Q-SENN: Quantized Self-Explaining Neural Networks

Thomas Norrenbrock, Marco Rudolph, Bodo Rosenhahn

Institute for Information Processing (tnt)
L3S - Leibniz Universität Hannover, Germany
{norrenbr, rudolph, rosenhahn}@tnt.uni-hannover.de

Abstract

Explanations in Computer Vision are often desired, but most Deep Neural Networks can only provide saliency maps with questionable faithfulness. Self-Explaining Neural Networks (SENN) extract interpretable concepts with fidelity, diversity, and grounding to combine them linearly for decision-making. While they can explain *what* was recognized, initial realizations lack accuracy and general applicability. We propose the Quantized-Self-Explaining Neural Network "Q-SENN". Q-SENN satisfies or exceeds the desiderata of SENN while being applicable to more complex datasets and maintaining most or all of the accuracy of an uninterpretable baseline model, outperforming previous work in all considered metrics. Q-SENN describes the relationship between every class and feature as either positive, negative or neutral instead of an arbitrary number of possible relations, enforcing more binary human-friendly features. Since every class is assigned just 5 interpretable features on average, Q-SENN shows convincing local and global interpretability. Additionally, we propose a feature alignment method, capable of aligning learned features with human language-based concepts without additional supervision. Thus, what is learned can be more easily verbalized. The code is published: https://github.com/ThomasNorr/Q-SENN

Introduction

The ability to comprehend the decision-making process of deep learning models is gaining significance, especially for safety-critical applications like autonomous driving or medical diagnosis, where practitioners and legal requirements necessitate a thorough understanding of the decision and its reasoning (Molnar 2020; Bibal et al. 2021). However, the high dimensionality of image data has made it challenging to develop interpretable models for computer vision, with many problems still unsolved. Alvarez Melis and Jaakkola (2018) proposed *Self-Explaining Neural Networks* (SENN) which can be described as linear combination of interpretable concepts extracted from the input, but their realization lacked accuracy and applicability to complex datasets beyond MNIST. They had three desiderata for the concepts: **Fidelity** refers to relevancy, **Diversity** to non-overlapping concepts, and **Grounding** to the alignment with a human concept. We propose the *Quantized-SENN* (Q-SENN), which satisfies these requirements. Q-SENN is motivated by former works such as the SLDD-Model (Norrenbrock, Rudolph, and Rosenhahn 2022), *PIP-Net* (Nauta et al. 2023), *ProtoPool* (Rymarczyk et al. 2022) or the *Concept Bottleneck Model* (CBM) (Koh et al. 2020), as they all aim to obtain human-understandable concepts as input to an interpretable classifier C. Our proposed Q-SENN improves the interpretability of C compared

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

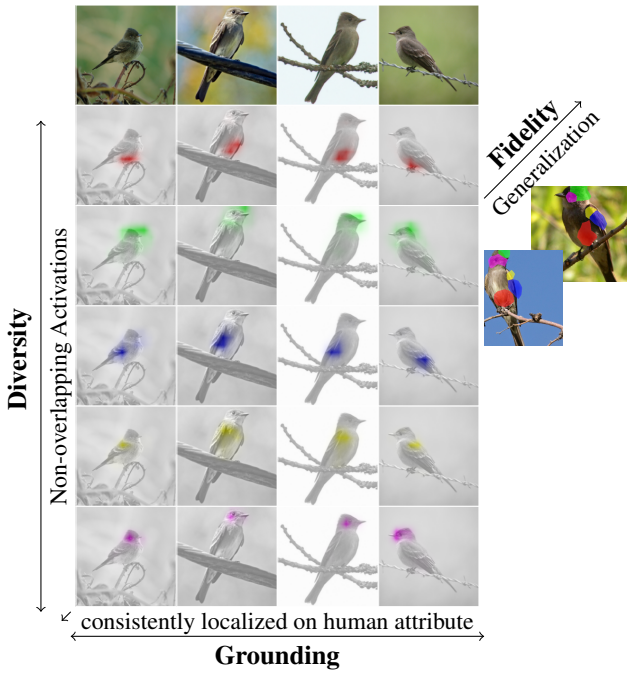


Figure 1: Q-SENN optimizes for Diversity, Grounding and Fidelity: The global explanation shows one example class being recognized through 5 interpretable features that show high Diversity and Grounding, consistently localizing on the same meaningful human attributes across images, *e. g.*, *belly, crown, upper tail, upper wing and eye.* When measuring Fidelity, the features generalize to unseen data and the local explanation fits the class explanation. Visualization techniques are based on overlaying color-coded feature maps, described in the supplementary material.

to all competitors as it is extremely sparse, using just very few, usually 5 to be easy for all humans to follow (Miller 1956), features to recognize a class, and ternary. That way, the relationship between a feature and a class is either a positive ($+1$), negative (-1) or neutral (0) assignment with no sublevels. For example, a dog might be positively related to the *four-legged* feature, neutral to most features, such as color, and negatively related to features that might be required to differentiate it from other classes in the dataset, *e. g.*, *feline* if multiple classes of cats are present. Additionally, the features used for C exhibit improved desiderata of SENN, thus making Q-SENN both globally and locally interpretable, as shown in Figure 1. Finally, Q-SENN is more broadly applicable, as it does not need additional annotations and scales to ImageNet (Russakovsky et al. 2015). We create our Q-SENN through an iterative process of calculating the sparse ternary layer and fine-tuning the model with these fixed feature-class assignments. We select a sufficiently low number of features to ensure a sharing between the classes. This leads to the emergence of more grounded, general features during one iteration and subsequent removal of an assignment based on spurious correlations on the next iteration. That way, the model converges to more robust assignments between grounded features and classes and shows exceptional robustness to spurious correlations. As the ternary structure prohibits features from differentiating between two assigned classes, they naturally become more diverse, general and binary leading to improved explanations (Lipton 1990) and generalization on all investigated datasets. The social science survey by Miller (2019) suggests classifying the explanations of Q-SENN as human-friendly due to them being contrastive, concise and general. Additionally, we introduce metrics to estimate the **Grounding** of learned features as human concepts and distinguish them from class detectors. Overall, we demonstrate that the proposed Q-SENN is an interpretable model where every class is assigned on average 5 diverse and more grounded features as depicted in Figure 1 with an exceptional robustness to spurious correlations. Finally, while the features of Q-SENN are more alignable with any human concept, thus show **Grounding**, the manual alignment of learned features with human concepts is laborious. To facilitate this process, we demonstrate a method to automatically align the features of our proposed Q-SENN with human interpretable concepts using CLIP (Radford et al. 2021) without requiring additional annotations and validate the method using the attributes contained in CUB-2011 (Wah et al. 2011).

