Understanding Multimodal Deep Neural Networks: A Concept Selection View

Chenming Shang (scm22@mails.tsinghua.edu.cn)

Hengyuan Zhang (zhang-hy22@mails.tsinghua.edu.cn)

Hao Wen (wenh22@mails.tsinghua.edu.cn)

Yujiu Yang (yang.yujiu@sz.tsinghua.edu.cn)

Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

Abstract

The multimodal deep neural networks, represented by CLIP, have generated rich downstream applications owing to their excellent performance, thus making understanding the decision-making process of CLIP an essential research topic. Due to the complex structure and the massive pre-training data, it is often regarded as a black-box model that is too difficult to understand and interpret. Concept-based models map the black-box visual representations extracted by deep neural networks onto a set of human-understandable concepts and use the concepts to make predictions, enhancing the transparency of the decision-making process. However, these methods involve the datasets labeled with fine-grained attributes by expert knowledge, which incur high costs and introduce excessive human prior knowledge and bias. In this paper, we observe the long-tail distribution of concepts, based on which we propose a two-stage Concept Selection Model (CSM) to mine core concepts without introducing any human priors. The concept greedy rough selection algorithm is applied to extract head concepts, and then the concept mask fine selection method performs the extraction of core concepts. Experiments show that our approach achieves comparable performance to end-to-end black-box models, and human evaluation demonstrates that the concepts discovered by our method are interpretable and comprehensible for humans.

Keywords: model interpretability; concept-based model; multimodal pre-trained model; model debugging; concept mining

Introduction

Deep neural networks (DNNs) (LeCun, Bengio, & Hinton, 2015) have achieved unprecedented success in a wide range of machine learning tasks, including computer vision (K. Han et al., 2022), natural language processing (Guo et al., 2024), and speech recognition (Radford et al., 2023). However, due to their complex and deep structures, they are often regarded as black-box models (Castelvecchi, 2016), which are too difficult to understand and interpret. In fields that demand high levels of trustworthiness, such as medicine (Huang, Shang, Xiong, Pang, & Jin, 2021), education (Zhang, Liu, Shang, Li, & Jiang, 2024), and finance (Ozbayoglu, Gudelek, & Sezer, 2020), how humans understand the DNNs has become increasingly crucial. It indicates whether we can trust the DNNs' decisions, and how we can rectify the errors when the DNNs make mistakes. Making deep learning models more interpretable is a significant yet challenging research topic.

Recently, there has been rapid development in multimodal pre-trained models (X. Han et al., 2023), among which Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) has achieved remarkable progress by employing contrastive learning to enable the shared representation space for vision and text. Specifically, CLIP consists of a text encoder and an image encoder, to extract textual and visual features, respectively. It leverages natural language as supervision for images and establishes the correspondence between images and text by maximizing the similarity between an image and its corresponding text description while minimizing the similarity between text descriptions from different images. Through such cross-modal learning and extensive training with large-scale data, CLIP can realize a range of interesting applications, for instance, serving as a backbone for image representation extraction which then be directly utilized for classification tasks (Kumar, Raghunathan, Jones, Ma, & Liang, 2022), highlighting the importance of understanding the underlying mechanism and capability of CLIP.

DNNs can be interpreted at various levels, including pixels (Chattopadhay, Sarkar, Howlader, & Balasubramanian, 2018), samples (Hammoudeh & Lowd, 2022), weights (Wortsman et al., 2020), individual neurons (Ghorbani & Zou, 2020), subnetworks (Amer & Maul, 2019) and representations (Hendricks, Hu, Darrell, & Akata, 2018). However, with the increasing complexity of model structures, deeper layers, diverse data formats, and large-scale datasets, traditional model interpretability methods have become challenging to apply and generate human-understandable explanations. In this paper, we aim to address such a challenge: how to present the reasoning process of CLIP in a more intuitive manner and allow humans to intervene in the model's results?

A promising approach for achieving interpretability in deep learning is through concept-based models (CBMs) (Schwalbe, 2022; Poeta, Ciravegna, Pastor, Cerquitelli, & Baralis, 2023), which map the visual representations to a set of human-generated high-level concepts to explain the black-box features of DNNs (Koh et al., 2020). These interpretable concepts are then used to make the final decision by a linear function, greatly enhancing our understanding of the decision-making process. Due to the considerable cost of the fine-grained and precise annotation for each concept in CBMs, multimodal pre-trained model-based CBMs (Yuksekgonul, Wang, & Zou, 2022; Oikarinen, Das, Nguyen, & Weng, 2023; Shang et al., 2024) have recently emerged as a research hotspot. However, these recent innovations involve humans predefining a set of complex concepts specific to particular categories, such as *small, black insect with six legs* (Yang et al., 2023), which introduces excessive human bi-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

