On Robustifying Concept Explanations

Elizabeth Chou and Amanda Boyd

October 4, 2023

ON ROBUSTIFYING CONCEPT EXPLANATIONS

Elizabeth J. Chou & Amanda Boyd & 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 [14]. What makes concept-based explanations particularly appealing is their alignment with semantically relevant patterns [31]. Research [14, 16] substantiates the preference for concept explanations over explanations derived from salient input features [25, 27] or prominent training examples [17]. The significance of concept explanations extends beyond interpretability alone. They also hold the potential to encode domain-specific prior knowledge effectively [32].

This paper centers on a category of interpretable methods known as concept bottleneck models (CBM)[18]. 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 first 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 [32, 21, 31, 9, 7, 28]. 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 [23, 2]. 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 [21, 32, 20] by harnessing pretrained multimodal models like CLIP [22] 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 scores, 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 scores, 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 Robust Average Concept Explanations (R-ACE), which we show is instrumental in improving reliability of explanations.

2 Preliminaries

Let the pretrained model-to-be-explained be $f : \mathbb{R}^D \to \mathbb{R}^L$. f output L scores for each possible category. Further, we use $f^{[k]}(\mathbf{x})$ to denote k^{th} hidden layer of the network. Given a dataset of examples: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ and a list of concepts $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$, CBE models aim to provide an "explanation" for the predictions of f using elements of $\mathcal{C}$. This is done via a two-stage procedure. In the first stage, concept activations are learnt from l^{th} layer representation of an example. Specifically, a vector v_k for k^{th} concept by is obtained as $\arg\max \mathbb{E}_{(x,y)\sim\mathcal{D}_k^{(k)}}[\ell(v^T f^{[l]}(\mathbf{x}), y)]$ where ℓ is the loss function (usually cross-entropy). The second stage obtains "explanations" by trying to emulate the prediction of f using concept scores obtained from the first stage. For this prcedure a variety of methos have been proposed [14, 26, 31]. 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. Oikarinen et al. [21] (O-CBM) estimates the learns to linearly project from the embedding space of CLIP [22] using its text description to the embedding space of the model-to-be-explained. 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.f.

Limitation: Unreliable Explanations. Major concerns regarding reliability of CBEs have been raised Ramaswamy et al. [23]. These issues have also been recently pointed by Anonymous [2]. A key issue is that explanations of the samr model can vary significantly with the choice of probe-dataset and the concept set bringing [2, 23]. Anonymous [2] provide a method called U-ACE focused on using uncertainty for improving CBE models.

3 Robust Average Concept Explanations

As outlined earlier, CBE models (CBMs) rely heavily on concept scores (concept scores) to generate explanations. The accuracy of concept scores is a key factor influencing the output of CBE models. Errors and uncertainty in concept scores can affect the subsequent stages and lead to poor explanations downstream if not accounted for. Additionally, assessing the significance of a concept is intractable if the concept is absent from the dataset.

Our approach has the following steps. (1) Estimate concept scores along with their error interval, (2) Compute and return a linear predictor model that is robust to input noise.

Estimation of concept scores and their error given an instance $\mathbf{x}$ denoted as $\vec{m}(\mathbf{x}), \vec{s}(\mathbf{x})$ respectively. Once concept scores 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 scores to their logit scores, i.e. $f(\mathbf{x}) \approx W_c\vec{m}(\mathbf{x})$. Since the concept scores contain noise, we require that W_c is such that predictions do not change under noise, i.e. $W_c\vec{s}(\mathbf{x}) \approx 0$.

We also add a sparsity constraint to W_c by using L_1 norm regularization. To further sparsify the model, the final wights are obtained by setting the values below a threshold to zero. The threshold is picked by hyper-parameter tuning; essentially we choose the highest threshold such that the cross-validation scores between the dense and sparse models remain close.

4 Experiments

We evaluate R-ACE on two synthetic and two real-world datasets. For evaluation of the method we follow the same procedure and datasets as Anonymous [2]. We analyse the reliability of our method against other baselines in Section 4.1. In Section 4.2, we assess the output of our method against known ground truth. Finally, we evaluate on real-world datasets. We experiment with the following models mentioned earlier: *TCAV* [14], *O-CBM* [21], *Y-CBM* based on [32] and *U-ACE* [2].

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.

4.1 Synthetic Data

Following Anonymous [2], we first experiment with a synthetic task where the model is trained to classify colours i.e. each input is a block image of a single colour. To add variability in the input, the random pixel noise is added to the inputs. The colours red and green are considered as the first label ($L = 1$) and the rest colours are considered label 2. The model is trained on a dataset with equal proportions of all colours. The concepts are defined by the colours with their literal name: *red, green, blue, white*. All the methods attribute positive importance for *red, green* and negative or zero importance for *blue, white* when explaining the first category. However, if the probe-dataset is not perfect i.e. has extraneous concepts, or a distribution shift then CBEs show considerably poor behaviour.

