Doubly Right Object Recognition: A *Why* Prompt for Visual Rationales

Chengzhi Mao[1] Revant Teotia[1] Amrutha Sundar[1] Sachit Menon[1]
Junfeng Yang[1] Xin Wang[2] Carl Vondrick[1]
[1]Columbia University [2]Microsoft Research
{mcz,rt2891,as6431,sm4934,junfeng,vondrick}@cs.columbia.edu, wanxin@microsoft.com

Abstract

Many visual recognition models are evaluated only on their classification accuracy, a metric for which they obtain strong performance. In this paper, we investigate whether computer vision models can also provide correct rationales for their predictions. We propose a "doubly right" object recognition benchmark, where the metric requires the model to simultaneously produce both the right labels as well as the right rationales. We find that state-of-the-art visual models, such as CLIP, often provide incorrect rationales for their categorical predictions. However, by transferring the rationales from language models into visual representations through a tailored dataset, we show that we can learn a "why prompt," which adapts large visual representations to produce correct rationales. Visualizations and empirical experiments show that our prompts significantly improve performance on doubly right object recognition, in addition to zero-shot transfer to unseen tasks and datasets.

1. Introduction

Computer vision models today are able to achieve high accuracy – sometimes super-human – at correctly recognizing objects in images. However, most models today are not evaluated on whether they get the prediction right for the right reasons [14, 19, 48, 53]. Learning models that can explain their own decision is important for building trustworthy systems, especially in applications that require human-machine interactions [2, 15, 37, 50]. Rationales that justify the prediction can largely improve user trust [54], which is a crucial metric that the visual recognition field should push forward in the future.

Existing methods in interpretability have investigated how to understand which features contribute to the models' prediction [33, 34, 44, 46, 47, 52, 60]. However, saliency explanations are often imprecise, require domain expertise to understand, and also cannot be evaluated. [20, 22] have instead explored verbal rationales to justify the decision-making. However, they require manual collections of the plausible rationales in the first place, which subsequently are limited to small-scale datasets and tasks [24, 56].

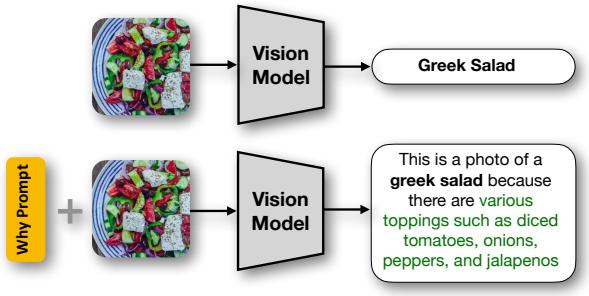

Figure 1. Visual reasoning for doubly right object recognition task. Motivated by prompting in NLP, we learn a *why* prompt from multimodal data, which allows us to instruct visual models to predict both the right category and the correct rationales that justify the prediction.

Scalable methods for explainability have been developed in natural language processing (NLP) through *prompting*. By adding additional instructions to the input, such as the sentence "think step-by-step," language models then output descriptions of their reasoning through the chain of thought process [57]. Since the explanations are verbal, they are easily understandable by people, and since the mechanism emerges without explicit supervision, it is highly scalable. In this paper, we investigate whether visual representations can also explain their reasoning through visual chain-of-thought prompts.

Our paper first introduces a benchmark for doubly right object recognition, where computer vision models must predict both correct categorical labels as well as correct rationales. Our benchmark is large, and covers many categories and datasets. We found that the visual representations do not have double right capability out-of-the-box on our benchmark. The recent large-scale image-language pre-trained models [41, 49] can retrieve open-world language descriptions that are closest to the image embedding in the feature space, serving as verbal explanations. However, the models often select the wrong rationales.

Instead, we propose a framework to explicitly transfer the chain-of-thought reasoning from NLP models into

vision models. We first query the large-scale language model [9] via the chain-of-thought reasoning for object category, where we obtain language rationales that explain discriminative features for an object. We then collect images containing both the category and the rationale features using Google image search. We then train visual *prompts* to transfer the verbal chain of thought to visual chain of thought with contrastive learning, where features of images and their rationales are pulled together. Our "why" prompts obtain up to 26 points gain at doubly right performance when evaluated on our benchmark. In addition, visualizations and quantitative results show that our why prompts zero-shot transfer to unseen tasks and datasets. We believe this "doubly right" object recognition task is a future direction which the visual recognition field should go forward for. Our data and code is available at `https://github.com/cvlab-columbia/DoubleRight`.