Our main **contributions** are as follows:

- We propose quantization for sparse decision layers in an iterative fine-tuning loop, leading to an Quantized-Self-Explaining Neural Network (Q-SENN) which is easy to interpret for humans.
- We measure the self-explaining quality via **Fidelity**, **Diversity** and **Grounding** and show significant improvements on these metrics compared to previous work.
- We demonstrate the high interpretability, increased accuracy and exceptional robustness to spurious correlations of our Q-SENN on several benchmark datasets and architectures for image classification.
- We propose a method for automatically aligning the features with human interpretable concepts without additional annotations using CLIP.

Related Work

Interpretable machine learning refers to models that are interpretable by design, as well as post-hoc methods that attempt to explain what a model has learned. Interpretability can be categorized as either local or global, referring to the interpretability of a single decision or the entire model, respectively (Molnar 2020).

Research towards global post-hoc interpretability attempts to align learned representations with human-understandable concepts (Kim et al. 2018; Bau et al. 2017; McGrath et al. 2022; Fel et al. 2023; Yuksekgonul, Wang, and Zou 2022; Zhang et al. 2018). While they usually rely on auxiliary data, we introduce a new such method for our proposed model with no need for additional annotations. Saliency maps are local explanations showing which part of an input image is relevant to a prediction. While post-hoc methods, *e. g.* Grad-CAM (Selvaraju et al. 2017), often lack properties such as shift invariance and faithfulness (Kindermans et al. 2019; Adebayo et al. 2018), interpretable models with built-in saliency maps (Böhle, Fritz, and Schiele 2022, 2023; Stalder et al. 2022; Zhang, Wu, and Zhu 2018) provide more faithful localizations. Interpretable models are becoming increasingly relevant, since desired properties, that additionally improve the effectiveness of post-hoc techniques such as ours, can be built-in. For example, Q-SENN's low number of total features and their easy-to-interpret sparse ternary assignment increases the value of aligning a single feature with a human concept using our proposed alignment method.

Another type of interpretable neural network is the *Self-Explaining Neural Network* (Alvarez Melis and Jaakkola 2018) (SENN). Its most simple form is described as

$$y(\boldsymbol{x}) = C(\boldsymbol{x})^T f(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x}$ refers to the input, $f(\boldsymbol{x})$ to concepts extracted from the input and $C(\boldsymbol{x})$ to class specific weights for the concepts which should be independent of small changes to $f(\boldsymbol{x})$. Additionally, SENN postulates three desiderata for the concepts: **Fidelity** refers to the preservation of relevant information about the input. **Diversity** describes the need for non-overlapping concepts. Finally, **Grounding** refers to an alignment of the concept with a human interpretable one. We consider our model as Self-Explaining Neural Network with constant, quantized and sparse $C(\boldsymbol{x})$. This exceeds the desired stability of $C(\boldsymbol{x})$ and makes the computationally expensive optimization obsolete, enabling the application to complex datasets.

Alternatively, models like *ProtoTree* (Nauta, van Bree, and Seifert 2021), *ProtoPNet* (Chen et al. 2019), *ProtoPShare* (Rymarczyk et al. 2021), *ProtoPool* (Rymarczyk et al. 2022) and *PIP-Net* (Nauta et al. 2023) use a deep feature extractor to learn prototypes from data. The similarities to these prototypes are then fed into interpretable models. Their interpretability is however unclear, as Kim et al. (2022) and Hoffmann et al. (2021) showed a discrepancy between the

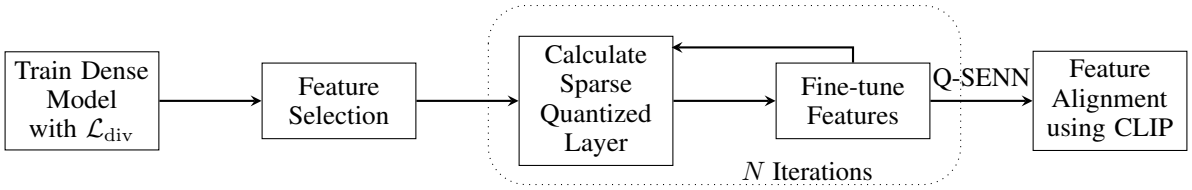


Figure 2: Overview of our proposed pipeline to construct a Q-SENN.

human and computed similarities. Our proposed model is to some extent similar to *ProtoPool* or *PIP-Net*, as it also shares a small set of total features between all classes and only uses very few per class. Instead of prototypes, our Q-SENN uses fewer features in total and per class, while showing increased accuracy. The *Concept Bottleneck Model* (CBM) (Koh et al. 2020) initially predicts the annotated concepts in a given dataset, and subsequently employs a basic model to forecast the target category from the identified concepts. This notion has undergone further investigation and expansion (Sawada and Nakamura 2022; Zarlenga et al. 2022). Both Margeloiu et al. (2021) and Ramaswamy et al. (2022) suggest that end-to-end training of the CBM can result in the encoding of additional information beyond the targeted concepts, leading to a reduction in interpretability. Marconato, Passerini, and Teso (2022) address this issue by disentangling the features and increase alignment. Our proposed method differs from CBM, as it does not use supplementary annotations and generates a decision function C that is considerably more interpretable, as it is sparse and quantized. Finally, Liu et al. (2023) showed that ternary networks can be used to compress models while maintaining high accuracy. For high levels of sparsity, Glandorf, Kaiser, and Rosenhahn (2023) demonstrate a similar result. Instead of compression, Q-SENN offers interpretability through its ternary and sparse final layer. While measuring the interpretability of deep neural networks is an open task, increased interpretability is typically measured by reducing model complexity, such as the number of operations (Yang, Rudin, and Seltzer 2017; Slack et al. 2019; Rüping et al. 2006) or the number of features (Rüping et al. 2006). This motivated recent work on interpretable machine learning (Nauta et al. 2023; Rosenhahn 2023) and the SLDD-Model (Norrenbrock, Rudolph, and Rosenhahn 2022).

