

# Learning Concise and Descriptive Attributes for Visual Recognition

An Yan<sup>\*◊</sup>, Yu Wang<sup>\*◊</sup>, Yiwu Zhong<sup>\*♣</sup>, Chengyu Dong<sup>◊</sup>, Zexue He<sup>◊</sup>,

Yujie Lu<sup>♣</sup>, William Yang Wang<sup>♣</sup>, Jingbo Shang<sup>◊</sup>, Julian McAuley<sup>◊</sup>

<sup>◊</sup>UC San Diego, <sup>♣</sup>University of Wisconsin-Madison, <sup>♦</sup>UC Santa Barbara

{ayan, yuw164, cdong, zehe, jshang, jmcauley}@ucsd.edu

{yujielu, william}@cs.ucsb.edu, yzhong52@wisc.edu

## Abstract

Recent advances in foundation models present new opportunities for interpretable visual recognition – one can first query Large Language Models (LLMs) to obtain a set of attributes that describe each class, then apply vision-language models to classify images via these attributes. Pioneering work shows that querying thousands of attributes can achieve performance competitive with image features. However, our further investigation on 8 datasets reveals that LLM-generated attributes in a large quantity perform almost the same as random words. This surprising finding suggests that significant noise may be present in these attributes. We hypothesize that there exist subsets of attributes that can maintain the classification performance with much smaller sizes, and propose a novel learning-to-search method to discover those concise sets of attributes. As a result, on the CUB dataset, our method achieves performance close to that of massive LLM-generated attributes (e.g., 10k attributes for CUB), yet using only 32 attributes in total to distinguish 200 bird species. Furthermore, our new paradigm demonstrates several additional benefits: higher interpretability and interactivity for humans, and the ability to summarize knowledge for a recognition task.

## 1. Introduction

Explaining black-box neural models is a critical research problem. For visual recognition, one line of research tries to classify objects with descriptions or attributes [12, 8, 39, 18, 22], which provide additional information beyond visual cues such as activation maps [41, 40]. However, they require in-depth human analysis and intensive annotation to obtain key attributes for a particular recognition task. Such a paradigm is costly and thus impractical to scale up when the number of classes and domains grows.

The recent advance of foundation models creates new

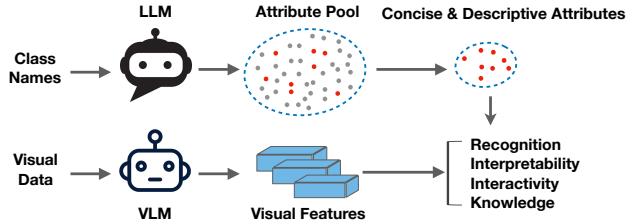


Figure 1: Our proposed paradigm for visual recognition via learning a concise set of descriptive attributes.

opportunities for building interpretable visual recognition models, as demonstrated by the powerful capabilities of models such as GPT-3 and ChatGPT in encoding world knowledge [5, 32, 21]. One can query useful visual attributes from LLMs and classify images via these attributes by converting visual features from vision-language models (VLMs) (e.g., CLIP [36]) into attribute scores [56]. One recent work [52] shows that a large set of attributes from LLMs (e.g., 50 attributes per class) can achieve comparable performance to image features in a linear probing setting. However, two key observations motivate us to re-think this formulation: (1) A large number of attributes dramatically hurts the interpretability of a model. It is unrealistic to manually check thousands of attributes to fully understand model decisions. (2) We surprisingly find that when the number of attributes is large enough (e.g., the dimension of image features), random words drawn from the entire vocabulary can perform equally well as LLM-generated attributes. Moreover, reducing the number of random words by 25% can still attain competitive performance. This indicates that redundant and noisy information exists in the massive LLM-generated attributes.

With our findings, we ask the research question: *Can we learn a concise set of representative visual attributes in the form of natural language to explain how visual recognition works? For example, can we find a few representative attributes to distinguish 200 bird species?* This is a non-trivial problem. Even for humans, it is not easy to summariz-

\* equal contributions.

size what are the representative visual attributes given many visual classes. To tackle this challenge, we propose a novel learning-to-search method, which uses image-level labels to guide the searching of discriminative attributes. Specifically, we train a learnable dictionary to approximate the embedding space of VLMs, and then find descriptive attributes in the latent text space via nearest neighbor search.

In summary, we propose a new paradigm for visual recognition (Figure 1), which seeks to learn a concise set of visual attributes in the form of natural language. Once learned, there are several benefits to our new paradigm: **(1)** Our discovered attributes are highly descriptive. On 8 visual recognition datasets, our model classifies images via these attributes and achieves comparable classification performance as image features, even if the number of attributes is much smaller than the dimension of image features. **(2)** The condensed sets of attributes enable strong interpretability for the model decision process through a few human-friendly text descriptions. **(3)** Additionally, our framework presents a natural language interface for humans to interact with. One can correct a wrong prediction during model inference, by perturbing the values of attribute scores where it made mistakes. **(4)** Lastly, these expressive attributes can be viewed as a concise form of knowledge to summarize useful features for a visual recognition task, without costly human effort.

Overall, our contributions are three-fold:

- Leveraging recent advances in foundation models, we propose a new paradigm for visual recognition by learning a concise set of attribute descriptions.
- To find these attributes, we propose a novel learning-to-search method which prunes the large attribute pool from large language models to a descriptive subset.
- We conduct extensive experiments across 8 visual recognition datasets to validate our recognition effectiveness and efficiency with additional benefits.

## 2. Methodology

In this section, we introduce our key components for a new paradigm of visual recognition. It mainly consists of three modules: **First**, in Section 2.1, given an image domain, we query large language models to obtain a large set of visual attributes for the categories of a task. **Second**, we use a semantic transformation (Section 2.2) to project the image features into attribute features via a vision-language model, where each dimension in the new space corresponds to an attribute concept, and a higher value represents higher correlation between the image and the attribute. **Finally**, given the large space of attributes, we propose a novel learning-to-search method (Section 2.4) to efficiently prune the attributes into a much smaller subset to obtain a concise model for classification.

