Making AI Think Lean: Sparse Concept Bottleneck Models for Interpretable Decisions

Anonymous ACL submission

Abstract

Concept Bottleneck Models (CBMs) provide a promising approach to enhance interpretability in machine learning models. These models excel at disentangling and anchoring visual representations into human-comprehensible concepts. We present an approach to enhance visual model interpretability by incorporating natural language text directly extracted from images. We introduce the Visual-Rationale Alignment Learning (VIRAL) framework, which incorporates natural language text directly extracted from images to improve the interpretability of visual models. Through the use of the Gumbel-Sinkhorn algorithm for sparse alignment and extensive experimental analysis, VIRAL demonstrates its effectiveness in providing human-understandable explanations for predictions, contributing to the development of more transparent and trustworthy AI multimodal systems.

1 Introduction

Data in the real world is complex and often exhibit intricate symmetries and patterns. This complexity suggests that a limited number of factors could explain the extensive variation seen in real-world data. The success of representation learning in machine learning is largely dependent on the recognition and utilization of these patterns and structures. Concept-based learning (Koh et al., 2020) has emerged as a powerful approach to address this problem by anchoring representations in human-understandable concepts, such as colors, shapes, textures, and objects, which are crucial for interpretation and categorization. By focusing on these interpretable concepts, concept-based learning aims to create a more robust and transparent framework for understanding and manipulating large datasets that drive advances in machine learning. The concept explanations (Koh et al., 2020; Yuksekgonul et al., 2022) provided by concept bottleneck models (CBMs) offer insight into the inner workings of a prediction model by identifying the most crucial concepts on which the model relies when making a decision. To generate a meaningful explanation, a range of possible concepts and a set of examples that the model has previously encountered are presented. The explanation then highlights the concepts that frequently appear in the examples and aids the model in making accurate predictions. However, concept explanations are susceptible to spurious correlations within the data, resulting in unreliable interpretations. Sparsity emerges as a viable strategy to address the challenges posed by these spurious correlations by constraining the number of concepts considered by the model. We introduce the Visual-Rationale Alignment Learning (VIRAL) framework, which incorporates natural language text directly extracted from images to improve the interpretability of visual models.

By minimizing the alignment loss, VIRAL encourages the model to align the visual features with the most relevant rationale features while promoting sparsity in the alignment. The sparsity induced by the Gumbel-Sinkhorn algorithm enhances the interpretability of the model by focusing on the most important concepts. The effectiveness of the VIRAL framework is demonstrated through extensive experiments on real-world datasets. The results show that VIRAL achieves promising interpretability, as measured by established metrics, while maintaining competitive performance compared to baseline models.

2 Related Work

We explore some related work in relation to Concept/Attributes, Concept Alignment, and Latent Matchings.

Concept/Attribute based frameworks Attributes or concepts have the potential to significantly increase the interpretability of machine learning models, particularly in the context of data transfer between tasks (Palatucci et al., 2009; Frome et al., 2013; Lampert et al., 2009). Practioners have successfully mapped specific attributes or high-level concepts, such as hues, contours, or abstract notions, to model features, thus enabling the provision of human-comprehensible explanations for model predictions. This methodology facilitates the elucidation of factors that influence a model's decisions, thereby fostering improved understanding, trust, and debugging of the model. Concept Bottleneck Models (CBMs) Koh et al. (2020) are a promising approach to improve interpretability in machine learning. Unlike attribute-based models, which depend on predefined attributes that require extensive domain knowledge and may not capture the full complexity of the data, CBMs integrate the learning of high-level concepts directly into the model by incorporating a bottleneck layer with a dimension smaller than that of the input and output layers, forcing the network to learn a compressed representation of the input data. This integration enables CBMs to automatically discover and utilize meaningful intermediate concepts that are both interpretable and relevant to the prediction task.