SLDD-Model

The *Sparse-Low-Dimensional-Decision-Model* (SLDD-Model), uses on average $n_{wc} = 5$ features out of $n_f^* = 50$ total features in the final layer to recognize a single class, since humans can follow a decision based on five cognitive aspects (Miller 1956). After training a model with the *Feature Diversity Loss* $\mathcal{L}_{\text{div}}$, which penalizes highly activated and weighted features that localize on the same region, a subset of features is selected in a supervised manner and *glm-saga* (Wong, Santurkar, and Madry 2021) is used to compute the regularization path, from which they chose a solution with on average 5 nonzero weights per class. The model is then fine-tuned with the final layer fixed to this solution, *s. t.* the features adapt to it. The resulting SLDD-Model shows competitive accuracy while being more interpretable. We use the SLDD-Model as baseline for our experiments and demonstrate that our method increases all desiderata of a SENN by quantizing the sparse matrix and iteratively optimizing it.

Method

We evaluate our proposed Q-SENN in image classification. The task involves the classification of an image $\boldsymbol{I} \in \mathbb{R}^{3\times w\times h}$ of dimensions w and h into a single class $c \in \{c_1, c_2, \ldots, c_{n_c}\}$ using a deep neural network Φ as feature extractor and classifier network C. The network Φ extracts feature maps $\boldsymbol{M} \in \mathbb{R}^{n_f\times w_M\times h_M}$ and aggregates them into a feature vector $\boldsymbol{f} \in \mathbb{R}^{n_f}$. The final output is obtained by applying C to the feature vector, resulting in the output vector $\boldsymbol{y} \in \mathbb{R}^{n_c}$ as $\boldsymbol{y} = C(\boldsymbol{f})$.

Q-SENN

The proposed pipeline to create a Q-SENN is shown in Figure 2. We first train a dense model and then select n_f^* out of the n_f features, following the SLDD-Model, to compute a low-dimensional sparse quantized final layer for the selected features using *glm-saga*. The model is then fine-tuned with the final layer fixed to that computed solution, *s. t.* the features adapt to the assigned classes. The cycle of computing an interpretable final layer and fine-tuning the model is repeated N times. Afterwards, the learned features can be aligned with human concepts in form of natural language to create a model with verbalized explanations. One such method using CLIP is presented at the end of this paper.

Q-SENN is considered as SENN with constant quantized $C(\boldsymbol{f}) = \boldsymbol{y} = \boldsymbol{W}^Q\boldsymbol{f} + \boldsymbol{b} = \alpha\boldsymbol{W}^1\boldsymbol{f} + \boldsymbol{b}$ with the ternary sparse weight matrix $\boldsymbol{W}^Q \in \{-\alpha, 0, \alpha\}^{n_c\times n_f^*}$, or matrix of ones $\boldsymbol{W}^1 \in \{-1, 0, 1\}^{n_c\times n_f^*}$, and bias $\boldsymbol{b} \in \mathbb{R}^{n_c}$. Therefore, we want the features $\boldsymbol{f}$ of our model to show Fidelity, Diversity and Grounding. In this work, Fidelity is measured as accuracy, as preserving relevant information is required to accurately predict the class. For Diversity, the Feature Diversity Loss $\mathcal{L}_{\text{div}}$ (Norrenbrock, Rudolph, and Rosenhahn 2022) is utilized to ensure a high local diversity of the feature maps $\boldsymbol{M}$ that are used for the same class. The training loss is calculated as

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{CE} + \beta\mathcal{L}_{\text{div}} \tag{2}$$

with $\beta \in \mathbb{R}_+$ as weighting factor between cross-entropy loss and $\mathcal{L}_{\text{div}}$. Additionally, Q-SENN uses an average of just $n_{wc} = 5$ features per class, *s. t.* redundant features are prevented and humans can follow the decision. For Grounding,

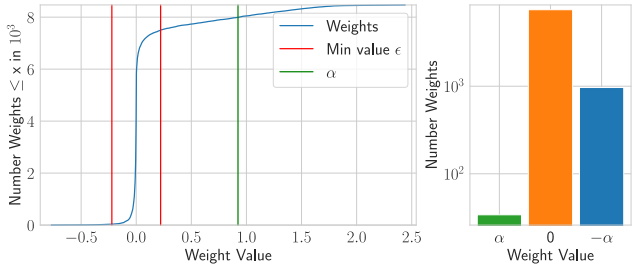


Figure 3: Exemplary result (right) of quantization on cumulative distribution (left) of nonzero weights in $\boldsymbol{W}^{\text{sp}}$ for CUB-2011($n_w = 1000$): Weights are set to 0 or $\pm\alpha$. α is the average of all remaining values above 0.

a reduced feature vector $\boldsymbol{f^*} \in \mathbb{R}^{n_f^*}$ is used with $n_f^* < n_c$ features, enforcing them to capture general attributes shared by multiple assigned classes. Additionally, we quantize the sparse weight matrix $\boldsymbol{W}^{\text{sp}} \in \mathbb{R}^{n_c \times n_f^*}$, *s. t.* it only contains either 0 or $\pm\alpha$ with $\alpha \in \mathbb{R}$. This leads to increased local and global interpretability, as the relationship between feature and class can be described as a positive, neutral, or negative assignment. It also hinders features from becoming class detectors, which hurts the interpretability as a concept. Quantized weights counteract this, as increasing the activation does not move the prediction towards one specific class. Therefore, the model is biased to learn more binary concepts, which lead to contrastive human-friendly explanations. Finally, we designed the pipeline to do N iterations of calculating the sparse matrix $\boldsymbol{W}^Q$ and fine-tuning it. The features become more general during the sparse and low-dimensional fine-tuning, as they have to recognize concepts used for multiple classes across the entire dataset. Therefore, assignments based on spurious correlations become suboptimal. The iteration leads to the removal of such assignments, *s. t.* more robust assignments remain. This leads to more grounded and robust features, thus supporting Grounding and Fidelity. To facilitate the convergence of $\boldsymbol{W}^Q$, the learning rate decays while iterating.