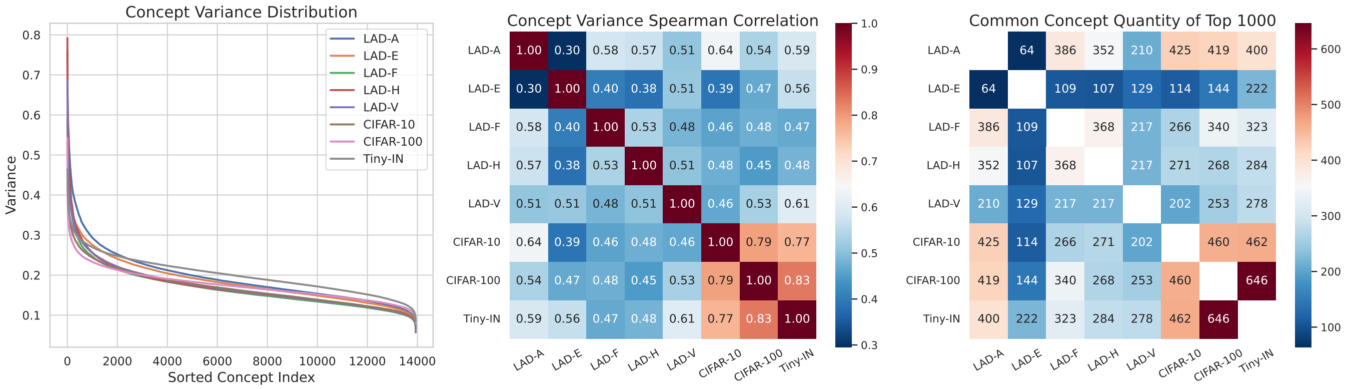

Figure 1: *Left*: the distribution of the sorted concept variances. *Middle*: the Spearman correlation coefficients of concept variances between any two datasets. *Right*: the number of concepts shared in the top 1000 concepts with the highest variances between any two datasets.

ases. It resembles generating a set of descriptions for images belonging to specific categories rather than truly understanding and explaining the decision-making process of CLIP.

To address the aforementioned challenges, we propose a two-stage Concept Selection Model (CSM) to understand concepts emerging from CLIP without introducing any human priors. Initially, we establish a concept library that CLIP can comprehend. We employ a powerful CLIP as a concept annotator to label the dataset with concepts, during which we observe the long-tail distribution of concepts. Subsequently, the concept greedy rough selection algorithm is applied to extract head concepts, and then the concept mask fine selection method performs the extraction of core concepts for specific classification tasks. Eventually, we conduct experiments using the filtered core concepts, which achieve comparable performance to end-to-end black-box models on multiple datasets. In addition, human evaluation demonstrates that the concepts discovered by our method are interpretable and comprehensible for humans.

Setup and Observation

Image Dataset and Concept Library

We use the image classification task to understand the role of different concepts in the decision-making process of CLIP. For this purpose, we have chosen 2 common object datasets: CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton, et al., 2009), as well as a fine-grained object dataset: Large-scale Attribute Dataset (LAD) (Zhao et al., 2019), including 5 sub-datasets: LAD-Animal, LAD-Electronic, LAD-Food, LAD-Hair, and LAD-Vehicle. Additionally, the Tiny-ImageNet (Tiny-IN) dataset (Le & Yang, 2015) is also employed as a reference to measure concept distribution.

In order to enable the CLIP to automatically select and determine which concepts play the most important role in the decision-making process, thereby understanding the model's internal mechanisms instead of relying on pre-defined concepts by humans, which involve excessive human prior knowledge, we need to establish a concept library that satisfies the following properties: (*i*).**Comprehensiveness**: the concept library is expected to encompass commonly used concepts in the open world, capturing a broad range of information and knowledge. (*ii*).**Atomicity**: each concept in the concept library should be atomic rather than composite, which is divided into the most basic semantic units. We find that the scene graph consisting of visual concepts is an appropriate foundation. The scene graph derives from the Visual Genome dataset (Krishna et al., 2017) containing 100K images with corresponding text descriptions that provides a highly representative depiction of the real world, as shown in Fig.2-Up. Meanwhile, in the scene graph, an atom is defined as an individual visual concept, corresponding to a single scene graph node, which ensures the principle of minimal semantic units. Atoms are further subtyped into objects, relationships, and attributes. We pick the nouns and adjectives in them as the concept library with a quantity of 13,933.