### 2.1. Generating Attribute Concepts via LLMs

The first step of our framework is to obtain a set of appropriate attribute concepts. Given a dataset with different categories, (e.g., CUB with 200 bird classes), what are the distinctive visual attributes to recognize them? Manually labeling and designing these attribute concepts can be costly, and can not scale to large numbers of classes. Large Language Models (LLMs), such as GPT-3 [5] and ChatGPT, provide an alternative solution. We can view these language models as implicit knowledge bases with exceptional world knowledge on a variety of tasks and topics, which humans can easily interact with through natural language to query knowledge. To this end, prompt engineering, or the ability to ask good questions to language models, is still important. To effectively query knowledge from LLMs with regard to classifying images, we design two types of prompts.

**Instance Prompting for Class-level Features.** For each class  $c$  in a given task, our first design choice is to query class-level information from LLMs. We prompt a language model with the instance prompt:

*Q: What are the useful visual features to distinguish  $Y_c$  in a photo?*

where  $Y_c$  corresponds to the name of class  $c$  in the form of natural language.

**Batch Prompting for Group-level Features.** For certain datasets (e.g., CIFAR-100 and ImageNet), there is inherently a hierarchy that some categories belong to the same group. For example, in CIFAR-100, there is a superclass for every five categories. Hence, we propose batch prompting, where we ask the language model to reason about the distinctive visual features among a batch of categories:

*Q: Here are  $N_g$  kinds of  $Y_g$ :  $\{Y_{c_1}, Y_{c_2}, \dots, Y_{c_M}\}$ . What are the useful visual features to distinguish them in a photo?* where  $N_g$  is the number of classes in a group  $g$ ,  $Y_g$  is the name of the group,  $Y_{c_i}$  corresponds to the name of each class  $c_i$  in the form of natural language.

We present more details regarding our prompt design, robustness check of different prompts, and examples of the generated attributes in Appendix A.

### 2.2. Semantic Projection

After obtaining a pool consisting of  $N$  attribute concepts  $\mathcal{C} = \{a_1, a_2, \dots, a_N\}$ , the second challenge is how we can best leverage these attributes to build interpretable image classifiers. Recent advances of vision-language models such as CLIP bridge the gap between images and text, by pre-training models with large scale image-text pairs. Intuitively, converting from images to text is a discretization process that will unavoidably lose rich semantic information stored in an image.

To better preserve information, we use a semantic projection that transforms a visual feature into an attribute

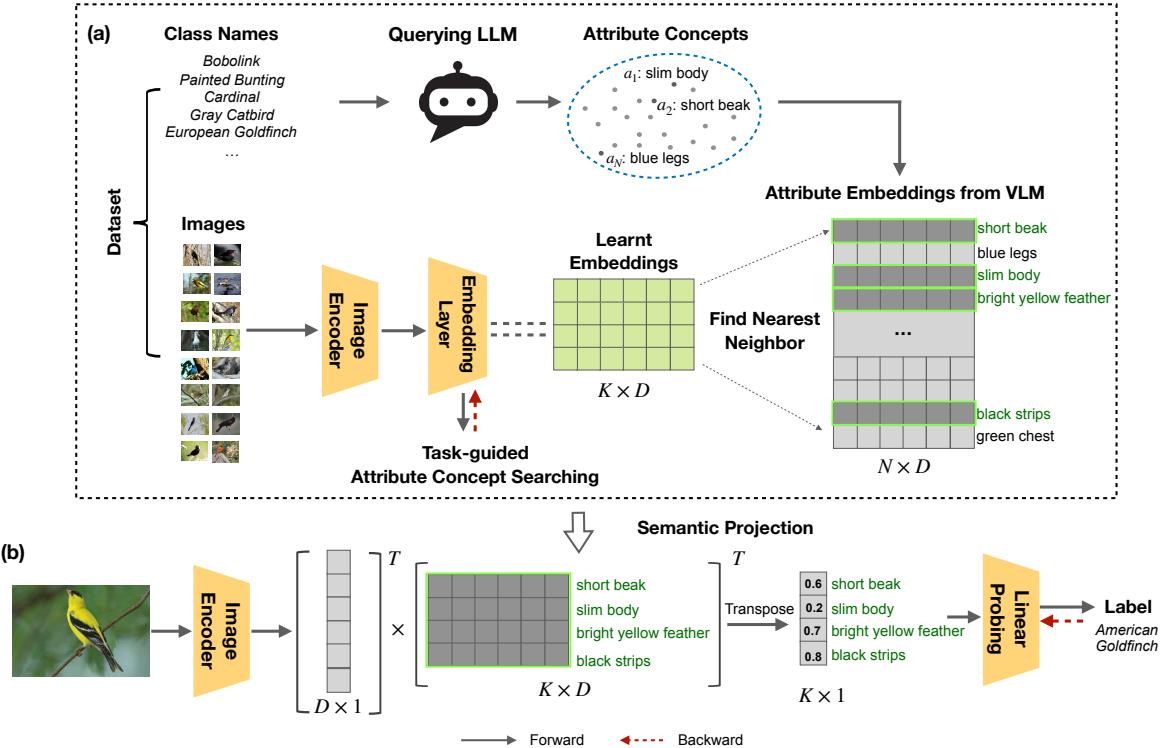


Figure 2: The framework of our model. (a) Querying attributes from LLMs and finding a concise set of representative attributes; (b) An example using the attributes for interpretable visual recognition.