Concept Alignment Concept alignment (Rane et al., 2023), a subfield of AI alignment, aims to ensure that AI systems and humans share a common understanding of concepts. Recent research (Rane et al., 2023; Wynn et al., 2023; Sucholutsky and Griffiths, 2023) has highlighted the importance of concept alignment for safe and beneficial AI development, exploring its relationship with value alignment. Further studies have delved into how humans and AI learn concepts, identifying pathways towards mutual understanding and suggesting methodologies to enhance concept alignment. This work contributes to these efforts by proposing a novel approach to facilitate concept alignment, with potential to address limitations of existing methods.

Learning with Matchings In many machine learning scenarios, 'learning with matchings' is crucial. It involves identifying optimal correspondences between item sets, such as matching users with products, aligning multilingual lexicons (Conneau et al., 2017; Hoshen and Wolf, 2018; Mukherjee et al., 2018), or tracking objects across video frames (Burke et al., 2020). This method leverages data structures and relationships to address complex challenges. The goal is to develop models that predict the best matchings,

3 Sparse Concept Bottleneck Model and Visual-Rationale Alignment (VIRAL)

This section introduces the Sparse Concept Bottleneck Model and Visual-Rationale Alignment (VIRAL) framework, which incorporates rationale selection, visual feature alignment, and sparsity constraints to enhance interpretability and performance in image classification tasks. Given a data set $\mathbf{X} \in \mathbb{R}^{N \times H \times L \times c}$ of N images, each with dimensions $H \times L$ and c channels, and a corresponding set of textual descriptions $\mathbf{t}_i$ for each image $\mathbf{x}_i$, VIRAL aims to align visual representations with the most informative and pertinent textual fragments, referred to as rationales. To incorporate rationales and improve interpretability, we introduce a rationale selector g_ϕ that operates on the textual descriptions $\mathbf{t}_i$ associated with each image $\mathbf{x}_i$. The framework, similar to Concept Bottleneck Models (CBMs) (Koh et al., 2020), employs a dual encoder architecture: a text encoder $f^{txt}(.)$ and an image encoder $f^{img}(.)$. The schema of VIRAL is shown in Fig 1 The rationale selector assigns

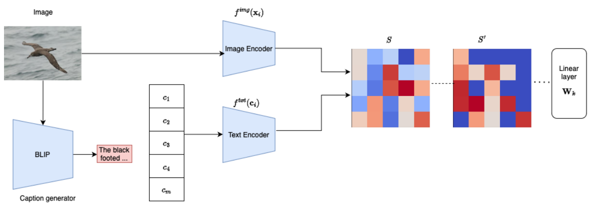

Figure 1: Overview of VIRAL, which processes an input image and its concept annotations through encoders to extract features, mapped into a common embedding space.

relevance scores to words or phrases in the text, identifying the most informative fragments. Let $\mathbf{r}_i$ denote the rationale for the i-th image, obtained by applying the rationale selector to the text. The rationales $\mathbf{c}_i$ provide a focused representation of the text, highlighting key aspects for understanding the image. The text encoder $f^{txt}(\cdot)$ uses rationales or concepts $\{\mathbf{c}_i\}_{i=1}^M$. These rationales or concepts represent the most informative aspects of the text for interpreting the images. On the other hand, the image encoder $f^{img}(\cdot)$ translates each image $\mathbf{x}_i$ into an image-based feature vector $f^{img}(\mathbf{x}_i)$. To capture the alignment between the image feature vectors and the rationale/concept vectors, a similarity matrix $S \in \mathbb{R}^{N \times M}$ is constructed.

$$\mathbf{S} \approx f^{\text{txt}}(\mathbf{c}_i)^T f^{\text{img}}(\mathbf{x_i}) \in \mathbb{R}^{N \times M} \quad (1)$$