CLIP based Concept Annotator

Consider a dataset of image-label pairs $\mathcal{D} = \{(x,y)\}$, where $x \in \mathcal{X}$ is the image and $y \in \mathcal{Y}$ is the label. We have N concepts to describe the essential information of the world, which can be denoted as discrete tokens $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$. Multi-modal pre-trained alignment model (e.g., CLIP) has an image encoder $\Phi_I : \mathcal{X} \rightarrow \mathbb{R}^d$ and a text encoder Φ_T, which can map images and text into a shared d-dimensional feature space respectively. We encode the discrete tokens with the CLIP text encoder then perform L_2 normalization[1] to obtain the concept embeddings $\{w_1, w_2, \ldots, w_N | w_i = \Phi_T(e_i), i = 1, 2, \ldots, N\}$ with the length of 1 and the dimension of d. We concatenate these concept embeddings to a concept projection matrix $W_{N\times d} : \mathbb{R}^d \rightarrow \mathbb{R}^N$ in arbitrary order, also identified as a concept library.

Correspondingly, we utilize the CLIP image encoder to get the visual representations $f = \Phi_I(x)$. Considering that

[1] The symbol of L_2 normalization is omitted for convenience.

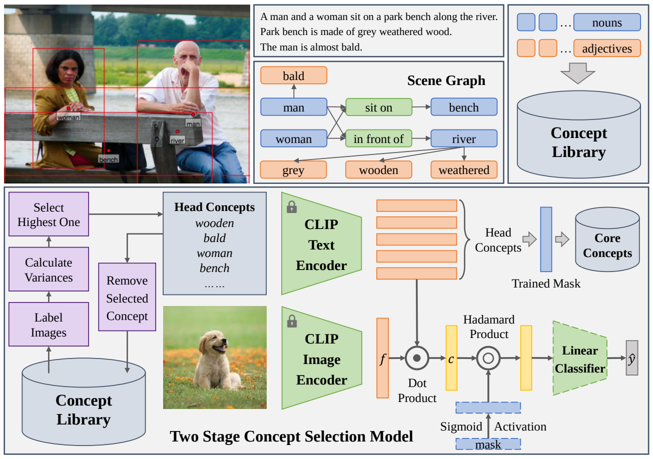

Figure 2: *Up*: In the Visual Genome dataset, each image is accompanied by corresponding textual descriptions, which are transformed into scene graphs, with each word represented as an atomic node. *Down*: Two-stage concept selection model: a rough selection is utilized to obtain the head concepts from the concept library, and subsequently a fine selection is applied to identify the core concepts from the head concepts.

CLIP has aligned the images with the textual data when pretraining, the visual representations share the feature space with any concept embedding, and the projection length $\|f\|_2 \cdot \cos\langle w_i, f\rangle$ can reflect the presence of a particular concept in the image. Since undergoing L_2 normalization on w_i, the concept existence can be directly formulated in terms of the dot product $c = W \cdot f$. Thus far, we have utilized CLIP to annotate arbitrary concepts on any given image.

Concept Distribution

We utilize CLIP-ViT-L/14 to annotate 13,933 concepts on the training set images of 8 datasets. For each concept, we calculate its variance across different samples within the same dataset. A higher variance indicates a more notable difference in the presence of that concept across different images, suggesting that the concept may play a greater role in the task. We sort the concept variances and visualize the distribution of variances for each concept after ranking, as shown in Fig.1-Left. We discover that different datasets exhibit remarkably similar distributions, all demonstrating a long-tail effect. Specifically, among the 13,933 concepts, only approximately 1,000 concepts have variances above 0.3 and 4,000 concepts have variances above 0.2, which implies that despite the numerous concepts available, the model relies on only a small fraction of them to make decisions.

We then analyze the fluctuations of these concepts across different datasets. In Fig.1-Middle, we present the Spearman correlation coefficients of concept variances between any two datasets. We observe that all coefficients are positive, and the correlation coefficients between common object datasets are consistently above 0.77, indicating a strong positive correlation. This suggests that the model tends to depend on the same concepts when dealing with general image classification tasks. In addition, we find that the correlation coefficients between fine-grained object datasets are relatively smaller, which indicates that the model focuses on different concepts. The correlation coefficient reflects the similarity between datasets to a certain extent, for example, the coefficient between the LAD-Animal and LAD-Electronic is only 0.3, showing a lower correlation, which is consistent with the significant differences in images between these two datasets.

Furthermore, we present the number of concepts shared in the top 1000 concepts with the highest variances between any two datasets, reflecting the fluctuations of the head concepts, as shown in Fig.1-Right. The head concepts exhibit a similar trend to the overall concepts, with more concepts shared among common object datasets and fewer shared among fine-grained object datasets. As the differences between datasets increase, the number of shared concepts decreases.

