Very Short and Accurate Explanations by Design

Anonymous ICCV submission

Paper ID *****

Abstract

Deep models are "black-box", meaning that their decision-making process is often not transparent to users. To address this issue, several post hoc methods have been proposed for explaining the model's predictions. However, post hoc explanations are often unreliable and not faithful to the model. Interpretable-by-design methods, such as Information Pursuit (IP) and its variants, map the input data to a small set of interpretable concepts by asking a set of queries, and make a prediction based on the sequence of query answers. Such models are faithful by design because their predictions are based on the explanations, i.e. the sequence of query answers. However, they require either a very complex algorithm for selecting which queries to ask or fully annotated datasets for training a query-answering system. This paper proposes IP-OMP-ConceptQA, an interpretable-by-design method that combines an efficient query selection method (OMP) with an accurate zero-shot query answering system (Concept-QA). Experiments on vision data sets show that IP-OMP-ConceptQA outperforms existing methods in terms of accuracy, interpretability, faithfulness, and efficiency in scenarios where very short explanations are desired.

1. Introduction

The lack of transparency of machine learning models has raised the question of whether these "black-box" models can be trusted [6]. For instance, when using such models in healthcare applications, an answer to questions such as "Why does a deep neural network classify a tumor detected in an MRI as benign or malignant?" can have life-saving consequences. In such cases, understanding how a prediction is made is just as important as achieving high accuracy.

Related work. Most existing methods for interpreting the predictions of a machine learning model are *post hoc*, i.e., they aim to explain the prediction after it has been made [9, 12, 14]. Post hoc explanations typically assign an importance score to each input feature which depends on the sensitivity of the model's output with respect to each feature. However, such explanations often do not reliably or faithfully represent the model's decision-making process [1, 13].

Interpretable-by-design algorithms, such as Concept Bottleneck Models (CBMs) [7] or Information Pursuit (IP) [2], address these issues by producing an explanation that is interpretable to users as part of their prediction process.

CBMs use a concept predictor network to map each input image to a feature vector whose entries measure whether a human-interpretable concept is present or not in the image. A linear classifier is then applied to this concept vector to predict the class, and the concepts with the highest classifier weights are chosen as an explanation for the prediction. CBMs bring significant advantages relative to post hoc explanation methods in terms of both faithfulness and interpretability. However, their accuracy is hampered by the use of linear classifiers, which is done to facilitate the selection of the concepts that form an explanation. In addition, CBMs require predicting a large number of concepts, all at once, while in practice very few concepts may be sufficient to provide an explanation for a prediction. Moreover, CBMs require a fully annotated dataset to train a concept predictor.

IP addresses these issues by playing a Twenty Questions game in which very few concepts are queried, one at a time, until a reliable prediction can be made based on the selected concepts. Implementing IP requires three ingredients (see Fig. 1): a *querier* that selects which concepts to query and in what order, an *answerer* that predicts whether a concept selected by the querier is present or not in the image, and a *classifier* that predicts the class from the sequence of query-answer pairs. IP selects queries whose answers maximize mutual information with the labels (this requires learning a generative model and may need an exponential number of samples [11]), answers the queries using a fully supervised concept predictor (this requires fully annotated datasets as for CBMs), and uses a nonlinear network for classification (which improves classification accuracy relative to CBMs).

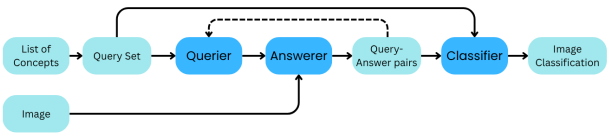

Figure 1. Illustration of the IP framework.

Variational Information Pursuit (VIP) [3] improves IP's query selection efficiency by jointly training a querier and classifier networks so that the querier selects concepts that best improve the classifier's accuracy. However, VIP still uses a fully supervised concept predictor. VIP-ConceptQA [5] addresses this issue by using CLIP and GPT to generate pseudo-labels to train a zero-shot concept question answerer (ConceptQA). However, both VIP and VIP-ConceptQA are limited to small- to medium-scale tasks, because the querier and classifier must learn from an exponential number of query-answer pairs, resulting in a slow training process.