Given the computation of $\mathbf{S}$ across all image-concept pairings, each image is endowed with a unique representation based on its similarity to each concept or rationale. This approach diverges from the complex projections used in related Concept-Based Model (CBM) methodologies, such as those proposed by Bachman et al. (2019); Tschannen et al. (2019). We contend that the similarity vector itself serves as an effective and robust image-concept representation, thereby obviating the need for additional computational layers often deemed superfluous in the literature (Wong et al., 2021). In a K-class classification scenario, we integrate a linear layer $\mathbf{W}_k \in \mathbb{R}^{N \times K}$ with the similarity matrix $\mathbf{S}$. This configuration yields the network output:

$$Y = \mathbf{S}\mathbf{W}_k^T \in \mathbb{R}^{N \times K} \tag{2}$$

The prediction loss $\mathcal{L}_{pred}$ is defined based on the linear model in equation 2. It measures the discrepancy between the predicted class probabilities $\hat{\mathbf{y}}$ and the true class labels $\mathbf{y}$. We use the cross-entropy loss to compute $\mathcal{L}_{pred}$:

$$\mathcal{L}_{pred} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \log \hat{y}_{ik} \tag{3}$$

where y_{ik} is the true label of the i-th image for the k-th class (0 or 1), and $\hat{y}_{ik}$ is the predicted class label.

The alignment loss $\mathcal{L}_{align}$ encourages the model to learn meaningful alignments between image and rationale features, captured by the similarity matrix $\mathbf{S}$. The similarity matrix S, computed using a similarity function, is typically dense with most elements nonzero. To focus on significant alignments, we use Gumbel-Sinkhorn (Mena et al., 2018), combining Gumbel-Softmax (Gumbel, 1954) with the Sinkhorn algorithm (Cuturi, 2013). The Gumbel-Softmax trick adds stochasticity, enabling a differentiable approximation of discrete choices. Alg 1 demonstrates how to obtain $\mathbf{S}'$.

Algorithm 1 Compute Selected Similarity Matrix using Gumbel-Sinkhorn

Input: Image features $f^{\text{img}}(\mathbf{x}_i)$, concept features $f^{\text{txt}}(\mathbf{c}_i)$, learnable matrix $\mathbf{W}$, temperature τ

Output: Selected Similarity Matrix $\mathbf{S}'$

Compute Similarity Matrix: $\mathbf{S} = f^{\text{img}}(\mathbf{x}) f^{\text{txt}}(\mathbf{r}_i)^T$;

Apply Gumbel-Max Trick:

- Generate Gumbel noise $\mathbf{G} \sim \text{Gumbel}(0,1)^{M \times N}$
- $\tilde{\mathbf{W}} = \text{softmax}((\mathbf{W} + \mathbf{G})/\tau)$

Compute Selected Similarity Matrix: $\mathbf{S}' = \tilde{\mathbf{W}} \odot \mathbf{S}$;

To quantify the effectiveness of the transformation from an original matrix $\mathbf{S}$ to a sparse matrix $\mathbf{S}'$ achieved through a Gumbel-Softmax mechanism, the alignment loss function, $\mathcal{L}_{\text{align}}$ is introduced. This loss function measures the fidelity of $\mathbf{S}'$ in capturing the essential structural characteristics of $\mathbf{S}$, while adhering to the sparsity constraints imposed by the Gumbel-Softmax process. The alignment loss can be expressed as follows:

$$\mathcal{L}_{\text{align}} = \|\mathbf{S} - \mathbf{S}'\|_F^2, \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.his formulation not only highlights the differences between the matrices but also penalizes larger discrepancies more severely, ensuring that $\mathbf{S}'$ closely aligns with the patterns and values found in $\mathbf{S}$.

The alignment regularization is added to the concept prediction loss $\mathcal{L}_{pred}$ and the alignment loss $\mathcal{L}_{align}$ to form the final objective function:

$$\mathcal{L} = \mathcal{L}_{pred} + \lambda_{align}\mathcal{L}_{align} \tag{5}$$

where λ_{align} is a hyperparameter.

4 Experimental Evaluation

Experimental Setup. We evaluated three different benchmark data sets to assess the proposed hierarchical framework, namely CUB (Wah et al., 2011), SUN (Xiao et al., 2010), and AwA (Xian et al., 2017) with their description in Tab 1.