Method: Concept Selection Model

From the observation, we conclude that only a few concepts produce major changes when reasoning in CLIP. Therefore, in this section, we propose a two-stage Concept Selection Model (CSM) to filter out the head concepts and further identify the core concepts.

Greedy Rough Selection

Selecting the optimal core concepts from a vast concept library is a combinatorial optimization problem, which becomes highly demanding when coupled with the optimization problem of linear classifiers. Therefore, based on the aforementioned observation, we adopt a greedy strategy to select the head concepts from the concept library in advance, narrowing down the scope of the concept library to avoid getting trapped in the local optima in the subsequent fine selection.

Specifically, for a given dataset, we calculate the variance of the activation value of each CLIP-annotated concept across all images, which reflects the differences in a concept's presence, and we append the concept with the highest variance to the head concepts. An important consideration is that con-

Algorithm 1 Greedy Rough Concept Selection

```
Input: $K$ images visual representations $V \in \mathbb{R}^{K\times d}$, $N$ concepts concept embeddings $W \in \mathbb{R}^{N\times d}$ identified as the concept library, head concept size $M$.
Output: head concept embeddings $W_{\text{head}} \in \mathbb{R}^{M\times d}$.
Initialization: $W_{\text{head}} \leftarrow \emptyset$
for $i \in [1,\ldots,M]$ do
    $C \leftarrow W \cdot V^T$
    $t^* \leftarrow \arg\max_{t=1}^{N} \text{Var}(C[t])$
    $W_{\text{head}} \leftarrow W_{\text{head}} \cup \{W[t^*]\}$
    for $k \in [1,\ldots,K]$ do
        $V[k] \leftarrow V[k] - \cos\langle V[k], W[t^*]\rangle \cdot W[t^*]$
    end for
end for
```

Table 1: Comparison of different concept selection methods. Bold indicates the best and underline indicates the 2nd-best result.

Method	CIFAR-10	CIFAR-100	LAD-A	LAD-E	LAD-F	LAD-H	LAD-V	**Mean**
Linear Probing	0.9805	0.8703	0.9823	0.9397	0.8823	0.5176	0.9302	0.8718
Random	0.7212	0.8323	0.8872	0.8269	0.7074	$\underline{0.2482}$	0.8102	0.7191
Human Prior	N/A	N/A	0.9659	0.9049	0.7749	0.2145	0.8934	N/A
ConceptNet	$\underline{0.8704}$	$\underline{0.8418}$	$\underline{0.9744}$	$\underline{0.9190}$	$\underline{0.8130}$	0.1269	**0.9172**	$\underline{0.7804}$
CSM	**0.9708**	**0.8510**	**0.9767**	**0.9270**	**0.8516**	**0.4080**	$\underline{0.9129}$	**0.8426**

cepts with similar lexical meanings are likely to generate high variances, for example, *dog* and *puppy* could lead to concept redundancy, which we want to avoid. Therefore, after identifying the concept with the highest variance in each iteration, we need to eliminate its activation from all images to prevent the repetition of synonymous concepts. The details are depicted in Algorithm 1, where we pick the top $M = 1000$ head concepts. Note that here we only utilize the image information from the training set without their labels, so the rough selection is weakly correlated with the specific downstream classification task and strongly related to the data itself.

Mask Fine Selection

After obtaining the head concepts, we proceed with the fine selection of core concepts. Specifically, we establish a learnable mask m, the size of which is equal to the number of head concepts, indicating the importance weights. Each weight in this mask is activated through a sigmoid function $\sigma(\cdot)$, mapping it to a value between 0 and 1, representing the significance of the respective concept, and values closer to 1 signifying higher importance. We perform a Hadamard product ($\odot$) of this importance weight with the concept activations generated by the head concepts, resulting in a weighted concept bottleneck. We then connect a linear classifier $\Psi_m : \mathbb{R}^M \rightarrow \mathcal{Y}$ for classification after this bottleneck, as illustrated in Fig.2-Down, where we solve the following optimization problem:

$$\hat{y} = \Psi_m\left(\sigma(m) \odot c_{\text{head}}\right) = \Psi_m\Big(\sigma(m) \odot \big(W_{\text{head}} \cdot \Phi_I(x)\big)\Big) \quad (1)$$

$$\min_{\Psi, m} \underset{(x,y)\in\mathcal{D}}{\mathbb{E}} \Big[\mathcal{L}\big(\hat{y}, y\big)\Big] + \lambda \cdot \Omega(\Psi_m) \quad (2)$$

where $\mathcal{L}(\hat{y}, y)$ is the cross-entropy loss function, $\Omega(\Psi)$ is a complexity measure, and λ is the regularization strength.