IP-OMP [4] addresses this issue by mapping both the image to be classified and the query to CLIP space and using Orthogonal Matching Pursuit (OMP) to select queries. Specifically, OMP represents the embedded image as a linear combination of *a few* embedded queries, those with the highest CLIP dot products with the embedded image minus the contribution from prior queries. That is, IP-OMP selects the queries via sparse coding, bypassing the need to train a querier. A linear classifier is then trained on the sparse codes (IP-OMP-SparseCode) or the sequence of CLIP dot products (IP-OMP-CLIP), both interpreted as zero-shot answers to the queries. In practice, however, these continuous-valued answers are not interpretable relative to ConceptQA binary answers. Therefore, IP-OMP is more efficient than VIP-ConceptQA, but at the cost of reduced interpretability.

In short, existing interpretable-by-design methods either use inefficient query selection methods (CBM, IP, VIP, VIP-ConceptQA), or non zero-shot query answering methods that require huge annotation effort (CBM, IP, VIP) or inaccurate/uninterpretable zero-shot query answering methods (IP-OMP), as shown in Table 1.

Table 1. Prior explainable-by-design methods use computationally intense query selection methods or inaccurate zero-shot answering methods. Our method (IP-OMP-ConceptQA) combines an efficient query selection method (OMP) with an accurate zero-shot query answering method (Concept-QA) to improve classification accuracy when very short explanations are desired.

	Efficient Query Selection?	Accurate Zero-Shot Answers?
CBM [7]	No selection	Not zero-shot
IP [2]	No	Not zero-shot
VIP [3]	Somewhat	Not zero-shot
VIP-ConceptQA [5]	Somewhat	Yes
IP-OMP [4]	Yes	No
IP-OMP-ConceptQA	Yes	Yes

Paper contributions. This paper proposes an interpretable-by-design approach to image classification that combines an efficient query selection method (IP-OMP) with an accurate and interpretable zero-shot query answering method (Concept-QA). The proposed approach (IP-OMP-ConceptQA) outperforms existing methods in scenarios where very short explanations are desired.

2. Methodology

The proposed IP-OMP-ConceptQA framework is illustrated in Fig. 2 and consists of a *querier* (IP-OMP), an *answerer* (ConceptQA), and a neural network *classifier*.

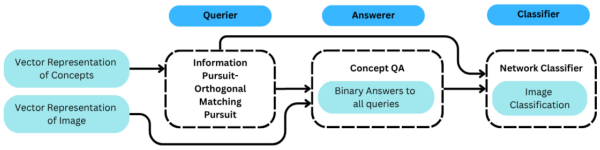

Figure 2. Illustration of the IP-OMP-ConceptQA framework.

2.1. IP-OMP Querier

Queries. An interpretable-by-design framework requires a set of interpretable queries that are sufficiently adequate and well-suited to the task at hand. While a querier could select all possible queries, doing so would decrease interpretability due to redundant or non-relevant information. Additionally, works such as Label-free CBMs demonstrate that narrowing down the list of concepts results in better performance and interpretability [10]. As a result, this approach uses a querier, IP-OMP, that selects a concise list of informative, task-relevant queries to guide the classification process, and more importantly, in a time-efficient manner.

IP-OMP. The IP-OMP querier starts with a dictionary of atoms $D = [d_1, d_2, \ldots, d_j] \in \mathbb{R}^{m \times n}$, where each atom d_j is the CLIP embedding of the text corresponding to the jth visual query [4]. For example, the query set for the CUB200 dataset consists of text descriptions of certain visual features that could be seen in birds (e.g., "while feather", "brown beak"), and the dictionary atoms are their CLIP embeddings.