These data sets cover a wide range of diversity in both the number of samples and their practical use. For vision models, we utilize CLIP (Radford et al., 2021) with a standard backbone, specifically

Table 1: Description of Datasets

Dataset	Attr.	Ex.	Labels
AwA (Animals with Attr.)	85	30,475	50
SUN (Scene Und.)	102	14,340	717
CUB (Caltech Birds)	312	11,788	200

Model	Dataset (Accuracy %)		
	CUB	SUN	AwA
Baseline (Images)	76.70	42.90	76.13
Label-Free CBMs	74.59	–	71.98
CLIP Embeddings	81.90	65.80	79.40
CDM^{H} (Panousis et al., 2023)	80.30	66.25	75.22
VIRAL (Ours)	**81.40**	**67.45**	74.70

Table 2: Classification Accuracy for Various Models. Bold values denote the best performance per dataset.

ViT-B/16. To avoid recalculating embeddings for images/patches and text data in each iteration, we pre-compute these embeddings using the chosen backbone. These embeddings are then loaded and used during the training phase to calculate the necessary metrics. For high-level conceptual analysis, we consider the class names of each dataset. We use BLIP (Li et al., 2022) to generate precise and contextually rich captions for diverse image datasets. The BLIP model, with its dual capabilities in image comprehension and natural language processing, is central to our automated caption generation strategy. We keep the value of $\lambda_{align} = 0.75$ and $\tau = 0.5$.

4.1 Classification Performance

This section evaluates the classification accuracy of VIRAL. Our evaluation compares several models to assess the classification and concept sparsification capabilities of our proposed model: (i) a baseline model without interpretability features, (ii) state-of-the-art Label-Free Concept Bottleneck Models (CBMs) (Oikarinen et al., 2023), (iii) tasks using CLIP embeddings, and (iv) classifications leveraging concept set similarity (CDM). We also highlight VIRAL's contributions to model interpretability and efficiency.

Table 2 presents the accuracy achieved by VIRAL and various baseline methods across three data sets. As observed, VIRAL consistently achieves competitive accuracy on all datasets. Notably, it surpasses the Label-Free CBM on all datasets, demonstrating the effectiveness of our sparse models. Although we primarily focused on accuracy, it is important to note that VIRAL also offers concept sparsification and interpretability advantages, which we analyzed separately.

Interpretability Metrics. In the absence of human annotators, we propose to assess the interpretability and groundability of our concept representation using Concept Consistency which measures image coherence and alignment per concept. Consistency is quantified by the average pairwise similarity of images linked to a concept, indicating that well-grounded concepts in the visual domain exhibit similar features. Concept consistency is computed by extracting visual features using a pre-trained CLIP models followed by calculating pairwise cosine similarities of these features. The average similarity score indicates visual consistency and concept alignment with its visual representations. This metric is evaluated across all concepts, providing insight into the model's ability to maintain consistent and interpretable concept representations.

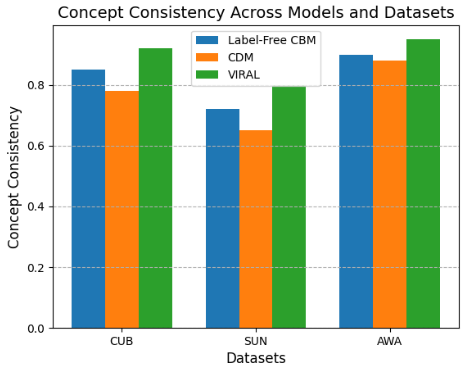

Figure 2: This figure evaluates three models—Label-Free CBM, CDM, VIRAL—across CUB, SUN, and AwA. The Concept Consistency, measures average pairwise similarity of concept-linked images, showing each model's ability to maintain coherent concept representations.