concept space. Given an image  $I$ , we convert the  $D$ -dimensional image feature  $\mathbf{V} \in \mathbb{R}^D$  into an  $N$ -dimensional attribute concept vector  $\mathbf{A} \in \mathbb{R}^N$ :

$$\begin{aligned} \mathbf{V} &= \Theta_V(I), \mathbf{T}_i = \Theta_T(a_i) \\ s_i &= \cos(\mathbf{V}, \mathbf{T}_i), i = 1, \dots, N \\ \mathbf{A} &= (s_1, \dots, s_N)^T \end{aligned} \quad (1)$$

where  $\cos(\cdot, \cdot)$  is the cosine similarity between two vectors,  $s_i$  is the cosine similarity between two vectors.  $\Theta_V$  and  $\Theta_T$  are the visual and text encoder of a VLM.  $\mathbf{T}_i$  is the embedding of the  $i$ -th attribute in the attribute concept pool,  $i \in \{1, \dots, N\}$ .  $\mathbf{A}$  is the semantic vector of image  $I$ .

### 2.3. The Hypothesis of Attribute Concept Space

Conceptually, our semantic projection resembles principal component analysis, where we aim to find a set of bases in the form of natural language, and by projecting the images into these bases we obtain a new attribute concept space where each dimension in the space corresponds to a visual attribute concept. However, the large bag of attribute concepts we obtained from large language models is not the optimal language basis. As of today, LLMs are models that noisily condense world knowledge from the web,

and are not optimized for visual recognition or visual reasoning tasks. We hypothesize that there exist subsets of attributes that can still achieve high classification performance with a much smaller size. Intuitively, most attributes in the large attribute concept pool are irrelevant to classify a certain class. For example, attributes that describe dogs are less likely to be suitable attributes to recognize birds or cars. Practically, formatting a compact attribute set is also helpful for humans to interact with the model and understand its behavior better. A small number of attributes is much easier for diagnostic purposes and making decisions with these neural models, which is the ultimate goal of building interpretable models.

### 2.4. Task-Guided Attribute Concept Searching

Finding an expressive set of language bases is non-trivial. The massive attributes from LLMs are noisy, and finding a few representative attributes for hundreds of classes in a task can be challenging and costly, even for human experts with domain knowledge. An exhaustive search is also impractical given the large text space.

Inspired by dictionary learning and vector quantization techniques [43], we present a learning-to-search method that learns a dictionary to approximate an expressive sub-

set of attributes given fixed  $K$ . Specifically, we first define an embedding matrix  $\mathbf{E} \in \mathbb{R}^{K \times D}$ , where  $K$  is a  $K$ -way categorical that equals the number of attributes, and  $D$  is the dimensionality of embedding vectors  $\mathbf{V}$  and  $\mathbf{T}_i$  (*i.e.*, the latent dimension of VLMs), where  $\mathbf{V}$  and  $\mathbf{T}_i$  is the image embedding and the  $i$ -th attribute embedding shown in Eq.(1). Since our goal is to find  $K$  attributes to be expressive, we propose a task-guided attribute concept searching method to optimize for a particular task. For visual recognition tasks, we use a classification head to project the dictionary into  $K_C$  classes and guide the learning process with the categorical cross-entropy loss:

$$\mathcal{L}_{ce} = -\frac{1}{M} \sum_{i=1}^M \sum_{c=1}^{K_C} y_{i,c} \log(p_{i,c}) \quad (2)$$

where  $M$  is the number of images in a mini-batch,  $y_{i,c}$  is the binary indicator of the  $i$ -th image in the mini-batch belonging to class  $c$ , and  $p_{i,c}$  is the predicted probability of the  $i$ -th image belonging to class  $c$ .

But simply training with the guidance of the cross-entropy loss is suboptimal, as the embeddings  $\mathbf{E}$  are not in the same space of  $\mathbf{T}$ . Thus, we use the Mahalanobis distance as a constraint to encourage the embeddings to be optimized towards the latent space of vision-language models. Given a sampled probability distribution  $\mathbf{T}$ , the Mahalanobis distance of  $\mathbf{E}_j$  from  $\mathbf{T}$  is defined as

$$\mathcal{D}_{mah}^j = \sqrt{(\mathbf{E}_j - \boldsymbol{\mu}) \mathbf{S}^{-1} (\mathbf{E}_j - \boldsymbol{\mu})} \quad (3)$$

where  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_D)$  is the mean vector and  $\mathbf{S}$  is the positive-definite covariance matrix of  $\mathbf{T}$ . Then the regularization term is defined as:

$$\mathcal{L}_{mah}^j = \frac{1}{K} \sum_{j=1}^k \mathcal{D}_{mah}^j \quad (4)$$

Overall, our model is optimized with a mixture of two losses:

$$\mathcal{L}_{loss} = \mathcal{L}_{ce} + \lambda \sum_{j=1}^K \mathcal{L}_{mah}^j. \quad (5)$$

After training, we have the embedding matrix  $\mathbf{E}$  which will be used for searching the attributes from the attribute concept pool  $\mathcal{C}$ . Note that for  $\mathbf{E} \in \mathbb{R}^{K \times D}$ , each row of  $\mathbf{E}$  is a  $D$ -dimensional vector. We denote the  $j$ -th row of  $\mathbf{E}$  as  $\mathbf{E}_j$ . We use greedy search as follows:

$$\begin{aligned} \mathbf{T}_j^* &= \arg \max_{i \in \{1, \dots, N\}} \cos(\mathbf{T}_i, \mathbf{E}_j), \\ \text{s.t. } \mathbf{T}_j^* &\neq \mathbf{T}_k^*, \forall 1 \leq k < j, \\ \text{where } j &\text{ is from 1 to } K, \end{aligned} \quad (6)$$

As  $j$  iterates from 1 to  $K$ , we can find  $K$  attribute embeddings  $\mathbf{T}_j^*, j \in \{1, \dots, K\}$ , which corresponds to  $K$  expressive attribute concepts and are the condensed features containing the necessary knowledge for the task. With the selected attributes, we can calculate the semantic vector of each image as in Eq. (1), where each dimension of the vector is a similarity score between the image and an attribute. We evaluate the performance of these semantic vectors with linear probes, and the obtained linear model is used for inference and analysis.