2. Related Work

Explainability. Visual recognition models achieve high performance on the classification tasks, yet they often provide vague and unreliable interpretations and rationales [8]. There are two lines of research for interpreting image classification using neural networks: feature visualization and language explanation. Feature visualization methods [6, 32–34] find the inputs that maximize the outputs of learned features. Gradient-based feature visualization [28, 42, 44, 46, 47, 52, 60] highlight the input features/pixels in images that are most important for the network to make decisions. Since the visualizations are often abstract, it is hard for a non-expert to understand. In addition, saliency map [44] highlights regions that may contain overlapping concepts such as color, texture, and shape, which is hard to disentangle. The second line of research uses language-based explanation methods [20, 22] to generate visual explanations. However, those methods require human annotations, which limits their ability to evaluate on a larger scale.

External Knowledge. Visual models often learn spurious correlations without external knowledge [29]. External knowledge allows models to learn the right features and obtain better transferability [26, 45]. Large-scale pretrained language models, such as GPT-3 [9], contains knowledge and commonsense learned from the Internet [38]. [57] shows that designing the right prompt, such as the chain of thought, improves the model's ability for language reasoning, which we leverage as an external knowledge source. Other sources of external knowledge include interactions [1, 11, 55], physics [31, 59], etc. [30] provides descriptions together with the objective category, which improves recognition performance. However, since they do not annotate rationales in their approach, they cannot measure CLIP's ability to produce the right rationales.

Visual Attributes Several works have studied visual attributes in images [17, 25, 27, 35, 36, 39, 63]. [58] used the visual attributes for animal classification and [23] used them for face verification. However, some of the attributes are spurious correlated [51] with the prediction task. In this work, we aim to generate visual rationales that produce robust visual attributes like shape and parts, instead of spurious features such as the background.

Visual Reasoning. [61] proposed a new cognition benchmark that the model needs to predict both the answer and the rationale to be correct. Visual Question Answering [3] performs visual understanding as a QA task, such as questions about COCO images. However, the questions are not asking for rationales to justify object recognition. [62] shows that multi-modal vision language models can perform zero-shot image-language tasks, such as image captioning. However, all existing visual reasoning work does not directly evaluate object recognition on the rationales they provided.

Prompting. Prompt tuning is a lightweight adaptation method for language task [64–66]. Recently, the computer vision field has adapted the language and proposed the visual prompts to adapt vision models [4, 5, 21, 43]. Due to their lightweight, it has been shown effective for continuous learning [12]. One advantage of the visual prompt is that it does not require model access at test time, which is flexible [4]. While existing visual prompt methods focus on improving the recognition task performance, we propose to use this lightweight prompt to improve the models' ability to provide visual rationales.

3. Language Rationales to Visual Rationales

We first introduce how to obtain the language rationales for discriminating an object. We then show how we translate the language descriptions into visual images and construct the dataset. We then propose to construct a "why prompt" to adapt the large visual models to produce the right rationales for their predictions.

3.1. Rationale from Language Model

A standard way of getting visual rationales for image predictions is through manual annotation. However, manual annotation is expensive, especially when applied to large-scale datasets. Recent advances in large-scale language models, such as the GPT3, demonstrate the ability to provide various commonsense knowledge in language. By providing chain-of-thought instructions to the language model, [57] shows that language models can perform the task of interest as well as produce explanations. We can use large-scale language models as a tool to collect rationales by asking the right language prompts. In addition, since the rationales are presented in language, it is easy to understand, even for non-experts.