Unreliability due to misspecified concept set. Robustness to irrelevant and nuisance concepts is paramount for two key reasons. To evaluate such robustness of different methods, we experiment with different CBE methods on the synthetic data while varying the concepts made available in the probe-set. Specifically, following Anonymous [2] we expanded the concept set to include common fruit names as concepts along with the four initial colour concepts while using an in-distribution probe-dataset. Since CLIP embeddings of fruits also contain colour information, addition ofr fruits into the probe-set adds nuisance concepts. In Figure 1, we plots the score of the most salient fruit concept with increasing number of fruit (nuisance) concepts and note that R-ACE is far more robust to the presence of nuisance concepts.

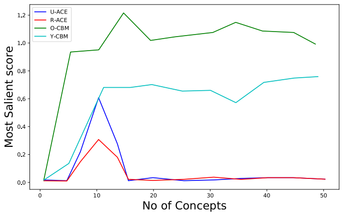

Figure 1: Score of most salient fruit against addition of concepts. We can see that R-ACE outperforms all other methods

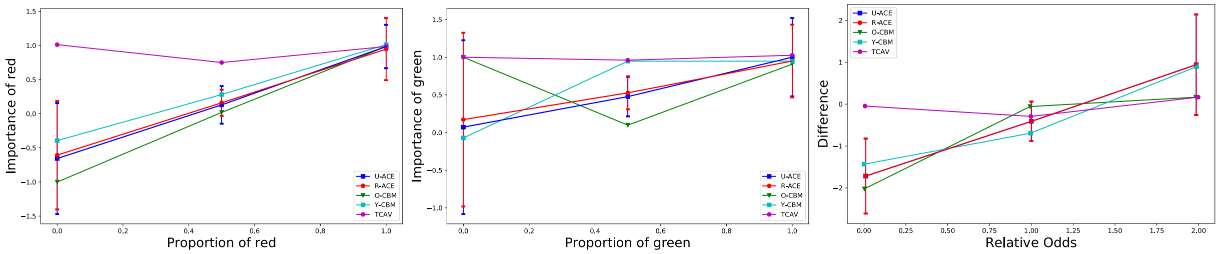

Figure 2: Left and middle plots show the importance of red and green concepts while the rightmost plot shows their importance score difference. R-ACE estimated large uncertainty in importance score when red or green concept is missing from the dataset as seen in the left of the left and middle plots.

Unreliability due to dataset shift. We varied the probe-dataset to include varying population of different colours. 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 2. In Figure 2, we plot the importance of a concept as assessed by the methods against their prevalence in the probe dataset. 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). If one relied on these scores, then one might assume that one colour is more important than another depending on the probe-dataset used. Both R-ACE and U-ACE provide a confidence band which can be used to understand the statistical significance of the predicted importances.

4.2 Assessment with known ground-truth

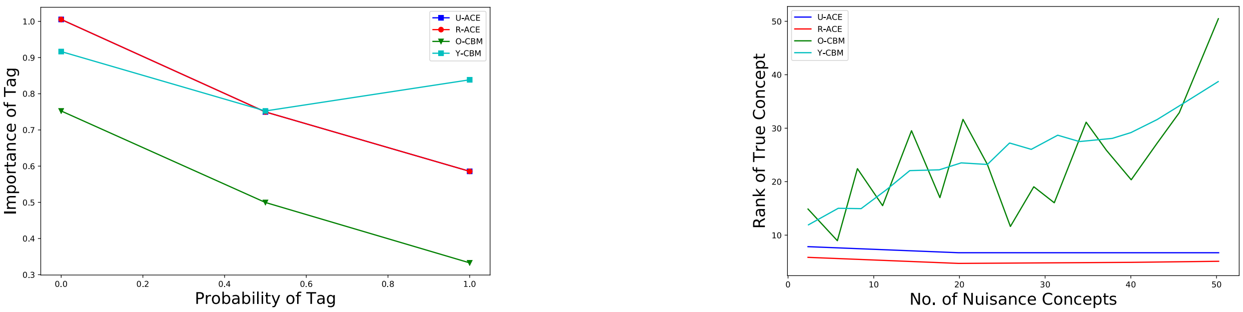

Figure 3: Left: Importance of a tag concept for three model-to-be-explained. X-axis shows the probability of tag in the training dataset of model-to-be-explained. Right: Average rank of true concepts with irrelevant concepts (lower is better).

Our objective in this section is to establish that R-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. Following Anonymous [2] we make a dataset using two labels from STL-10 dataset [10]: *car, plane* and paste a tag: *A* or *B* in the image. The probability that the examples of *car* are added the *A* tag is p and 1-p for the *B* tag. Similarly for the examples of *plane*, the probability of *A* is p and *B* 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.