We obtain the trained mask and sort it by magnitude, then select the top N^* concepts as the core concepts W_{core}. At next stage, we only use these core concepts to retrain the concept-based model by solving the following optimization problem:

$$\min_{\Psi} \underset{(x,y)\in\mathcal{D}}{\mathbb{E}} \bigg[\mathcal{L}\Big(\Psi\big(W_{\text{core}} \cdot \Phi_I(x)\big), y\Big)\bigg] + \lambda \cdot \Omega(\Psi) \quad (3)$$

where $\Psi : \mathbb{R}^{N^*} \rightarrow \mathcal{Y}$ is a linear classifier. By following the steps, we have systematically filtered and obtained the top concepts as head concepts and the core concepts from the concept library. As a result, we have developed a concept-based model capable of performing classification tasks utilizing the core concepts.

Experimental Results

In the comprehensive study of our approach (CSM) from the perspectives of **Accuracy (Q1-A1)** and **Interpretability (Q2-A2)**, we aim to answer the following questions:

- **Q1.1**: What is an appropriate quantity of core concepts?
- **Q1.2**: Does CSM offer advantages compared to other concept selection methods?
- **Q1.3**: For what tasks do concept based models outperform black-box models?
- **Q2.1**: Which concepts are selected as core concepts?
- **Q2.2**: How do people understand and intervene in the model's reasoning and decision-making process?
- **Q2.3**: Does the concept based model actually function as we understand it in practice?

Accuracy

Concept Quantity (A1.1). An intuitive judgment is that the number of concepts should be positively correlated with the number of categories. More categories require more concepts to support differentiation. Therefore, we conduct experiments on CIFAR-10 and CIFAR-100 according to the average number of concepts possessed by each category, as shown in Fig.3. Concept based model performs poorly when only a small number of concepts are available at the initial stage. As the number of concepts increases, the model's classification accuracy steadily improves, with a particularly noticeable improvement in the early stages when the concept quantity is small. When the number of concepts increased from approximately 0.1 concepts per class to 0.2 concepts per class, the

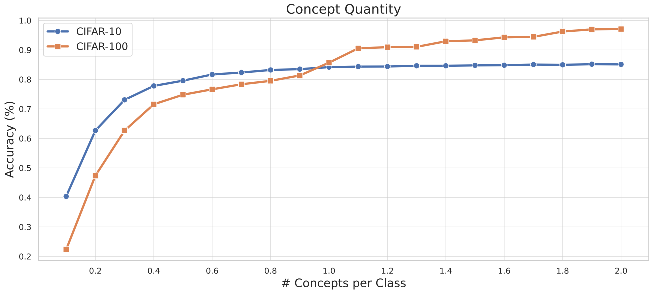

Figure 3: The variation in accuracy of the CSM as the concept quantity increases on CIFAR-10 and CIFAR-100.

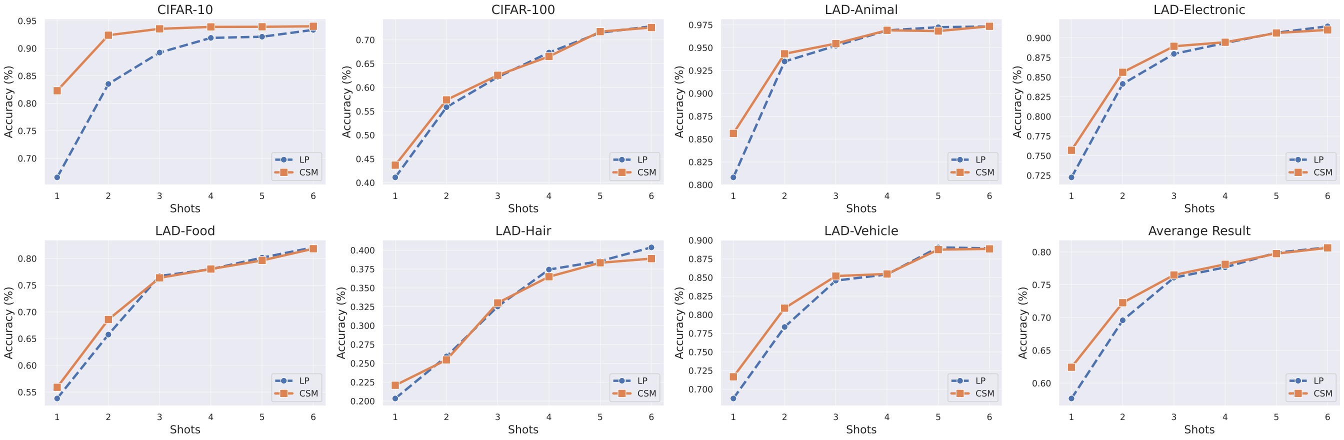

Figure 4: Accuracy comparison between CSM and linear probing. The x-axis indicates the number of labeled images per class.