To select queries for an given image, IP-OMP represents the CLIP embedding $x \in \mathbb{R}^m$ of this image as a sparse linear combination of the dictionary, i.e., it finds a sparse vector $\beta \in \mathbb{R}^n$ such that:

$$x \approx D\beta. \tag{1}$$

Then the dictionary atoms that have non-zero coefficients in β represent the queries that are selected for this image. A greedy algorithm for finding a sparse β is Orthogonal Matching Pursuit (OMP), which alternates between a *least-squares estimation step* where β is updated given an estimate for its support Λ, and a *support selection step* where the support Λ is updated given an estimate for β.

Specifically, the algorithm starts with an empty index set $\Lambda_0 = \emptyset$, i.e., none of the atoms/queries is initially selected.

Then, at the k-th step, a *least-squares estimation step* finds the sparse coefficients as:

$$\beta_k = \arg\min_{\beta} \|x - D_{\Lambda_k}\beta_k\|_2^2 = (D_{\Lambda_k}^\top D_{\Lambda_k})^{-1} D_{\Lambda_k}^\top x\,, \quad (2)$$

where D_{Λ_k} be the matrix whose columns are those from D with indices in the current set of indices $\Lambda_k \subset [n]$. Note that we conveniently let the solution be the zero vector when Λ_k is empty. Next, a *support selection step* identifies an index outside Λ_k whose corresponding atom has the highest correlation with the current residual, $r_k - D_{\Lambda_k}\beta_k$, i.e.:

$$j_k = \arg\max_j |\langle d_j, r_k\rangle|\,, \quad (3)$$

and updates the index set: $\Lambda_{k+1} = \Lambda_k \cup \{j_k\}$. This process is repeated until the specified number of atoms has been selected, i.e., the query set length has reached a desired value.

The work of [4] establishes a connection between IP and OMP. Specifically, [4] shows that under certain distributional assumptions on the data and the queries, finding the query whose answer has maximum mutual information with the random variable to be inferred is equivalent to finding the query with the highest dot product with the residual up to a normalization of the dot product. This modified query selection algorithm is called IP-OMP as combines the best aspects of IP and OMP: a query selection that is informative due to IP, and computationally efficient due to OMP.

2.2. ConceptQA Answerer

Once a query has been selected by IP-OMP, an answer for it needs to be found. The ConceptQA answerer takes both a query and an image as inputs, and determines whether the query is true or false based on the image. ConceptQA is trained using a set of image-concept pseudo-labels generated from GPT and CLIP using the following steps:

1. For every concept in the query set and image-label pair, GPT is asked whether the concept is important for determining the label.
2. If GPT replies "No," then the pseudo-label for a concept-image pair is "No."
3. If GPT replies "Yes", the dot product between the CLIP embedding of the concept and the CLIP embedding of the image is used to determine whether the concept is present in the image.

2.3. Classifier

Two different classifiers are tested and used to determine the class of the image: a Linear Classifier (Logistic Regression) and Network Classifier (Multilayer Perceptron). Both classifiers operate the sequence of query-answer pairs produced by ConceptQA and trained using the cross-entropy loss.

3. Experiments and Results

In this section, the proposed IP-OMP-ConceptQA model is evaluated on three widely used image classification data sets: CIFAR10 [8], CIFAR100 [8], and CUB200 [15]. The CIFAR10 data set consists of 60,000 images with 10 classes of 6,000 images each, including airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. CIFAR100 is an extended version of CIFAR 10 with 100 classes and 600 images per class, for a total of 60,000 images, while the CUB200 data set has 11,788 images of 200 bird categories. Each model was trained on each data set separately.