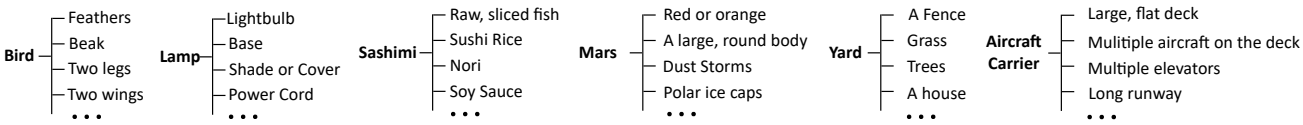

Figure 2. Examples of the rationales generated by prompting GPT3 with the chain of thought reasoning on the features of objects. We show one example from each of the six datasets—CIFAR10, CIFAR100, Food101, Caltech101, SUN, and ImageNet— we studied. The **bold** word is the category, and the list follows is the generated rationales. GPT3 can produce reasoning for why an object is as predicted.

Table 1. List of datasets that include both the model prediction and rationales. Animals with attributes (AWA) collects discriminative attributes for animals. CUB collects verbal rationales on only birds. BDD-X collects rationales for explaining driving scenarios. VAW and Broaden collect a large number of attributes. However, they are often not the right rationales for explaining why objects are classified, such as color. The * indicates the number of annotated video frames. The existing datasets are often small-scale, limited in domains, and not annotating rationales for the object recognition task. Our framework allows the automatic collection of diversified categories over a large scale. Note that the $^+$ name in our dataset, such as CIFAR-10$^+$, indicates that we collect the same set of categories in CIFAR-10 from google with our pipeline. Our benchmark is large, containing more categories and dataset variants than prior methods.

Dataset Name	Number of Categories	Number of Descriptions	Number of Images	For Right Rationales
AWA [24]	50	85	30,475	Yes
CUB [20, 56]	200	N/A	11,788	Yes
BDD-X [22]	906	1,668	26,000*	Yes
VAW [40]	N/A	650	72,274	No
Broaden [7]	584	1,300	62,476	No
Ours CIFAR-10$^+$	10	63	2,201	Yes
Ours CIFAR-100$^+$	100	540	18,318	Yes
Ours Food101$^+$	101	435	15,212	Yes
Ours Caltech101$^+$	101	516	16,849	Yes
Ours SUN$^+$	397	2,170	75,381	Yes
Ours ImageNet$^+$	1000	5,810	271,016	Yes

Motivated by the recent work in language prompting [9, 30], we propose to ask the language model: what are the valid rationales for an object prediction? Specifically, to obtain rationales for visual objects, we ask GPT3 the following question:

```
Q: What are useful visual features for
    distinguishing a {category name} in a
    photo?
A: There are several useful visual
    features to tell there is a {category
    name} in a photo:
-
```

where GPT3 automatically generates the answer for us, which scales to large, unseen categories. In Figure 2, we show random examples of the rationales generated by querying GPT3. Each category is provided with a list of discriminative features for the category, which is consistent with how people explain their predictions.

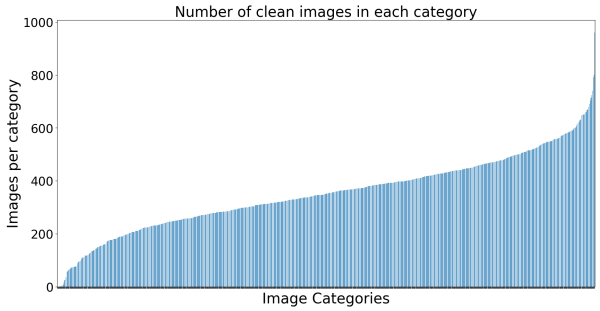

Figure 3. Histogram of the number of images per ImageNet category retrieved by Google search. There are natural imbalances in the data. Learning model under such imbalanced natural data distribution is an interesting problem to study.

3.2. Knowledge Transfer to Visual Domain

Since not every rationale exists in every image of the given object, we cannot directly apply this list of rationales to images based on their category. We need external knowledge to collect images containing specified rationales, so that we can correspond the language rationales to the visual rationals in images. Prior work [45] uses structured external knowledge from Wiktionary and WordNet to obtain images with text descriptions.

In the era of the Internet, people often Google a new word if they don't know what the object — the word refers to — looks like. Once people see the image examples from Google, they quickly understand what the word refers to in the real world. Our method is motivated by this natural procedure, where our system Googles "what the rationales look like" to visually understand them.