Table 2: Top 5 core concepts with the highest activation variance for different datasets.

CIFAR-10	CIFAR-100	LAD-Animal	LAD-Electronic	LAD-Food	LAD-Hair	LAD-Vehicle
fishing boat	*sundress*	*shepherd dog*	*range hood*	*sandwich*	*pile*	*streetcar*
horseback	*wildlife*	*pelican*	*lens cap*	*bacon*	*braid*	*parasailing*
frog	*streetcar*	*leopard*	*refrigerator*	*lettuce*	*ponytail*	*cart*
airplane	*shepherd dog*	*American bison*	*monitor*	*tomato*	*hijab*	*minibike*
yacht	*tree*	*mermaid*	*earphone*	*walnut*	*barber*	*seaplane*

accuracy on both CIFAR-10 and CIFAR-100 improved by more than 20%. For CIFAR-10, the improvement becomes less pronounced at about 0.6 concepts per class for CIFAR-10 and at about 1.1 concepts per class for CIFAR-100. The model's accuracy approaches convergence at around 2 concepts per class. Therefore, in subsequent experiments, we choose the concept quantity as the category quantity times 2.

Concept Selection Methods (A1.2). We choose 3 other concept selection strategies as comparisons: (*i*).**Random**: without distinguishing between good and bad concepts, an equal number of concepts to the CSM are selected directly and randomly from the concept library. (*ii*).**Human prior**: for the LAD dataset, important concepts for the human task are manually identified. We similarly apply CLIP-ViT-L/14 for annotation to compare the performance of concept selection. (*iii*).**ConceptNet**: ConceptNet is a knowledge graph dataset (Speer, Chin, & Havasi, 2017). We collect all concepts that have the *hasA*, *isA*, *partOf*, *HasProperty*, and *MadeOf* relations with categories to select concepts.

The results are shown in Tab.1, where the uninterpretable black-box linear probing[2] is also presented for reference (not bolding). Our method achieves the highest classification accuracy on most datasets, even close to the unexplainable black-box linear probing, especially excelling on challenging tasks such as LAD-Hair. The random selection method serves as a baseline and performs the worst. Surprisingly, the human prior concept selection method does not perform well, implying a disparity between the model's reasoning and decision-making process and the human's process. It is worth mentioning that our method does not utilize any external knowledge, and all concept selection is done automatically by the model. Despite this, it outperforms the ConceptNet concept selection method, which incorporates human prior knowledge, highlighting the effectiveness of CSM concept selection.

Few-shot Ability (A1.3). The essence of the concept-based model is to decouple the representations of the uninterpretable black-box model into a set of meaningful concepts, in other words, to inform the model which information should be paid more attention to. When the amount of data is limited, concepts assist the model in extracting the crucial components relevant to the classification from the representation of the black-box model, and this distinct guidance helps the model focus more on the task, leading to superior performance. Fig.4 illustrates the performance comparison under different data volume settings between CSM and linear probing on 7 datasets and their average result. Compared to the black-box model, our method achieves superior performance when little data is available, and exhibits a slight performance gap with larger amounts of sample sizes, which indicates that our method has maintained accuracy without sacrificing interpretability. Meanwhile, our method does not introduce additional human priors, so the performance improvement cannot be attributed to external knowledge injection, but rather to better decoupling of knowledge inherent in the model.

[2] Linear probing uses the representations extracted by CLIP image encoder to make predictions directly without interpretation.

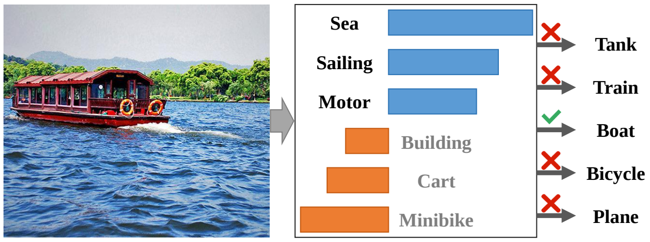

Figure 5: The decision process of the concept-based models.