### 3. Experiments

#### 3.1. Experimental Setup

**Datasets** We conduct our experiments on 8 different image classification datasets, including: CUB [44], CIFAR-10 and CIFAR-100 [24], Food-101 [4], Flower [31], Oxford-pets [33], Stanford-cars [23], Imagenet [9]. For Imagenet, it is not trivial to analyze all 1000 diverse classes. So we narrow the scope to 397 animal classes, with 509,230/19,850 samples for train/test. We denote this subset as Imagenet-Animals. For other datasets, most of them include images within a specific domain (CUB, Flower, Food, Oxford-pets, Stanford-cars), while CIFAR-10 and CIFAR-100 contain broader classes that lie across domains.

**Implementation Details** Our method involves two stages of training. The first stage consists of task-guided learning of a dictionary  $\mathbf{E}$  to approximate CLIP text embeddings and using this dictionary to find  $K$  attributes for visual recognition. For the Mahalanobis distance, the parameter  $\lambda$  is tuned with a grid search in  $\{1, 0.1, 0.01, 0.001, 0\}$ . The second stage is one-layer linear probing to classify semantic vectors. The batchsize is set to 4,096 for all datasets except 32,768 on Imagenet-Animals for faster converging. We set the number of epochs to 5,000 epochs with early stopping. The learning rate is set to 0.01 in all experiments with an Adam optimizer [20]. Unless specified, we use GPT-3 and CLIP ViT-B/32 for all performance comparison.

**Baselines** We compare with state-of-the-art works that leverage attributes either from human annotations or from LLMs. For a fair comparison, we use linear probes to evaluate all methods: (1) **CompDL** [56] builds semantic vectors using CLIP scores between human-designed attributes and images. (2) **LaBO** [52] is a recent work that builds semantic vectors with a large set of attributes from LLMs. (3) **Human** [44, 22]. Attribute labels for each image are annotated by humans. We compare with two versions: binary labels for each attribute, and calibrated labels with confidence scores given by annotators.

To validate the effectiveness of learning-to-search, we explore other baselines: (1) **K-means**. Perform K-means

Datasets	CUB			CIFAR-10			CIFAR-100			Flower		
$K$	32	200	400	8	10	20	64	100	200	32	102	204
LaBo	–	60.93	62.61	–	78.11	84.84	–	75.10	76.94	–	80.98	86.76
Ours	60.27	<b>63.88</b>	<b>64.05</b>	77.47	<b>80.09</b>	<b>87.99</b>	73.31	<b>75.12</b>	<b>77.29</b>	80.88	<b>87.26</b>	<b>89.02</b>
Datasets	Food			Oxford_Pets			Stanford_cars			Imagenet_Animals		
$K$	64	101	202	16	37	74	64	196	392	128	397	794
LaBo	–	79.95	81.33	–	76.91	84.33	–	72.33	74.39	–	74.88	75.49
Ours	78.41	<b>80.22</b>	<b>81.85</b>	76.29	<b>83.15</b>	<b>85.91</b>	72.07	<b>74.57</b>	<b>75.56</b>	74.48	<b>75.69</b>	<b>75.83</b>

Table 1: Comparison with state-of-the-art. LaBo is designed to use at least as many attributes as classes. We use “–” to denote non-applicability.

K (# of attributes)	8	16	32	312
Human Binary [44]	4.02	7.31	10.11	47.38
Human Calibration [22]	3.75	7.15	9.78	43.37
CompDL [56]	12.64	26.41	28.69	52.60
Ours	<b>31.67</b>	<b>48.55</b>	<b>60.27</b>	<b>65.17</b>

Table 2: Comparison with human annotations on CUB.

clustering on CLIP attribute embeddings, then find  $K$  attributes with nearest distance to each clustering center. Intuitively this can be a strong baseline, as  $K$  attributes close to each center can be distinctive. (2) **Uniform Sampling** from the large attribute pool. (3) **SVD**. After obtaining the attribute embeddings  $\mathbf{T}$ , we run SVD decomposition of  $\mathbf{T}$  to get the top  $K$  vectors and find attributes with the largest similarity with the  $K$  important vectors. (4) **Similarity**. We calculate the average score of each attribute across all images and then find the  $K$  attributes with the largest average scores. (5) **Img Features**. Black-box linear probing on latent image features with two linear layers and an intermediate dimension  $K$  as a reference.

### 3.2. Main Results

**Comparison with previous work** We first compare our method with LaBo [52]. It is designed to use  $M_c$  concepts per class with default number of 50, which corresponds to 10,000 attributes for CUB. For fair-comparison, we set  $M_c$  as 1 and 2 in the experiments. As shown in Table 1, our method outperforms LaBo with the same number of attributes on both the full and few-shot setting. Furthermore, our method can achieve similar accuracy with only a smaller number of attributes (e.g., 32 attributes for CUB). These results suggest that our learned attributes are discriminative enough to classify the images, despite given much fewer attributes.

We then further compare with human annotations from CUB. For  $K < 312$ , we select attributes based on their

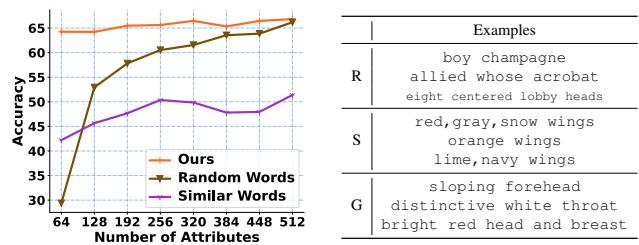


Figure 3: Performance comparison with random or similar words on CUB.