We propose to leverage this external knowledge from web search engines, a larger-scale knowledge base than Wiktionary [45]. We use Google image search to obtain images that contain our specified visual features, and we find that the returned images match the language queries with high accuracy. The search engine serves as a cross-modal translator, translating the knowledge from the language domain to the visual domain by retrieving images.

We use the following queries to query images that belong to our specified category and contain our specified attribute:

```
1. {category name} which has {attribute
    name}
2. {attribute name} of {category name}
3. a photo of {category name} because
```

Figure 4. Examples for the images in our collected ImageNet$^+$ dataset, where each image corresponds to one category label and one rationale. We show three random categories, and for each category, we show three random rationales. The rationales are consistent with the image features that decide the object's category.

```
there is {attribute name}
```

and we retrieve the top 50 images returned by Google search. More query sentences can be constructed if more images are needed, which we leave for future work. We remove the duplicated and incorrect images in preprocessing.

Google image search allows us to obtain images containing specific attributes. We perform the above automatic pipeline on ImageNet categories and obtain a dataset that contains images and the rationales that explain why this image is a particular category. For the dataset collected with our framework, we show a histogram of the number of images per ImageNet category in Figure 3. The data is S-curved due to the natural, non-interventional collection procedure. We also show examples of rationales and the retrieved images in Figure 4, where images can often be explained by rationales. Since our system is automatic, we in addition run our pipeline on the categories in CIFAR-10, CIFAR-100, Food101, Caltech101, and SUN datasets. Table 1 shows the scale of our collected images and compares them with prior datasets that are related to our work. Our pipeline allows richer rationales and is more extensive in the number of images and categories. To evaluate how accurate the Google image search can retrieve the images with the specified attribute as well as the category, we conduct a human study on the quality and consistency, which we describe in Section 4.4.

3.3. Learning the Why Prompt

Using external knowledge from both the language model and Google, we show we can collect image datasets containing both the category and rationales. Our pipeline allows us to evaluate the quality of this doubly right task on a large scale over several datasets for the first time. Our results show that doubly right object recognition is a challenging task for large visual models, such as CLIP.

We then seek a way to adapt and improve large pre-trained models so that they can provide the correct rationale for the predictions. Motivated by the prompting in natural language processing, we propose constructing a visual "why" prompt that instructs the model to produce the right rationales for the prediction. We optimize the visual prompt to ask "what are the visual rationales that explains the prediction of this image". We use the following input prompts or deep prompts to adapt the model.

Prompt Design. We study how to adapt transformer-based models since they are state-of-the-art. Our visual *why* prompts append additional tokens P_k to the input sequence of the vision transformer:

$$\mathbf{x} = [\mathbf{x}; P_0, P_1, ..., P_k] \tag{1}$$

The remaining transformer parameters and computations are kept the same as the original.

Deep prompt. Besides adding context to the input sequence, the deep prompting method [21] adds prompts to the intermediate layers in the transformer, which is a more powerful adaptation method than single-layer input prompt. Let $\mathbf{x}_i$ be the latent token sequence of the $i-$th layer in the transformer. We can add a prompt to each latent layer:

$$\mathbf{x}_i = [\mathbf{x}_i^0; P_i, P_i^1, ..., P_i^k] \tag{2}$$

$$\mathbf{x}_{i+1} = \text{Head}(\mathbf{x}_i) \tag{3}$$

where the Head indicates the transformer block. We find deep prompt is particularly useful for learning on large-scale datasets, such as ImageNet.

Learning Objective. We now introduce the training objective to instruct models for doubly right object recognition. We use cross-modal image-to-text contrastive learning to train the model for the right rationales. Using our above pipeline, we have collected a set of images and their rationales, which we denote as $\{(\mathbf{x}_i, \mathbf{t}_i)\}$. In contrast to the prior text prompts for object recognition, which uses:

```
This is a photo of [CATEGORY]
```

we instead create the training text prompt to be:

Table 2. Benchmarking six large-scale image-language pretrained models on doubly right recognition over six datasets. The accuracy for doubly right is gray-scaled. We **bold** the best accuracy. ↑ indicates higher number is better. For the same CLIP model, when model capacity increases from Res50 to H/14, the models' ability to get doubly right also increases, even though CLIP has never been trained on this metric. However, the doubly right accuracy on large-scale datasets is still low. For example, the double right accuracy is less than 1% on ImageNet. While larger model provides higher accuracy for object classification, they often provide wrong rationales, as indicated by the increase in RW percentage. Our evaluation suggests that this new doubly right object recognition task is challenging for existing large-scale visual models.