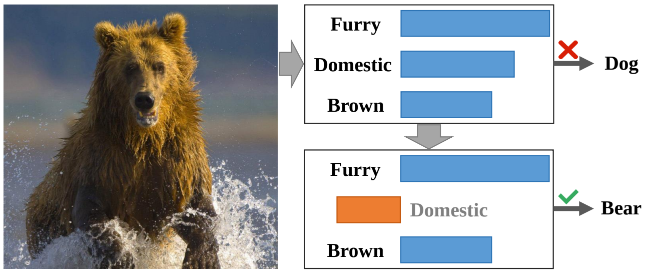

Figure 6: Model debugging through concept-based models.

Interpretability

Core Concepts (A2.1). Initially, we demonstrate the top 5 concepts with the highest activation value variances among the core concepts for each dataset, as shown in Tab.2. Consistent with our expectations, we observe that these concepts are tightly related to the core image compositions of the datasets, especially in fine-grained object datasets such as *leopard* in LAD-Animal, *monitor* in LAD-Electronic, *walnut* in LAD-Food, *ponytail* in LAD-Hair, and *streetcar* in LAD-Vehicle. The fact that these concepts are understandable to humans suggests that CSM can capture concept variations that are consistent with human expectations.

Model Debugging (A2.2). As an interpretable approach, one advantage of the concept-based model is its ability to help us understand the model's reasoning process, thereby making decisions more transparent. In Fig.5, we present the top 3 concepts *sea*, *sailing*, and *motor* with the highest normalized activation values in the CSM, as well as the bottom 3 activation concepts building, *cart*, and minibike. The model's classification of the image as a *boat* rather than a *tank*, *train*, *bicycle*, or *plane* is influenced by these concept activation values, which facilitates humans to comprehend the model's output.

Furthermore, after understanding the model's inference process, we can also intervene on the concepts of the misclassified samples to perform model debugging. For example, in Fig.6, the top 3 activation concepts are *furry*, *domestic*, and *brown*, which lead the model to misclassify the image as a *dog*. Upon analysis, it is determined that the concept *domestic* is incorrect, resulting in the classification error. Therefore, by manually setting the activation value of that concept to 0 or a negative value, we can achieve model debugging and assist the model in correctly classifying it as a *bear*.

Table 3: User study results.

Acc.	random	w/o GRS	w/o MFS	CSM
I2C	0.295	0.515	0.630	**0.695**
C2I	0.320	0.570	0.645	**0.680**
MD	0.080	0.260	0.215	**0.330**

Human Evaluation (A2.3). We evaluate the human interpretability in practical human-computer interactions through a user study and ablation of greedy rough selection (GRS) and mask fine selection (MFS) strategies. Specifically, we considered the following three scenarios: (*i*).**Image to Concepts (I2C)**: For a given image, we investigate whether humans perceive the concept with the highest CSM activation value. For each image, we display 1 concept from the top 5 highest activation values and randomly select 3 concepts from the remaining concepts, then ask users to determine which of the 4 concepts is most prominent in the image. (*ii*).**Concepts to Image (C2I)**: For a given combination of concepts, we assess whether humans can successfully predict the corresponding image. For each set of the top 5 highest concept activations, we display the corresponding image, another 1 image from the same category, and 2 additional images from different categories, then let users judge which of the 4 images best matches the concept activations. (*iii*).**Model Debugging (MD)**: When an image is misclassified, we examine whether humans could accurately identify the incorrect concept and correct it for model debugging. For a misclassified image, we provide the top 4 core concepts and prompt users to choose the concept most likely to be misidentified, then set the activation value of that concept to 0 and re-evaluate the accuracy.

For the user study, we randomly sample 5 categories from LAD-A and LAD-V, with 20 images per class. The study involves a total of 20 participants, and each answers 10 questions for every one of the above settings, the results are presented in Tab.3. Our method achieves the best results in all settings, as marked in bold. In tasks of concept-image correspondence, rough selection plays a crucial role, without which results in a severe drop in accuracy. In model debugging, fine selection is more pivotal, possibly due to the involvement of category information during mask training.

Conclusion

By utilizing CLIP as a concept annotator, we observe the long-tail distribution of concepts, based on which we propose a concept selection model that explores concepts without introducing any human priors, enabling a deeper understanding of the reasoning process in multimodal DNNs. This concept-based decoupling method brings enhanced few-shot capability and permits applications such as model debugging. In the future, it will become a promising research topic to explore the nature of hierarchy and correlations between concepts, leveraging simple concept combinations to form higher-level concepts, and contributing to more transparent AI systems.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61991451) and the Shenzhen Science and Technology Program (JCYJ20220818101001004).

References

Abid, A., Yuksekgonul, M., & Zou, J. (2022). Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International conference on machine learning* (pp. 66–88).