Table 3: Examples from Random (R), Silimlar (S), GPT-3 (G) attributes

accumulated confidence score for all samples. As shown in Table 2, human annotated attributes are more noisy than CLIP similarities. With the same attributes, CLIP scores from CompDL build more expressive features. Furthermore, our LLM-suggested attributes significantly outperform human designs, e.g. by using 16 attributes we achieve similar performance as 312 attributes defined by humans.

**Large-scale attributes behave like random words** We present our finding that LLM-generated attributes in a large quantity behave like random words. Specifically, we compare our method of using GPT-3 attributes with random or similar words. Here, we constructed random words by randomly choosing 1-5 words from the entire English vocabulary, and semantically similar words by combining 1-3 random colors with the noun “wings” as suffix. As shown in Figure 3, when  $K = 512$ , random words perform as well as GPT-3 attributes in terms of classification accuracy. Even reducing  $K$  from 512 to 256 does not significantly hurt its performance. But when  $K$  is small (e.g., 64), the performance of random words drops dramatically. We conjecture that it is because text embeddings randomly drawn from CLIP are nearly orthogonal bases [45]. Given an image feature  $\in \mathbb{R}^D$ , projection with a set of  $K=D$  orthogonal bases can perfectly preserve its information. We further

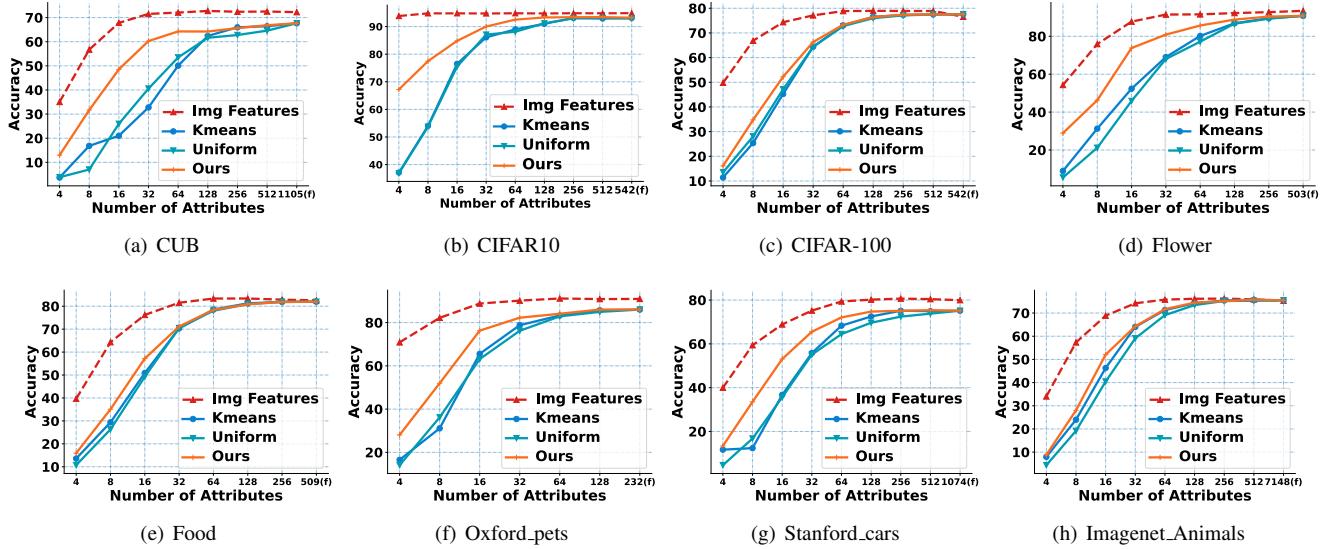


Figure 4: Overall Performance on all datasets. X-axis: number of attributes, Y-axis: Accuracy (%), “(f)” means “full”, i.e., all attributes in the pool are used. Uniform refers to uniform sampling.

explore how similar words (e.g., red wings, yellow wings) behave. Embeddings of similar words in a trained language model are not orthogonal bases hence the projection will lose information when  $K$  is large (e.g., intuitively it is hard to classify 200 bird species using only the color combination of wings). But as  $K$  gets smaller, since those similar words have close semantic meanings, they start to outperform random words. Overall, these findings motivate us to find a concise set of meaningful attributes while maintaining competitive performance.

**Number of attributes and selection methods** Finally, we study performance change under different number of attributes in Figure 4. First, our method is competitive with image features when  $K$  is large. Reducing number of attributes  $K$  to the number of classes  $C$  (e.g., 512 to 128 for CUB) does not significantly hurt performance, even for baseline methods. This validates our hypothesis that there is plenty of redundant information in the semantic space when the number of attributes is large (as used in LaBO [52]). It is possible to find a subset of expressive attributes for visual recognition. Second, we also consistently outperform other methods such as K-means clustering and uniform sampling, demonstrating the effectiveness of our task-guided searching method. Third, a heuristic design such as K-means performs similar as uniform selection. Note that though there is a performance gap between image features and using attributes, the gap can be minimized by using a stronger VLM, as the classification accuracy of attributes relies on the accurate estimation of the correlation between images and attributes ( see more results in appendix D ).

Datasets	CUB			CIFAR-100			
	K	8	16	32	8	16	32
GPT-3		31.67	48.55	60.27	34.77	52.24	66.30
GPT-3-Imagenet		30.81	49.29	60.41	33.80	51.01	65.61

Table 4: Ablation study w.r.t. different concept pools.