Quantization For quantization, we start with the regularization path computed by *glm-saga*. Figure 3 gives an overview with an exemplaric distribution of the nonzero weights in one of the sparse matrices $\boldsymbol{W}^{\text{sp}}$ before the first iteration. The very small magnitude of most weights indicates overfitting on the training set, while only the entries with higher magnitude generalize (Gale, Elsen, and Hooker 2019). To get more such entries in a sparse matrix, $\boldsymbol{W}^{\text{sp}}$ that resulted from *glm-saga* with the lowest regularization is used for quantization. As meaningful assignments between features and classes are desired, we set a threshold ϵ, below which, in terms of magnitude, all weights in $\boldsymbol{W}^{\text{sp}}$ are zeroed out. This threshold is calculated, *s. t.* exactly the desired number of weights n_w, in this paper usually $5 * n_c$, remain:

$$\epsilon = \text{sort}(|\boldsymbol{W}^{\text{sp}}|)_{n_w} \quad . \tag{3}$$

In Equation 3, $|\boldsymbol{W}^{\text{sp}}|$ represents the matrix obtained by taking the absolute value of each element in $\mathbf{W}$. The expression $\text{sort}(|\mathbf{W}|)$ describes the sorted list of absolute values of the elements in $\boldsymbol{W}^{\text{sp}}$ in descending order. Afterwards all entries whose absolute value is $\geq \epsilon$ are set to the average of the remaining values α. This leads to the quantized sparse Matrix $\boldsymbol{W}^Q \in \{-\alpha, 0, \alpha\}^{n_c \times n_f^*}$:

$$\alpha = \frac{\sum_{i,j} \mathbf{1}_{|\boldsymbol{W}^{\text{sp}}_{i,j}| \geq \epsilon} \left|\boldsymbol{W}^{\text{sp}}_{i,j}\right|}{\sum_{i,j} \mathbf{1}_{|\boldsymbol{W}^{\text{sp}}_{i,j}| \geq \epsilon}} \tag{4}$$

$$\boldsymbol{W}^Q_{i,j} = \begin{cases} \alpha & \text{if } \boldsymbol{W}^{\text{sp}}_{i,j} \geq \epsilon \\ -\alpha & \text{if } \boldsymbol{W}^{\text{sp}}_{i,j} \leq -\epsilon \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

The indicator function $\mathbf{1}_{|\boldsymbol{W}^{\text{sp}}_{i,j}| \geq \epsilon}$ takes a value of 1 if $\left|\boldsymbol{W}^{\text{sp}}_{i,j}\right| \geq \epsilon$, and 0 otherwise. Overall, the quantization scheme ensures that the resulting $\boldsymbol{W}^Q$ contains the desired number of entries which can be meaningfully averaged to get assignments, as they all initially indicated a generalizing connection between feature and class.

Experiments

This section presents the experimental results of our proposed method. The effectiveness of our approach is evaluated using Resnet50 (He et al. 2016), DenseNet121 (Huang et al. 2017), and Inception-v3 (Szegedy et al. 2016) as backbones. We used the default datasets for fine-grained image classification and prototype-based methods, as well as ImageNet-1K (Russakovsky et al. 2015) to showcase the applicability to large-scale datasets. For measuring the robustness against spurious correlations, TravelingBirds (Koh et al. 2020) is used. It is a dataset based on CUB-2011 (Wah et al. 2011), where the background is artificially spuriously correlated to the class in the training set. In the test set, the correlation is not maintained. Examples are in the supplementary material and in Figure 4. Models that maintain their accuracy on this dataset are therefore less susceptible to spurious correlations. An overview of the datasets, including CUB-2011 (in tables abbreviated C), TravelingBirds (T), Stanford Cars (Krause et al. 2013) (S), FGVC-Aircraft (Maji et al. 2013) (A) and the large-scale ImageNet-1K (I) is provided in Table 2. Notably, CUB-2011 includes both attribute and class labels, enabling us to assess the alignment of learned features with semantically meaningful concepts. Unless stated otherwise, reported values are averaged across five seeds. Implementation details and extensive results with standard deviations can be found in the supplementary material. In all tables, glm-saga$_5$ refers to applying *glm-saga* to a conventionally trained dense model with no feature selection and selecting a solution with $n_{wc} \leq 5$ and *Dense* refers to a conventionally trained model without feature selection and a densely connected final layer. For Q-SENN, the number of features per class is set to $n_{wc} = 5$ and the number of total features to $n_f^* = 50$. Table 1 gives an overview of all considered metrics. We evaluate Fidelity with accuracy on fine-grained and large-scale (I) image classification, as well as robustness to spurious correlation (T), Diversity with diversity@5 and Grounding with dependence γ and alignment r. Notably, Q-SENN maintains

Method	Accuracy ↑			diversity@5 ↑			No Concept	Alignment r ↑		Sparse	Dependence γ ↓		
	C	T	I	C	T	I	Supervision	C	T	$n_{wc} \leq 5$	C	T	I
Dense	86.7	39.4	76.1	51.6	51.5	52.0	✓	1.0	1.0	✗	10.4	22.3	3.2
glm-saga$_5$	77.7	36.0	58.0	46.7	47.7	50.0	✓	1.0	1.0	✓	49.2	52.7	53.6
CBM-joint	82.2	36.9	N/A	71.0	73.7	N/A	✗	3.0	2.8	✗	6.1	5.7	N/A
SLDD-Model	85.7	64.1	72.7	67.2	65.8	58.8	✓	1.9	1.7	✓	46.9	46.7	47.3
Q-SENN (Ours)	**85.9**	**67.3**	**74.3**	**78.1**	**75.4**	**77.2**	✓	**2.2**	**2.1**	✓	**30.0**	**30.7**	**30.7**

Table 1: Comparison across all desired metrics with Resnet50. Best results among comparable interpretable models are in bold. *N/A* indicates inapplicability due to missing annotations, while values with ✗ cannot be reasonably compared to the remaining rows due to sparsity or supervision. Accuracy, diversity@5 and Dependence γ are measured in percent. Note, that our proposed Q-SENN maintains most or all of the accuracy of dense models while simultaneously heavily improving the interpretability.

Dataset	C	A	S	T	I
# Classes n_c	200	100	196	200	1000
# Training	5 994	6 667	8 144	5 994	1 281 167
# Testing	5 774	3 333	8 041	5 774	50 000

Table 2: Number of classes, training and testing samples of the used datasets. C, A, S, T and I abbreviate CUB-2011, FGVC-Aircraft, Stanford Cars, TravelingBirds and ImageNet-1K.

more than 97% of the dense model's accuracy on ImageNet-1K, while most interpretable competitors lack the capacity or annotations to function at that scale. Overall, Q-SENN improves the baseline SLDD-Model on Fidelity, Diversity and Grounding, which are discussed in detail in the following sections.