	CIFAR-10$^+$				CIFAR-100$^+$				Food101$^+$			
Model	RR ↑	RW↓	WR↓	WW↓	RR↑	RW ↓	WR↓	WW↓	RR↑	RW↓	WR↓	WW↓
FLAVA	29.44	53.04	5.83	11.68	3.57	59.72	5.44	31.26	3.95	55.78	4.86	35.42
CLIP-Res50	30.65	50.85	8.51	9.97	4.51	58.94	7.76	28.78	6.53	61.47	5.31	26.69
CLIP-Res101	30.41	50.61	8.76	10.21	5.09	60.31	7.41	27.19	5.23	64.65	4.61	25.46
CLIP-B/32	36.98	46.71	9.00	7.30	5.23	59.93	7.30	27.53	5.48	63.57	5.41	25.53
CLIP-B/16	35.28	49.63	9.25	5.83	5.61	64.09	6.16	24.13	6.04	67.65	4.26	22.04
CLIP-L/14	**42.57**	44.52	7.06	5.84	**6.43**	63.71	7.73	22.13	**5.73**	70.07	4.30	19.91

	Caltech101$^+$				SUN$^+$				ImageNet$^+$			
Model	RR ↑	RW ↓	WR ↓	WW↓	RR↑	RW ↓	WR↓	WW↓	RR↑	RW↓	WR ↓	WW ↓
FLAVA	2.21	61.57	3.85	32.38	**0.95**	14.62	10.72	73.70	0.40	28.31	3.58	67.70
CLIP-Res50	4.13	61.60	5.31	26.69	0.78	19.87	10.93	68.41	0.61	40.42	3.69	55.26
CLIP-Res101	4.57	64.24	4.61	25.46	0.89	21.29	11.68	66.75	0.65	42.74	3.64	52.98
CLIP-B/32	4.51	65.92	5.42	24.15	0.86	23.32	10.72	65.16	0.68	42.69	3.87	52.76
CLIP-B/16	4.60	68.30	4.84	22.25	0.81	23.41	10.50	65.68	0.63	46.73	3.41	49.23
CLIP-L/14	**5.99**	66.55	5.96	21.50	0.94	24.27	11.21	63.58	**0.72**	50.00	4.34	44.94

Table 3. Gain of using why prompt to adapt models to perform doubly right object recognition. We evaluate the CLIP-H/14 model, except for ImageNet, where we use CLIP-B/32 due to the large dataset size. We use a deep prompt on ImageNet, which is indicated by *. We **bold** the best doubly right accuracy. Learning a why prompt significantly improves the models' ability to predict the right category as well as the right rationales, reducing the failures when the model predicts the right category with wrong rationales.

		RR↑		RW↓		WR↓		WW↓	
Datasets	Prompt Length	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
CIFAR-10$^+$	3	42.57	**70.82**	44.52	18.25	7.06	6.32	5.84	4.62
CIFAR-100$^+$	3	6.43	**22.27**	63.71	44.61	7.73	9.97	22.13	23.14
Food101$^+$	3	5.73	**25.25**	70.07	51.83	4.30	5.83	19.91	17.08
Caltech101$^+$	3	5.99	**23.64**	66.55	52.43	5.96	5.86	21.50	18.06
SUN$^+$	100	0.94	**6.70**	24.27	8.29	11.21	23.76	63.58	61.24
ImageNet$^+$	30*	0.68	**3.63**	42.69	21.70	3.87	7.66	52.76	25.34

```
This is a photo of [CATEGORY] because
    there is [RATIONALE]
```

This allows the model to learn what is the correct rationales that explain the category prediction.