### 3.3. Ablation Study

**Robustness to the attribute pool** First, we aim to explore the effects of different initialized attribute concept pools generated by LLMs. On CUB and CIFAR-100, we compare two attribute pools, attributes generated from classes in each dataset, and attributes generated from the full set of ImageNet classes. As shown in Table 4, even with the large and noisy attributes from ImageNet, our method can still efficiently find a small number of representative attributes for a task, and obtains competitive classification performance.

**Effectiveness of learning-to-search** Then, we discuss possible choices for selection out of the large attribute pool. Results are shown in Table 5 with the following observations: heuristic methods such as K-means and SVD are not optimal choices for identifying the most distinctive attributes. In fact, they are sometimes less effective than uniform sampling. This is likely because we need to identify the most distinguishing attributes for visual recognition, rather than the most diverse ones based on text embeddings. Overall, our method significantly outperforms other baseline selection methods, showing its efficacy.

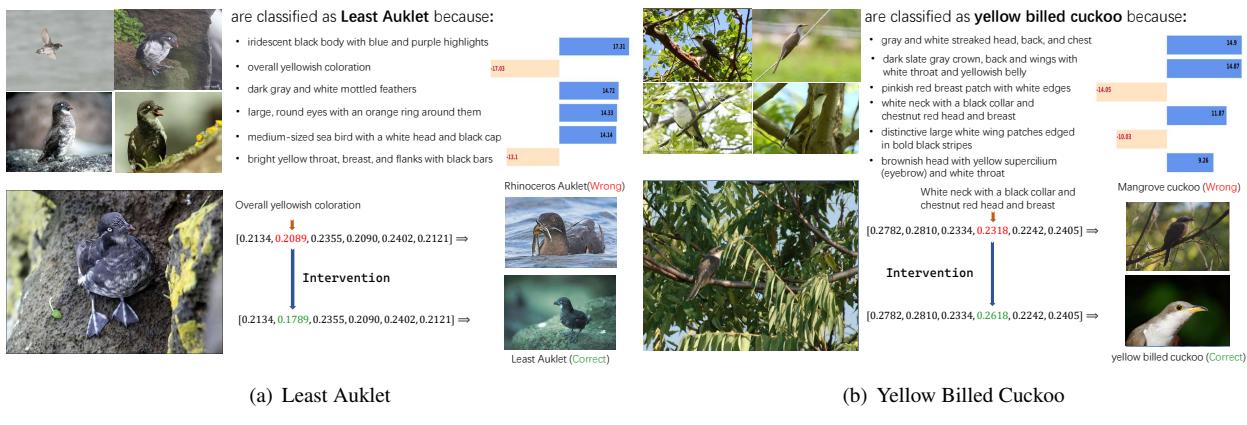


Figure 5: Examples on interpretability and interactivity. (1) The upper half of each figure show important attributes for two classes of birds. We choose 6 out of 32 attributes with highest importance scores, which are computed by multiplication between clip scores and weights in the linear probe, defined in Eq. (7). (2) The lower half of each figure demonstrates the intervention on the semantic vector (i.e., CLIP scores) to correct the prediction, we use  $\delta=0.03$  for all interventions on clip scores as an empirical value. The array of 6 scores are of the same order as the attributes.

Datasets	CUB			CIFAR-100		
K	8	16	32	8	16	32
K-means	16.83	21.02	32.76	25.39	45.26	64.41
Uniform	7.02	25.98	40.58	28.07	47.14	64.34
SVD	6.52	20.02	35.83	29.06	50.00	64.99
Similarity	4.73	9.72	18.00	26.75	45.61	62.79
Ours	<b>31.67</b>	<b>48.55</b>	<b>60.27</b>	<b>34.77</b>	<b>52.24</b>	<b>66.30</b>

Table 5: Ablation study w.r.t. different attribute selection strategies.

Dataset	CUB			
K	8	16	32	64
MAH	30.76	47.87	<b>60.27</b>	<b>64.25</b>
COS	28.96	47.35	58.27	63.25
CE	<b>31.67</b>	<b>48.55</b>	55.88	60.73
Dataset	CIFAR-100			
K	8	16	32	64
MAH	<b>34.77</b>	<b>52.24</b>	65.91	<b>73.31</b>
COS	31.98	51.15	65.02	72.80
CE	32.45	50.83	<b>66.29</b>	73.25

Table 6: Ablation study w.r.t. different regularization.

**Effectiveness of regularization** We compare the Mahalanobis distance (*MAH*) with two variations: (1) *COS*: For each vector  $\mathbf{E}_j$  and  $\mathbf{T}_i$  (of the  $i$ -th attribute) in the concept pool, we computed averaged cosine distance as follows:

$$\mathcal{L}_{cos} = \frac{1}{K^2} \sum_{j=1}^K \sum_{i=1}^K \frac{\mathbf{T}_i^\top \mathbf{E}_j}{\|\mathbf{T}_i\| \|\mathbf{E}_j\|}$$

(2) *CE*: Learning with Eq. (2) only. Results are in Table 6. Overall, Mahalanobis distance is an effective constraint to encourage the dictionary  $E$  to be close to the distribution of CLIP embeddings.

### 3.4. Analysis of Interpretability and Interactivity

We perform analysis and visualizations to show that:

(1) **Our learned attributes provide interpretability.** As shown in Figure 5, the upper half presents the images in a class  $c$  and high relevant attributes to recognize them. Specifically, we denote  $\mathbf{W} \in \mathbb{R}^{K_C \times K}$  as the weight of the

FC layer in linear probing, where  $K_C, K$  are the number of classes and attributes. Then for each image  $i$  and its semantic vector  $\mathbf{A} \in \mathbb{R}^K$ , we multiply the corresponding score vector of image  $i$  with the corresponding row of the FC layer  $\mathbf{W}_c$  to compute Importance Score  $\mathbf{IS} \in \mathbb{R}^K$ :

$$\mathbf{IS} = \mathbf{W}_c \otimes \mathbf{A} \quad (7)$$

where  $\otimes$  means element-wise multiplication. Then we present attributes with the top absolute values of  $\mathbf{IS}$  averaged over all samples in a class from the test set, with blue/orange bars indicating the positive/negative importance. Higher absolute values denote greater significance. Since all CLIP scores are positive [16], the positivity or negativity of high IS signifies their relevance to the class.