Fidelity

Fidelity refers to the preservation of relevant information about the input in the concepts, which enables high accuracy. Table 3 shows the impact of our method on accuracy with respect to backbones in detail. On all conventional datasets, our proposed changes to the SLDD-Model maintain or improve accuracy, while increasing interpretability. Q-SENN also clearly surpasses prototype-based models, using their reduced image size of 224, shown in Table 5: The accuracy is increased with fewer features in total and per class, thus while improving interpretability. Compared to all competitors, our Q-SENN shows exceptional robustness to spurious correlations, setting a new SOTA, while maintaining 78% of the accuracy of CUB-2011 on TravelingBirds compared to the 45% of the dense model. *PIP-Net* and the SLDD-Model show that sparsely using grounded features during training, as opposed to glm-saga$_5$ or *ProtoPool*, is crucial for that robustness. Considering the ablations in Table 4, quantization seems to improve accuracy on conventional datasets, whereas the iteration leads to more robustness as spurious correlations get ignored with Q-SENN combining the strengths. We focus on Resnet50 in the remaining part of the paper as the efficacy of Q-SENN is apparent for all backbones.

Figure 4: Exemplary local explanations in comparison: Q-SENN offers explanations based on interpretable features.

Diversity

Diversity refers to the representation of inputs with non-overlapping concepts. For measuring the Diversity, we utilize diversity@5 (Norrenbrock, Rudolph, and Rosenhahn 2022) defined on $\boldsymbol{S}_5$ obtained by applying softmax to the 5 highest weighted feature maps $\boldsymbol{M}_5$ for the predicted class as:

$$\text{diversity@5} = \frac{\sum_{i=1}^{h_M} \sum_{j=1}^{w_M} \max(S_{ij}^1, S_{ij}^2, \ldots, S_{ij}^5)}{5}. \quad (6)$$

It measures the local diversity of the 5 highest weighted features. As shown in the supplementary material, a higher diversity@5 of class-independent features also indicates less correlated features. It is apparent in Table 1 that the proposed Q-SENN shows the highest diversity@5, mainly pushed by quantization as the features can not recognize entire classes and further increased through iterations, shown in Table 4. Similar observations can be made on other architectures, provided in the supplementary material. The high diversity@5 leads combined with the average number of just 5 features per class to a globally more interpretable model. This is demonstrated in Figure 1: The shown class is recognized through 5 features that all consistently localize on different semantically meaningful regions, such as *belly, crown, upper tail, upper wing and eye*, and are therefore easier to interpret and align. This makes feature alignment easier and also enables improved local explanations, as shown in Figure 4. More examples with comparison to a dense model and global explanations are shown in the supplementary material.

Method	CUB-2011			FGVC-Aircraft			Stanford Cars			TravelingBirds		
	Inc.	Dense.	Res.	Inc.	Dense.	Res.	Inc.	Dense.	Res.	Inc.	Dense.	Res.
Dense	83.5	87.2	86.7	91.4	92.5	92.0	92.6	93.6	93.5	47.9	45.1	39.4
glm-saga$_5$	76.9	69.8	77.7	89.1	86.2	89.9	88.9	81.9	89.1	45.7	42.7	36.0
CBM-joint	80.1	-	82.2	N/A	N/A	N/A	N/A	N/A	N/A	51.8	-	36.9
SLDD-Model	80.8	84.4	85.7	90.6	92.1	92.0	91.3	92.2	**92.9**	**51.9**	56.1	64.1
Q-SENN (Ours)	**81.7**	**85.4**	**85.9**	**90.8**	**92.5**	**92.1**	**91.8**	**92.8**	**92.9**	50.5	**59.6**	**67.3**

Table 3: Fidelity measured as accuracy dependent on backbone. Best result among more interpretable models are in bold. Inc., Dense., and Res. abbreviate Inception, DenseNet and Resnet.

Method	Accuracy ↑			diversity@5 ↑			Alignment r ↑		Dependence γ ↓			Binary Features ↑		
	C	S	T	C	S	T	C	T	C	S	T	C	S	T
Q-SENN (Ours)	**85.9**	92.9	67.3	**78.1**	**81.6**	**75.4**	2.2	2.1	30.0	29.3	30.7	80.4	90.0	**84.0**
w/o Quantization	85.4	92.6	**70.0**	64.4	65.3	61.8	**3.1**	**2.8**	46.8	47.7	47.3	53.2	59.2	70.8
w/o Iteration	**85.9**	**93.0**	63.9	70.9	73.5	68.9	1.7	1.5	**29.2**	**28.6**	**30.5**	**90.4**	**96.4**	78.4

Table 4: Ablation Study on Impact of iteration and quantization. Binary Features are the percentage of features for which mean shift applied on the training distribution returns exactly 2 clusters.

Method	Accuracy ↑			Total Feat.↓			Feat./ Class↓		
	C	S	T	C	S	T	C	S	T
PIP-Net	82.0	86.5	70.1	731	669	825	12	11	5.6
ProtoPool	85.5	88.9	42.9	202	195	202	202	195	202
Q-SENN	84.7	91.5	76.7	**50**	**50**	**50**	**5**	**5**	**5**
$n_f^* > 50$	**86.4**	**92.2**	**78.4**	202	195	202	**5**	**5**	**5**

Table 5: Comparison with state-of-the-art prototype-based methods with Resnet50 as backbone on the same data: With reduced sizes, Q-SENN shows increased accuracy.

Grounding

We propose two metrics to measure Grounding, the alignability with human concepts. We evaluate the correlation between the features and attributes and their generality, as we aim for features that can be aligned with a human concept rather than an entire class. To estimate this generality, the average dependence γ of the predicted class on the most important feature is measured. The dependence γ with the effect e_i of every feature for the predicted class for every sample i is calculated as:

$$e_i = \{|\boldsymbol{W}_{\hat{c},0} \cdot f_0|, |\boldsymbol{W}_{\hat{c},1} \cdot f_1|, \ldots, |\boldsymbol{W}_{\hat{c},n_f} \cdot f_{n_f}|\} \quad (7)$$

$$\gamma = \frac{1}{n_T} \sum_{i=1}^{n_T} \frac{max(e_i)}{\sum_{j=1}^{n_f} e_{ij}} \quad . \quad (8)$$

The dependence $\gamma \in [0, 1]$ measures how strongly the decision of a model usually depends on the most important feature and is designed to estimate how class-specific the features of a model are. The *Pytorch* pretrained weights for Resnet50 on ImageNet-1K validate this metric: Using *glm-saga*, *V2* shows 66% accuracy with 1.1 features per class, while *V1* shows 58% with 4.7 features. Thus, the features for *V2* are already class detectors with no conceptual meaning. This is reflected in the dependence with 44% (i. e. 44% of the prediction is on average based on one out of 2048 features) for *V2* compared to 3% for *V1*. Because we aim for grounded features, we use *V1* for our experiments.