To compare against reported results of Anonymous [2], we generate concept explanations using the concept set reported by them. These include seven car-related concepts and three plane-related concepts along with the two tags: *A, B*. We obtain the importance score of the concept A 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 3. Since the co-occurrence probability of A with *car* class goes from 1, 0.5 to 0, we expect the importance score of A should change from positive to negative as we move right. We note that R-ACE, along with others, show the expected decreasing importance of the tag concept. The result corroborates that R-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 overcomplete experiment of Section 4.1 and generated explanations as animal (irrelevent) concepts are added. Right panel of Figure 3 shows the average rank of true concepts (lower the better). We note that R-ACE generates expected explanations even with 50 nuisance concepts.

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]. Following the experimental setting of Ramaswamy et al. [23], we generate explanations using PASCAL [8] or ADE20K [34] that are part of the Broden dataset collection [5]. The dataset contains images with dense annotations with more than 1000 attributes. We removed the 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 evaluate quality of explanations by their closeness to the explanations generated using the baseline of [24, 23]. The baseline estimates explanation using concept annotations and hence regarded as close to the ground-truth. For the top-20 concepts identified this way, we compute the average absolute difference in importance scores estimated using any estimation method. Table 1 presents the deviation in explanations averaged over all the 50 scene labels.

Dataset shift.

Ramaswamy et al. [23], Anonymous [2] provide concrete evidence of the significant variability in concept explanations when using ADE20K or PASCAL as the probe dataset for the same model. The explanations change due to a couple of factors: firstly, the population of concepts can differ between datasets, impacting their perceived importance when employing standard methods, and secondly, the variance in explanations. As shown in earlier section, R-ACE estimated scores have better variance . Furthermore, R-ACE attributes high uncertainty, and consequently near-zero importance, to concepts that are rare or absent in the probe dataset. Given these factors, we anticipate that R-ACE can address the data-shift problem effectively. 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.

Dataset↓	TCAV	O-CBM	Y-CBM	U-ACE	R-ACE
ADE20K	0.13	0.19	0.16	**0.09**	**0.09**
PASCAL	0.41	0.20	0.18	0.11	**0.08**

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

TCAV	O-CBM	Y-CBM	U-ACE	R-ACE
0.41	0.32	0.33	0.19	**0.12**

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 CBMs use predefined human-interpretable concepts as intermediate features for predictions [18, 4, 14, 35]. These additionaly provide the ability for human intervention in terms of weighing the concepts [3]. However, traditional CBMs rely on large labeled data with concept annotations but recent research have suggested incorporated large pretrained multimodal models like CLIP [22] to handle the task of producing the concept annotations [21, 32]. Despite these advancements, ensuring the reliability of CBMs remains a challenge, especially concerning the information leakage problem [13, 19]. Concept Embedding Models (CEM) [11] attempt to tackle the trade-off between accuracy and interpretability. Close to this work, Kim et al. [15] proposed the Probabilistic Concept Bottleneck Models (ProbCBM) emphasizing the need to model uncertainty in concept prediction. Anonymous [2] have also studies the idea of using uncertainty to make explanations more reliable. Our proposed method and presentation is inspired heavily from Anonymous [2]. The key difference lies in how we incorporate the uncertainity during training.

Concept based explanations CBEs entails learning concepts and subsequently decomposing individual predictions or overall label features of model-to-be-explained using a probe dataset of concept annotations. for explanation. Our proposed method is a form of concept based explanations (CBE) [14, 4, 35, 12]. Existing CBE methods, as pointed out by Ramaswamy et al. [23], have major drawbacks: a) the concepts can sometimes be more intricate to learn than the labels themselves, potentially rendering explanations non-causal; and b) concepts learned are sensitive to the probe dataset and hence are not reliable. Additionally, these methods often employ a number of concepts that far exceed what an average human can easily comprehend. Achtibat et al. [1] advocate for an approach that emphasizes important features, addressing the "where" aspect, and identifies the concepts used for prediction. Further uses of CBEs include explanations for out-of-distribution detector models [9], and explanations for NLP models using counterfactual texts [30]. Similar to us Moayeri et al. [20] explored use of CLIP to interpret representations of a different model trained on uni-modal data.

6 Conclusion

We delved into incorporation of uncertainity into concept explanation methods, with while using data-efficient training. We highlight the the reliability issues encountered by current concept explanation estimators, arguing for a need to incorporate uncertainity during learning eplanations. Consequently, we introduced a novel uncertainty-aware and data-efficient estimator and rigorously tested it using controlled experiments and challenging real-world evaluations. The outcomes of these assessments unequivocally demonstrated the superiority of concept explanations generated by our method, R-ACE, showcasing their heightened reliability and robustness.

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] Anonymous. Concept explanations should be uncertainty aware. *arXiv preprint*, 2023.

[3] 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.

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*, 2021.

[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.