(2) **Our concise set of attributes enables simple interactivity.** As shown in the lower half of Figure 5, we can correct the model’s wrong predictions during inference by changing only a single similarity score between an image

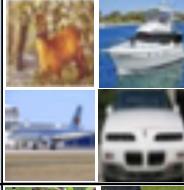
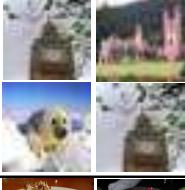
	CUB		<ul style="list-style-type: none"> <li>distinctive white throat</li> <li>bright red head and breast</li> <li>pinkish red breast patch with white edges</li> <li>bright yellow, green and blue plumage</li> <li>Red face with a black cap and bib</li> <li>Short legs for perching on reeds</li> <li>white and black spotted breast</li> <li>sloping forehead</li> </ul>	CIFAR10		<ul style="list-style-type: none"> <li>antlers (in males)</li> <li>pointed bow and stern</li> <li>propellers or jet engines</li> <li>moist slimy skin</li> <li>long head with a mane and tail</li> <li>landing gear</li> <li>portholes along the hull</li> <li>four wheels</li> </ul>
	CIFAR100		<ul style="list-style-type: none"> <li>a seat for the rider</li> <li>catkins (flowers) in spring</li> <li>many windows in the façade</li> <li>five pairs of walking legs</li> <li>smooth oval shaped sepals</li> <li>four-limbed primate</li> <li>headboard and footboard</li> <li>towers with conical roofs</li> </ul>	Flower		<ul style="list-style-type: none"> <li>Shiny wax coating on the spathe</li> <li>large, yellow or orange flower head</li> <li>bright pink color</li> <li>large, white petals with a yellow center</li> <li>pink to purple colored petals with red lips</li> <li>bright red and yellow petals</li> <li>pink, white, or lavender flowers with five petals</li> <li>deep purple or blue flowers</li> </ul>
Food	Food		<ul style="list-style-type: none"> <li>elbow macaroni noodles</li> <li>Shredded pork meat in the middle of the sandwich</li> <li>large pieces of clams visible in the chowder</li> <li>usually served in a warm wrap or burrito shell</li> <li>sliced into thin wedges or cubes</li> <li>thinly sliced raw fish</li> <li>tender squid rings inside</li> <li>a crisp, fried pastry dough exterior</li> </ul>	Oxford Pets		<ul style="list-style-type: none"> <li>black and tan coloring</li> <li>short coat of glossy black fur</li> <li>Long legs and neck</li> <li>Shade of red or wheaten color</li> <li>large, round eyes</li> <li>Pointed ears</li> <li>white blaze on face and chest</li> <li>greyish blue fur with silver tips</li> </ul>
Imagenet Animals	Imagenet Animals		<ul style="list-style-type: none"> <li>male finches have a bright red breast</li> <li>brownish-yellow fur</li> <li>small, four-limbed canid</li> <li>long, black shiny body</li> <li>the carapace is rough and bumpy</li> <li>white spots on the crab's shell</li> <li>English setters are bred in England</li> <li>long, wirehaired coat</li> </ul>	Stanford Cars		<ul style="list-style-type: none"> <li>signature Lincoln split headlamps</li> <li>large front grille with the signature BMW kidney shape</li> <li>large size with a wheelbase of 149.4 inches</li> <li>"4Runner" badge on the rear liftgate</li> <li>signature SRT8 grille with crosshair pattern</li> <li>Porsche logo on front grille and trunk lid</li> <li>S6 badge on the trunk lid</li> <li>unique HUMMER H2 logo on front grille</li> </ul>

Figure 6: A concise set of 8 descriptive attributes learned for each dataset with sampled images.

and the attribute that the CLIP model made a mistake on. This is a significant simplification compared with previous work [22] where they need to manipulate scores from a group of concepts for the CUB dataset. We present more user studies in appendix E.

### 3.5. Visualization of Our Discovered Attributes

We show our learned descriptive attributes with  $K = 8$  in Figure 6. Intuitively, we can observe these attributes are distinctive for each domain. Take birds recognition (CUB) as an example, the eight attributes covered most of the body parts of a bird (head, breast, legs, etc.). As we are condensing knowledge from hundreds of bird classes, each attribute broadly covers many categories. A bright red head and breast can be a noticeable visual attribute for many bird species, such as the Northern Cardinal and the Vermilion Flycatcher. Overall, explaining a domain with a few descriptive attributes is challenging, even for an expert with sufficient domain knowledge. But our model is able to automatically provide a level of knowledge to help humans understand how visual recognition works.

We then present case studies on CIFAR-10 with 4 attributes and CLIP scores of 10 random images from each class in Figure 7. In general, each image is activated in an distinguishable way in the heat map. Some attributes can distinguish a few classes, for example, cat and dog have



Figure 7: Case study on CIFAR-10. The numbers are CLIP similarity scores between each image and attributes.

higher activation on “fur coat” compared to automobile or truck. Thus “fur coat” may be an important feature to differentiate animals and vehicles.

## 4. Related work

**Interpretable Deep Learning** Interpretability is a critical research problem for deep learning with black-box models [11, 34, 37, 38, 13, 2, 50]. Some works study model behavior and explore if deep models could encode concepts for understanding [19, 28, 49, 29]. For image classification, preliminary attempts aim to describe objects with attributes [12, 26, 25] or building concept bottleneck mod-

els [22, 56, 55, 6]. These methods require in-depth human analysis and intensive labeling, which are impractical to scale to more classes and domains.