Test accuracy versus explanation length. Fig. 3 compares IP-OMP-ConceptQA against four other approaches, IP-OMP-CLIP, IP-OMP-SparseCode [4], VIP-CLIP [3] and VIP-ConceptQA [5], on all three datasets using test accuracy vs. explanation length (average number of query-answer pairs) as the evaluation metric. Notice that IP-OMP-ConceptQA outperforms all other methods on all data sets for very short explanation lengths. Specifically, on CIFAR10, IP-OMP-ConceptQA achieves the highest accuracy among all methods for explanation lengths up to six, and is competitive with VIP-ConceptQA for longer explanations. On CIFAR100, IP-OMP-ConceptQA outperforms all methods for explanation lengths up to nine, and is still competitive with VIP-Concept QA for longer explanations, with a drop in performance of less than 4%. On CUB200, IP-OMP-ConceptQA maintains superior performance up to eleven queries, and is competitive with VIP-Concept QA for longer explanations, with a drop in performance of less than 6%.

Evaluation of IP-OMP-ConceptQA explanations. Fig. 4 presents a qualitative evaluation of the explanations produced by IP-OMP-ConceptQA on two images from the CUB data set. Given an image to be classified, the sequence of questions selected by IP-OMP, the (binary) answers by ConceptQA and the (continuous) values of the sparse coefficients by IP-OMP, which can be considered as soft answers to the questions, are shown. For each query, a positive coefficient (yes, the query is present) is shown in green and a negative coefficient (no, the query is not present) is shown in red. Notice that the values of the sparse coefficients are frequently in the range (-0.3,0.3), making it difficult for humans to interpret whether the concept is present or not. On the other hand, Concept QA provides binary answers in $\{-1, 1\}$, which are easier to interpret for humans.

4. Conclusion

This work presented IP-OMP-ConceptQA, an interpretable-by-design method that combines an efficient query selection method (querier) with an accurate zero-shot query answering system (answerer) to produce an interpretable representation (sequence of query-answer pairs) for classifica-

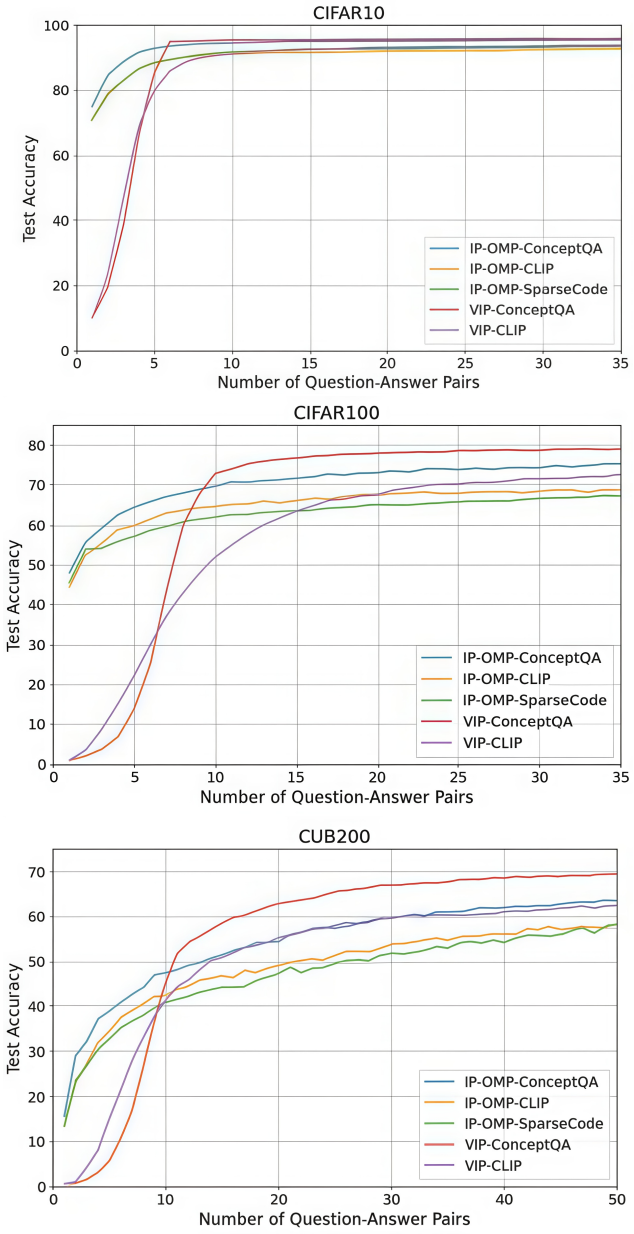

Figure 3. Test accuracy of five algorithms as a function of the number of query-answer pairs on CIFAR10, CIFAR100 and CUB200.