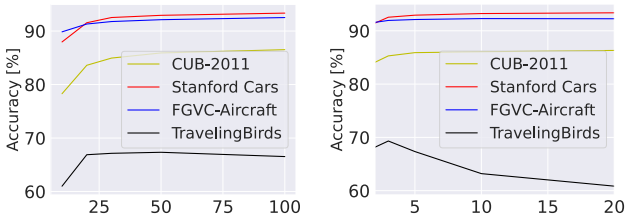
Additionally, the correlation between the attributes contained in CUB-2011, with ρ_{a+} denoting the indices where attribute a is present and ρ_{a-} the opposite, and the features on the training set $\boldsymbol{F}^{\text{train}} \in \mathbb{R}^{n_T \times n_f}$

$$A_{a,j}^{\text{gt}} = \frac{1}{|\rho_{a+}|} \sum_{i \in \rho_{a+}} F_{i,j}^{\text{train}} - \frac{1}{|\rho_{a-}|} \sum_{i \in \rho_{a-}} F_{i,j}^{\text{train}} \quad (9)$$

is measured, with the average maximum alignment r per feature as comparable metric:

$$r = \frac{1}{n_f^*} \sum_{j=1}^{n_f^*} \frac{n_T}{\sum_{l=1}^{n_T} F_{l,j}^{\text{train}} - \min_l F_{l,j}^{\text{train}}} \max_i A_{i,j}^{\text{gt}} \quad . \quad (10)$$

A higher value of r indicates features that are more aligned with the attributes in CUB-2011. We norm the difference by the average activation to be less dependent of assumptions about the underlying distribution. Because the models shown are not trained to focus on the given attributes, it can be assumed that models with higher r generally learn features more correlated to human concepts. Table 1 shows improvements in alignment r and dependence γ across datasets. Note that a dense matrix often leads to a low dependence γ at the cost of global interpretability and CBM was trained to predict a subset of the probed attributes. The impact of both proposed changes is apparent in Table 4: The iteration increases the alignment with given attributes, as the features do not have to adapt to a sparse structure based on spurious correlations. The problem of high dependence γ however, is reduced through the quantization. In summary, only the proposed Q-SENN uses more aligned, less class-specific features. We additionally included the fraction of binary-like features.



(a) Number of Features n_f^* (b) Features per class n_{wc}

Figure 5: Relationship between Accuracy and interpretability-related parameters for Q-SENN with Resnet50.

A feature is considered as binary if the mean shift clustering, using a GPU-accelerated implementation by (Schier, Reinders, and Rosenhahn 2022), applied to the feature distribution on the training set returns exactly 2 clusters. As expected, the quantization leads to more binary features. For reference, a Dense Resnet50 on CUB-2011 has 15% binary features and CBM 67%. The clustering is visualized and analyzed in the supplementary material. The analysis shows that the few binary features of the Dense model, and only its features, are class-detectors instead of general features. For Resnet50, the Dense model has effectively no feature, one across 5 seeds, whose active cluster, the one with a higher mean, has more samples than the most frequently occuring training class. In contrast, every binary feature of our Q-SENN has this property, e.g. 80.4% of all features on CUB-2011. Notably, even the CBM, trained to predict binary attributes, has a few features that do not have this property.

Interpretability Tradeoff

This section analyzes the impact of changing n_f^* and n_{wc} for our Q-SENN. The result is visualized in Figures 5a and 5b: More features n_f^* and to a lesser degree more weights per class n_{wc} lead to increased accuracy. However, for TravelingBirds there is no tradeoff. In fact, more interpretability, especially through sparsity, leads to a higher accuracy because the model does not have the capacity to learn all the spurious correlations but has to focus on the more general features. Further ablations regarding the convergence of $\boldsymbol{W}^Q$, computational cost, type of binary feature and impact of quantization levels are shown in the supplementary material.

Alignment Without Annotations

In this section, we propose an alternative way of aligning the learned features of the Q-SENN with human concepts, when no additional annotations are available. Note that it is easier for Q-SENN, as its features show Grounding, thus are more alignable with any human concept. CLIP (Radford et al. 2021) is utilized as it can compute the similarity between a given text prompt and an image due to its contrastive training. The method is first described and then validated by demonstrating high agreement between alignments computed from additional labels provided in CUB-2011 and via our method. Finally, we show that it even indicates certainty.

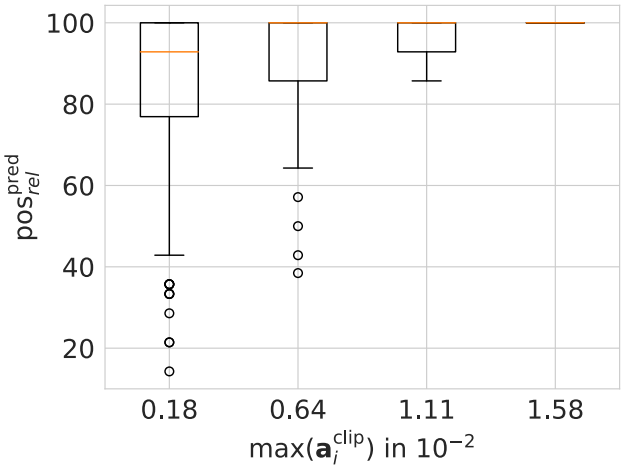


Figure 6: Relationship between the alignment value $\max(\mathbf{a}_i^{\text{clip}})$ and $\text{pos}_{rel}^{\text{pred}}$ for the proposed Q-SENN for all features over 5 seeds: With increasing $\max(\mathbf{a}_i^{\text{clip}})$, the predicted alignment for the feature i is more similar to the ground truth.

CLIP

In this work, we utilize CLIP (Radford et al. 2021), which is a multimodal neural network designed to map images and captions into a shared feature space. The model is trained using a contrastive loss function, which encourages high cosine similarity between the embedding of an image $\boldsymbol{I}_{\text{clip}} \in \mathbb{R}^{n_f^{\text{clip}}}$ and its corresponding caption $\boldsymbol{T}_{\text{clip}} \in \mathbb{R}^{n_f^{\text{clip}}}$, while minimizing the cosine similarity $\boldsymbol{I}_{\text{clip}} \boldsymbol{T}_{\text{clip}}$ between that image and other captions in the training set. This approach enables the model to learn joint representations that capture meaningful relationships between images and their associated textual descriptions, facilitating downstream tasks such as image captioning and retrieval. We used the ViT-B32 (Dosovitskiy et al. 2021) model of CLIP in our experiments.