Recent works [30, 35, 52] tackle this problem by using GPT-3 as a knowledge base to query visual attributes or concepts. Specifically, [30, 35] generate descriptions with LLMs, and use them for knowledge-aware prompting for each class to improve zero-shot performance of CLIP [36]. For example, given the class name “bee”, it will augment it with attributes such as “A bee with black and yellow body”. Our work differs in that our goal is to learn representative attributes for visual recognition without using class names. LABO [52] extends the idea of concept bottleneck models by generating thousands of concepts from LLMs. Inspired by our finding that there is great redundancy in the large-scale attributes, we aim to learn a concise set of attributes that are initially generated from LLMs for each task, while maintaining the classification performance as possible. Concise attributes also enable stronger interpretability and interactivity, and can help humans to summarize critical knowledge for visual recognition in an automatic way.

**Foundation Models** Recently, foundation models [3], which are pre-trained with a large amount of data and large model sizes, have revolutionized machine learning research and many fields. These models are shown to be adaptable to a wide range of downstream tasks for computer vision [15, 46, 58], natural language processing [10, 7, 57, 48] and cross-modal research [27, 42, 17, 14]. One direction is to train LLMs such as GPT3 [5] and ChatGPT with massive text to serve as a powerful knowledge base with high interactivity and beyond. Another direction is to build VLMs [36, 51, 54, 53, 1], which connect vision and language by pre-training with image-text pairs and learning a joint embedding space for both. In this work, we use LLMs as a knowledge base for querying visual related knowledge, and use VLMs to bridge vision and text, presenting a new paradigm for interpretable visual recognition in the era of foundation models.

## 5. Discussion

There are many interesting topics to explore with our new paradigm. First, our framework is a plug-and-play model that can be readily applied to many other vision tasks, by simply changing the task-guided learning objective to a particular task, e.g., classification losses for object detection, video understanding, and 3D classification. Furthermore, a concise set of descriptive attributes enables interactivity for vision models and empowers human-machine cooperation in a user-friendly way through natural language interfaces. Lastly, we show the potential of summarizing knowledge for challenging vision tasks in the new era of LLMs, which could have broad impact for various domains.

## 6. Conclusion

In this work, we propose a new paradigm for visual recognition that leverages a concise set of descriptive attributes. Motivated by our insightful finding that significant redundancy exists in massive LLMs-generated attributes, we design a simple yet effective searching method guided by image-level labels, to identify an informative subset. Our new paradigm is validated across 8 datasets to achieve strong classification accuracy with multiple benefits and broad impacts, including efficiency, interpretability, human interactivity, and knowledge summarization.

## Acknowledgments

We would like to sincerely thank the anonymous reviewers and chairs for their careful review of our work, with helpful and constructive suggestions to improve the paper.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 9
- [2] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nat. Mach. Intell.*, 3(12):1061–1070, 2021. 8
- [3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 9
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV(6)*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer, 2014. 4
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 9
- [6] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nat. Mach. Intell.*, 2(12):772–782, 2020. 9
- [7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 9
- [8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 4
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 9
- [11] Jack Dunn, Luca Mingardi, and Ying Daisy Zhuo. Comparing interpretability and explainability for feature selection. *CoRR*, abs/2105.05328, 2021. 8
- [12] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009. 1, 8
- [13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018. 8
- [14] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 9
- [15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 9
- [16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 9
- [18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 1
- [19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR, 2018. 8
- [20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4
- [21] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*, 2023. 1
- [22] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 1, 4, 5, 8, 9
- [23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 4
- [24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4
- [25] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372. IEEE Computer Society, 2009. 8
- [26] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE Computer Society, 2009. 8
- [27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 9
- [28] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *IJCNN*, pages 1–10. IEEE, 2020. 8
- [29] Thomas McGrath, Andrei Kapishnikov, Nenad Tomasev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *CoRR*, abs/2111.09259, 2021. 8
- [30] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 9
- [31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE Computer Society, 2008. 4
- [32] TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022. 1
- [33] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE Computer Society, 2012. 4
- [34] P Jonathon Phillips, Carina A Hahn, Peter C Fontana, David A Broniatowski, and Mark A Przybocki. Four principles of explainable artificial intelligence. *Gaithersburg, Maryland*, 2020. 8
- [35] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022. 9, 12
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 9

- [37] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pages 1135–1144. ACM, 2016. 8
- [38] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535. AAAI Press, 2018. 8
- [39] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR, 2015. 1
- [40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1
- [41] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016. 1
- [42] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics, 2019. 9
- [43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3
- [44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4, 5, 12
- [45] Zihan Wang, Chengyu Dong, and Jingbo Shang. "average" approximates" first principal component"? an empirical analysis on representations from neural language models. *arXiv preprint arXiv:2104.08673*, 2021. 5
- [46] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 9
- [47] Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. Visual clues: Bridging vision and language foundations for image paragraph captioning. *arXiv preprint arXiv:2206.01843*, 2022. 12
- [48] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. 9
- [49] An Yan, Xin Eric Wang, Tsu-Jui Fu, and William Yang Wang. L2c: Describing visual differences needs semantic understanding of individuals. *arXiv preprint arXiv:2102.01860*, 2021. 8
- [50] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7922–7931, 2019. 8
- [51] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *CoRR*, abs/2204.03610, 2022. 9
- [52] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv preprint arXiv:2211.11158*, 2022. 1, 4, 5, 6, 9
- [53] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *CoRR*, abs/2205.01917, 2022. 9
- [54] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. 9
- [55] Mert Yüksekgönül, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *CoRR*, abs/2205.15480, 2022. 9
- [56] Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn primitive concepts? *arXiv preprint arXiv:2203.17271*, 2022. 1, 4, 5, 9
- [57] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 9
- [58] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 9