tion. Experiments on vision data sets showed that IP-OMP-ConceptQA outperforms existing methods in terms of accuracy, interpretability, faithfulness, and efficiency in scenarios where very short explanations are desired. Specifically, IP-OMP is better than VIP at selecting queries when the explanation length is short (small number of query-answer pairs), and VIP is better than IP-OMP when the explanation length is long. Regarding answering queries, Concept QA consistently improves accuracy for both VIP and IP-OMP.

Future work could focus on creating a hybrid algorithm

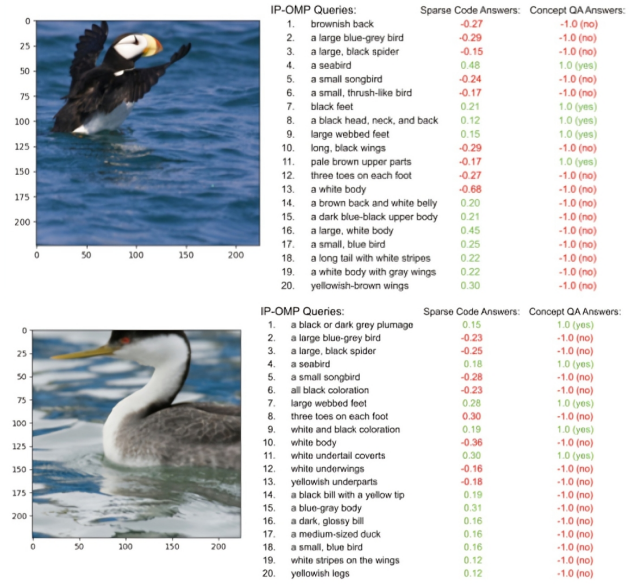

Figure 4. ConceptQA and SparseCode explanations for two images from the CUB data set. The input image is shown on the left, the queries selected by IP-OMP are shown on the center, and the answers by SparseCode and ConceptQA are shown on the right.

that would use IP-OMP-ConceptQA for the initial set of queries and then transition to VIP-ConceptQA for the remaining queries. Ideally, this algorithm would combine the best features of all current algorithms, leading to more efficient, faithful, and accurate predictions.

References

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

[2] Aditya Chattopadhyay, Stewart Slocum, Benjamin D. Haeffele, René Vidal, and Donald Geman. Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[3] Aditya Chattopadhyay, Kwan Ho Ryan Chan, Benjamin David Haeffele, Donald Geman, and René Vidal. Variational information pursuit for interpretable predictions. In *ICLR*, 2023. 2, 3

[4] Aditya Chattopadhyay, Ryan Pilgrim, and René Vidal. Information maximization perspective of orthogonal matching pursuit with applications to explainable AI. In *NeurIPS*, pages 2956–2990, 2023. 2, 3

[5] Aditya Chattopadhyay, Kwan Ho Ryan Chan, and René Vidal. Bootstrapping variational information pursuit with large language and vision models for interpretable image classification. In *ICML*, 2024. 2, 3

[6] David Gunning and David Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58, 2019. 1

[7] P. W. et al. Koh. Concept bottleneck models. In *Proceedings of Machine Learning Research*, 2020. 1, 2

[8] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 3

[9] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017. 1

[10] T. et al. Oikarinen. Label-free concept bottleneck models. In *ICLR*, 2023. 2

[11] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, pages 5171–5180, 2019. 1

[12] M. T. Ribeiro, S. Singh, and C. 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*, 2016. 1

[13] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[14] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2013. 1

[15] C. et al. Wah. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 3