Feature Alignment

For alignment without annotations, we use the features calculated on the training set $\boldsymbol{F}^{\text{train}} \in \mathbb{R}^{n_T \times n_f^*}$ and a set of n_p possible relevant concepts a for the model such as the attributes contained in CUB-2011. The proposed method can identify which of the prompts might be related to the activation of a specific feature. First, the entire training set is embedded using CLIP to obtain $\mathbf{D}_{\text{clip}} \in \mathbb{R}^{n_T \times n_f^{\text{clip}}}$. Afterwards, the possibly relevant attributes are converted into prompts, *e. g.*, "This is a photo of a red wing of a bird" for the attribute "red wing" of type "wing color". The conversion is described in detail in the supplementary material. These prompts are then encoded to $\mathbf{P}_{\text{clip}} \in \mathbb{R}^{n_p \times n_f^{\text{clip}}}$. The multiplication

$$\mathbf{S} = \mathbf{D}_{\text{clip}} \mathbf{P}_{\text{clip}}^T \tag{11}$$

leads to the similarity matrix $\mathbf{S} \in \mathbb{R}^{n_T \times n_p}$ between images and prompts. In order to relate the image similarities to the learned features, we use the feature values to weight the

Alignment Method	$\text{pos}^{\text{full}}\downarrow$	$\text{pos}^{\text{pred}}\downarrow$	$\text{pos}^{\text{pred}}_{rel}\uparrow$
Random	156.5	7.3	50.0%
Static w/ min. pos^{full}	122.9	8.0	41.7%
Proposed Method	**43.1**	**2.7**	**86.2%**

Table 6: Average validation metrics when aligning features of Q-SENN compared to baselines: Our proposed alignment method handily beats the optimal static baseline based on the ground truth. Rather than just identifying relevant concepts, our proposed method matches it to the correct feature.

image similarity according to the activation and obtain one weighting factor $v_{i,j}$

$$v_{ij} = \frac{\boldsymbol{F}^{\text{train}}_{i,j} - mean(\boldsymbol{F}^{\text{train}}_{:,j})}{std(\boldsymbol{F}^{\text{train}}_{:,j})} \tag{12}$$

for image i and feature j. With this normalized weighting $\mathbf{V}$ the alignment matrix $\mathbf{A}^{\text{clip}} \in \mathbb{R}^{n^*_f \times n_p}$ is obtained:

$$\mathbf{A}^{\text{clip}} = \mathbf{V}^T\mathbf{S} \quad . \tag{13}$$

The matrix can be used similarly to $\mathbf{A}^{\text{gt}}$ as it measures how much the perceived similarity to a specific prompt of CLIP varies along the feature dimension.

Note that the proposed method is generally independent of the used CLIP-like model, and could therefore profit of models that are able to caption more fine-grained subregions.

Validation

For validation, we test how well our method can identify the most related concept from a given set of concepts. As ground truth, the alignment matrix $\mathbf{A}^{\text{gt}}$ computed with Equation 9 is used. To compare whether our method can *order* the attributes in the same way, we use all attributes contained in CUB-2011 as set of possible attributes and compute $\mathbf{A}^{\text{clip}}$. For each feature i $\text{pos}^{\text{full}}_i$ then measures the position at which the most aligned attribute a in $\mathbf{a}^{\text{clip}}_i$ is ranked in $\mathbf{a}^{\text{gt}}_i$:

$$\text{pos}^{\text{full}}_i = \{j : \text{argmax}(\mathbf{a}^{\text{clip}}_i) = \text{argsort}(\mathbf{a}^{\text{gt}}_i)_j\} \quad . \tag{14}$$

Here, argmax returns the index with the highest value and $\text{argsort}(\mathbf{a}^{\text{gt}}_i)$ returns the indices that would sort $\mathbf{a}^{\text{gt}}_i$ in descending order. We additionally calculate the above metric for the predicted type by limiting $\mathbf{a}^{\text{gt}}$ and $\mathbf{a}^{\text{clip}}$ to its attributes. It focuses on the ability to identify which expression of a given type is more related to a feature, while pos^{full} measures if the method can identify type and expression, which can be more affected by biases of CLIP. To be less dependent on the number of attributes of the predicted type n_{type}, we also report the relative position

$$\text{pos}^{\text{pred}}_{rel,i} = 1 - \frac{\text{pos}^{\text{pred}}_i - 1}{n_{type} - 1} = \frac{n_{type} - \text{pos}^{\text{pred}}_i}{n_{type} - 1}, \tag{15}$$

which ranges from the worst case at 0% to optimality at 100%. The results in Table 6 validate the proposed method and show the superiority to any static baseline.

Optimal Static Baseline To compute the optimal static sorting of attributes for comparison with our alignment method, we calculated $pos_{i,j} = \text{argsort}(\mathbf{a}^{\text{gt}}_i)_j$ for all features i and attributes j and sorted the attributes, *s. t.* the average position

$$avg_j = \frac{1}{n^*_f}\sum_{i=0}^{n^*_f} pos_{i,j} \tag{16}$$

of the attribute ascends. This returns the optimal static list which minimizes pos^{full}.

Analysis of Alignment Value

Figure 6 shows the relationship between the alignment value $\max(\mathbf{a}^{\text{clip}}_i)$ and $\text{pos}^{\text{pred}}_{rel}$ for the proposed Q-SENN for all features over 5 seeds. The box plots are created by assigning every feature to one of the 4 evenly sized bins according to their alignment value $\max(\mathbf{a}^{\text{clip}}_i)$. The other metrics show similar trends. The figure demonstrates that our method conveys certainty, as $\text{pos}^{\text{pred}}_{rel}$ increases with $\max(\mathbf{a}^{\text{clip}}_i)$. Since the features do not need to be aligned with any of the given attributes, Figure 6 suggests that the features with higher uncertainty do not directly correspond to any of the given attributes, but rather concepts not annotated in CUB-2011. Therefore, the proposed method with no required additional annotation is more accurate for features aligned with the given attributes than the reported average in Table 6 suggests, with the median absolute value of pos^{pred} for the top 3 bins being the first position, shown in the supplementary material.