Amer, M., & Maul, T. (2019). A review of modularization techniques in artificial neural networks. *Artificial Intelligence Review*, *52*, 527–561.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, *538*(7623), 20.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 ieee winter conference on applications of computer vision (wacv)* (pp. 839–847).

Ghorbani, A., & Zou, J. Y. (2020). Neuron shapley: Discovering the responsible neurons. *Advances in neural information processing systems*, *33*, 5922–5932.

Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., … Yang, Y. (2024). *Connecting large language models with evolutionary algorithms yields powerful prompt optimizers.*

Hammoudeh, Z., & Lowd, D. (2022). Training data influence analysis and estimation: A survey. *arXiv preprint arXiv:2212.04612*.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., … others (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 87–110.

Han, X., Wang, Y.-T., Feng, J.-L., Deng, C., Chen, Z.-H., Huang, Y.-A., … Hu, P.-W. (2023). A survey of transformer-based multimodal pre-trained modals. *Neurocomputing*, *515*, 89–106.

He, X., & Peng, Y. (2019). Fine-grained visual-textual representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(2), 520–531.

Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018). Grounding visual explanations. In *Proceedings of the european conference on computer vision (eccv)* (pp. 264–279).

Huang, J., Shang, C., Xiong, A., Pang, Y., & Jin, Z. (2021). Seeing health with eyes: Feature combination for image-based human bmi estimation. In *2021 ieee international conference on multimedia and expo (icme)* (p. 1-6). doi: 10.1109/ICME51207.2021.9428234

Kazhdan, D., Dimanov, B., Jamnik, M., Liò, P., & Weller, A. (2020). Now you see me (cme): concept-based model extraction. *arXiv preprint arXiv:2010.13233*.

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* (pp. 2668–2677).

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *International conference on machine learning* (pp. 5338–5348).

Kong, C., Chen, Y., Zhang, H., Yang, L., & Yang, E. (2022). Multitasking framework for unsupervised simple definition generation. *arXiv preprint arXiv:2203.12926*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., … others (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, *123*, 32–73.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*.

Le, Y., & Yang, X. (2015). Tiny imagenet visual recognition challenge. *CS 231N*, *7*(7), 3.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Li, D., Zhang, H., Li, Y., & Yang, S. (2023). Multi-level contrastive learning for script-based character understanding. *arXiv preprint arXiv:2310.13231*.

Losch, M., Fritz, M., & Schiele, B. (2019). Interpretability beyond classification output: Semantic bottleneck networks. *arXiv preprint arXiv:1907.10882*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Marconato, E., Passerini, A., & Teso, S. (2022). Glancenets: Interpretable, leak-proof concept-based models. *Advances in Neural Information Processing Systems*, *35*, 21212–21227.

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

Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T.-W. (2023). Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, *93*, 106384.

Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., & Baralis, E. (2023). *Concept-based explainable artificial intelligence: A survey.*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Ramaswamy, V. V., Kim, S. S., Fong, R., & Russakovsky, O. (2022). Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. *arXiv preprint arXiv:2207.09615*.

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 ieee conference on secure and trustworthy machine learning (satml)* (pp. 464–483).

Schwalbe, G. (2022). Concept embedding analysis: A review. *arXiv preprint arXiv:2203.13909*.

Shang, C., Zhou, S., Yang, Y., Zhang, H., Ni, X., & Wang, Y. (2024). *Incremental residual concept bottleneck models.*

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Wang, B., Li, L., Nakashima, Y., & Nagahara, H. (2023). Learning bottleneck concepts in image classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10962–10971).

Wang, D., Cui, X., & Wang, Z. J. (2020). Chain: Concept-harmonized hierarchical inference interpretation of deep convolutional neural networks. *arXiv preprint arXiv:2002.01660*.

Wortsman, M., Ramanujan, V., Liu, R., Kembhavi, A., Rastegari, M., Yosinski, J., & Farhadi, A. (2020). Supermasks in superposition. *Advances in Neural Information Processing Systems*, *33*, 15173–15184.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., & Yatskar, M. (2023). Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 19187–19197).

Yuksekgonul, M., Wang, M., & Zou, J. (2022). Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*.

Zhang, H., Li, D., Yang, S., & Li, Y. (2022). Fine-grained contrastive learning for definition generation. *arXiv preprint arXiv:2210.00543*.

Zhang, H., Liu, Z., Shang, C., Li, D., & Jiang, Y. (2024). *A question-centric multi-experts contrastive learning framework for improving the accuracy and interpretability of deep sequential knowledge tracing models.*

Zhao, B., Fu, Y., Liang, R., Wu, J., Wang, Y., & Wang, Y. (2019). A large-scale attribute dataset for zero-shot learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 0–0).