASCAL	0.46	0.52	0.52	**0.32**

Table 3: *Quality of explanation comparison.* Kendall Tau Distance between concept importance rankings computed using different explanation methods shown in the first row with ground-truth. The ranking distance is averaged over twenty labels. U-ACE is better than both Y-CBM and O-CBM as well as TCAV despite not having access to ground-truth concept annotations.

F Extension of Simulation Study

Under-complete concept set. We now generate concept explanations with concepts set to {*"red or blue", "blue or red", "green or blue", "blue or green"*}. The concept *"red or blue"* is expected to be active for both *red* or *blue* colors, similarly for *"blue or red"* concept. Since all the concepts contain a color from each label, i.e. are active for both the labels, none of them must be useful for prediction. Yet, the importance scores estimated by Y-CBM and O-CBM shown in the Figure 4 table attribute significant importance. U-ACE avoids this problem as explained in Section 3.2 and attributes almost zero importance.

Concept	Y-CBM	O-CBM	U-ACE
red or blue	-75.4	-1.8	0.1
blue or red	21.9	-1.9	0
green or blue	-1.4	1.6	0
blue or green	-23.1	1.6	0

Table 4: When the concept set is under-complete and contains only nuisance concepts, their estimated importance score must be 0.