5 Conclusion

In this paper, we present VIRAL, a multi-faceted framework that improves the interpretability of visual models by incorporating natural language text. VIRAL extracts meaningful rationales from texts associated with images, which serve as a bridge between visual features and human-understandable concepts. The Gumbel-Sinkhorn algorithm acts as a differentiable concept selector, aligning visual features with extracted rationales and focusing the model on key concepts.

6 Limitations

The VIRAL framework, while innovative, has several limitations that could impact its efficacy and application. First, VIRAL's effectiveness of VIRAL depends on the quality and relevance of the natural language text associated with the images. Noisy, irrelevant, or explanatorily weak texts can result in poorly captured underlying concepts leading to suboptimal alignment and interpretability. Furthermore, VIRAL is limited to text-based explanations, which may not suffice for expressing complex visual patterns or abstract concepts better conveyed through visual means.

Additionally, the performance and interpretability of VIRAL are sensitive to hyperparameter settings, including the temperature parameter in the Gumbel-Sinkhorn algorithm and weighting coefficients for the loss terms. The optimal configuration of these parameters necessitates extensive experimentation and domain expertise, potentially limiting the accessibility and adaptability of the model. The incorporation of the Gumbel-Sinkhorn algorithm also adds significant computational complexity, particularly when dealing with large datasets or high-dimensional feature spaces, which may impede scalability and real-time application.

Evaluating the interpretability provided by VIRAL poses challenges because interpretability assessments are often subjective and context-dependent. Although the existing metrics offer some insights, they may not fully encapsulate human perception and understanding, necessitating user studies or expert evaluations for a more comprehensive assessment. In addition, the effectiveness of VIRAL can vary across different domains, such as medical or satellite imagery, where domain-specific knowledge is crucial for extracting meaningful rationales. Adapting VIRAL to these domains may require specialized preprocessing or domain-specific language models.

Despite its capacity to align visual features with interpretable rationales, VIRAL might still leave explanatory gaps. The decision-making process in models can involve complex interactions and transformations that are not fully elucidated by rationales alone, highlighting the need for additional techniques or complementary explanations to bridge these gaps.

References

Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Neural Information Processing Systems*.

Michael Burke, Kartic Subr, and Subramanian Ramamoorthy. 2020. Action sequencing using visual permutations. *IEEE Robotics and Automation Letters*, 6:1745–1752.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv'e J'egou. 2017. Word translation without parallel data. *ArXiv*, abs/1710.04087.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems*.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Neural Information Processing Systems*.

Emil Julius Gumbel. 1954. Statistical theory of extreme values and some practical applications : A series of lectures.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *ICML*.

Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. 2018. Learning latent permutations with gumbel-sinkhorn networks. *ArXiv*, abs/1802.08665.

Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. Learning unsupervised word translations without adversaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 627–632, Brussels, Belgium. Association for Computational Linguistics.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. 2023. Label-free concept bottleneck models. In *International Conference on Learning Representations*.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Konstantinos Panagiotis Panousis, Dino Ienco, and Diego Marcos. 2023. Sparse linear concept discovery models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2767–2771.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Sunayana Rane, Mark K. Ho, Ilia Sucholutsky, and Thomas L. Griffiths. 2023. Concept alignment as a prerequisite for value alignment. *ArXiv*, abs/2310.20059.

Ilia Sucholutsky and Tom Griffiths. 2023. Alignment with human representations supports robust few-shot learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 73464–73479. Curran Associates, Inc.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. 2019. On mutual information maximization for representation learning. *ArXiv*, abs/1907.13625.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001.

Eric Wong, Shibani Santurkar, and Aleksander Madry. 2021. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning*.

Andrea Wynn, Ilia Sucholutsky, and Thomas L. Griffiths. 2023. Learning human-like representations to enable learning human values. *ArXiv*, abs/2312.14106.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2251–2265.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492.

Mert Yuksekgonul, Maggie Wang, and James Y. Zou. 2022. Post-hoc concept bottleneck models. *ArXiv*, abs/2205.15480.