Future Work

Q-SENN follows the popular SENN (Alvarez Melis and Jaakkola 2018) framework and delivers concise, contrastive and general, therefore human-friendly (Miller 2019), explanations, as shown in Figure 1. However, human-subject experiments are required to actually measure the interpretability for humans (Doshi-Velez and Kim 2017). As Q-SENN is generally applicable to classification tasks, it can also bring more interpretability to other tasks like semantic segmentation, where interpretability is discussed (Kaiser, Reinders, and Rosenhahn 2023).

Conclusion

We propose the *Quantized-Self-Explaining Neural Network* Q-SENN with iteratively optimized ternary feature-class assignments. The experiments support the classification as SENN by evaluating how the concepts learned without additional supervision satisfy the three desiderata Fidelity, Diversity and Grounding. In isolation, quantization primarily benefits Fidelity, dependence γ and Diversity, while just iterating increases feature alignment r and robustness to spurious correlations. Q-SENN combines and amplifies the individual strengths, surpassing competitors in various facets. Finally, we propose a method for aligning the learned concepts with human ones with no need for additional supervision using CLIP, enabling Q-SENN to be applied in various scenarios with need for interpretability. Future work might even incorporate it to guide Q-SENN towards more grounded concepts.

Acknowledgements

This work was supported by the Federal Ministry of Education and Research (BMBF), Germany under the AI service center KISSKI (grant no. 01IS22093C) and the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122). This work was partially supported by Intel Corporation and by the German Federal Ministry of the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (GreenAutoML4FAS project no. 67KI32007A).

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Alvarez Melis, D.; and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.

Bibal, A.; Lognoul, M.; Streel, A.; and Frénay, B. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29.

Böhle, M.; Fritz, M.; and Schiele, B. 2022. B-cos Networks: Alignment is All We Need for Interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10329–10338.

Böhle, M.; Fritz, M.; and Schiele, B. 2023. Holistically Explainable Vision Transformers. *arXiv preprint arXiv:2301.08669*.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fel, T.; Picard, A.; Béthune, L.; Boissin, T.; Vigouroux, D.; Colin, J.; Cadène, R.; and Serre, T. 2023. CRAFT: Concept Recursive Activation FacTorization for Explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2711–2721.

Gale, T.; Elsen, E.; and Hooker, S. 2019. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*.

Glandorf, P.; Kaiser, T.; and Rosenhahn, B. 2023. HyperSparse Neural Networks: Shifting Exploration to Exploitation through Adaptive Regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1234–1243.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hoffmann, A.; Fanconi, C.; Rade, R.; and Kohler, J. 2021. This Looks Like That... Does it? Shortcomings of Latent Space Prototype Interpretability in Deep Networks.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Kaiser, T.; Reinders, C.; and Rosenhahn, B. 2023. Compensation Learning in Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3266–3277.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Kim, S. S.; Meister, N.; Ramaswamy, V. V.; Fong, R.; and Russakovsky, O. 2022. Hive: evaluating the human interpretability of visual explanations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, 280–298. Springer.

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2019. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 267–280. Springer.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, 5338–5348. PMLR.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia.

Lipton, P. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27: 247–266.

Liu, B.; Li, F.; Wang, X.; Zhang, B.; and Yan, J. 2023. Ternary weight networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Maji, S.; Kannala, J.; Rahtu, E.; Blaschko, M.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. Technical report.

Marconato, E.; Passerini, A.; and Teso, S. 2022. GlanceNets: Interpretable, Leak-proof Concept-based Models. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Margeloiu, A.; Ashman, M.; Bhatt, U.; Chen, Y.; Jamnik, M.; and Weller, A. 2021. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*.

McGrath, T.; Kapishnikov, A.; Tomašev, N.; Pearce, A.; Wattenberg, M.; Hassabis, D.; Kim, B.; Paquet, U.; and Kramnik, V. 2022. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47): e2206625119.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2): 81.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

Nauta, M.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2023. PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification.

Nauta, M.; van Bree, R.; and Seifert, C. 2021. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14933–14943.

Norrenbrock, T.; Rudolph, M.; and Rosenhahn, B. 2022. Take 5: Interpretable Image Classification with a Handful of Features. In *Progress and Challenges in Building Trustworthy Embodied AI*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramaswamy, V. V.; Kim, S. S. Y.; Meister, N.; Fong, R.; and Russakovsky, O. 2022. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features. *arXiv*. Decomposition in unlabelled / labelled features. Sparse explanation, not prediction as it uses ground truth attribute labels for explanation. Does not have to be faithful with actually done prediction.

Rosenhahn, B. 2023. Optimization of Sparsity-Constrained Neural Networks as a Mixed Integer Linear Program. *Journal of Optimization Theory and Applications*, 199(3): 931–954. (open access).

Rüping, S.; et al. 2006. Learning interpretable models.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Rymarczyk, D.; Struski, Ł.; Górszczak, M.; Lewandowska, K.; Tabor, J.; and Zieliński, B. 2022. Interpretable image classification with differentiable prototypes assignment. In *European Conference on Computer Vision*, 351–368. Springer.

Rymarczyk, D.; Struski, Ł.; Tabor, J.; and Zieliński, B. 2021. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1420–1430.

Sawada, Y.; and Nakamura, K. 2022. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10: 41758–41765.

Schier, M.; Reinders, C.; and Rosenhahn, B. 2022. Constrained Mean Shift Clustering. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Slack, D.; Friedler, S. A.; Scheidegger, C.; and Roy, C. D. 2019. Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501*.

Stalder, S.; Perraudin, N.; Achanta, R.; Perez-Cruz, F.; and Volpi, M. 2022. What You See is What You Classify: Black Box Attributions. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wong, E.; Santurkar, S.; and Madry, A. 2021. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning*, 11205–11216. PMLR.

Yang, H.; Rudin, C.; and Seltzer, M. 2017. Scalable Bayesian rule lists. In *International conference on machine learning*, 3921–3930. PMLR.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2022. Post-hoc Concept Bottleneck Models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*.

Zarlenga, M. E.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Precioso, F.; Melacci, S.; Weller, A.; Lio, P.; et al. 2022. Concept Embedding Models. In *NeurIPS 2022-36th Conference on Neural Information Processing Systems*.

Zhang, Q.; Cao, R.; Shi, F.; Wu, Y. N.; and Zhu, S.-C. 2018. Interpreting CNN knowledge via an explanatory graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhang, Q.; Wu, Y. N.; and Zhu, S.-C. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8827–8836.