We use the pretrained image-language model, where we encode the image with an image encoder F_θ and the corresponding rationale with a text encoder T. To train the model to rank the correct rationales higher than the other negative rationales, we minimize the following image-to-text contrastive loss function,

$$\mathcal{L}_s(\mathbf{x},\mathbf{t},\mathbf{y}) = -\mathbb{E}_{i,j}\left[\mathbf{y}_{ij}\log\frac{\exp(\cos(\mathbf{z}_i^{(I)},\mathbf{z}_j^{(T)})/\tau)}{\sum_k \exp(\cos(\mathbf{z}_i^{(I)},\mathbf{z}_k^{(T)})/\tau)}\right], \tag{4}$$

where $\mathbf{z}_i^{(I)} = F_\theta(\mathbf{x}_i)$ and $\mathbf{z}_i^{(T)} = T(\mathbf{t}_i)$ are the features from the image and text, respectively. $\mathbf{y}_{ij}$ indicates which image-text are paired in the dataset and which are not. We set $\mathbf{y}_{ij} = 1$ when the image and text are from the same data point. cos denotes the cosine similarity function. τ is the temperature hyperparameter to scale the confidence of the prediction. We then use gradient descent to optimize the visual prompt such that this loss is minimal.

3.4. Evaluation Metric

We define the following metric to evaluate the models' predictions as well as their rationales. We formulate producing the rationales as a ranking task, where we provide the model sentence descriptions containing a pair-wise combination of all the categories and rationales, and the image-language model will retrieve the closer ones. Specifically, we present the model with a sentence in this format:

```
This is a photo of [CATEGORY] because
    there is [RATIONALE]
```

and ask the model to return the sentence that has the closest representation to the visual embeddings. Since one image may have multiple rationales to explain the category, following the standard practice [16], we use a top-K accuracy — if the top K rationales include the ground truth, then it is

Table 4. Zero-shot gain of using why prompt to adapt models to perform doubly right object recognition. For all datasets we use CLIP-H/14 model, except for ImageNet, where we use CLIP-B/32. We **bold** the best accuracy for doubly right object recognition. Our method can obtain better rationales on unseen datasets than baseline (CLIP) model.

Zero-Shot Transfer		RR↑		RW↓		WR↓		WW↓	
Training Datasets	Testing Datasets	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
CIFAR-100$^+$	CIFAR-10$^+$	42.47	**54.99**	44.52	33.58	7.06	4.87	5.84	6.57
CIFAR-100$^+$	Food101$^+$	5.73	**8.35**	70.07	44.61	4.30	9.97	19.91	23.14
CIFAR-100$^+$	Caltech101$^+$	5.99	**15.07**	66.55	56.34	5.96	6.21	21.50	22.38
Caltech101$^+$	CIFAR-10$^+$	42.47	**49.63**	44.52	39.17	7.06	5.60	5.84	5.60
Caltech101$^+$	CIFAR-100$^+$	6.43	**13.20**	63.71	49.11	7.73	10.38	22.13	27.30
Caltech101$^+$	Food101$^+$	5.73	**7.36**	70.07	61.05	4.30	4.49	19.91	26.69
SUN$^+$	CIFAR-10$^+$	42.47	**49.00**	44.52	40.04	7.06	5.78	5.84	5.18
SUN$^+$	CIFAR-100$^+$	6.43	**13.11**	63.71	49.63	7.73	8.32	22.13	28.93
SUN$^+$	Food101$^+$	5.73	**8.94**	70.07	51.90	4.30	6.67	19.91	32.48
SUN$^+$	Caltech101$^+$	5.99	**13.58**	66.55	55.14	5.96	5.39	21.50	25.88
ImageNet$^+$	CIFAR-10$^+$	36.98	**38.68**	46.71	43.80	9.00	9.25	7.30	8.27
ImageNet$^+$	CIFAR-100$^+$	5.23	**15.67**	59.93	39.89	7.30	9.51	27.53	55.57
ImageNet$^+$	Food101$^+$	5.48	**8.31**	63.57	46.59	5.41	5.97	25.53	39.12
ImageNet$^+$	Caltech101$^+$	4.51	**16.71**	65.92	45.42	5.42	7.66	24.15	30.20
ImageNet$^+$	SUN$^+$	0.86	**1.98**	23.32	7.02	10.72	14.96	65.16	76.90

counted as correct. The predicted category is based on the majority vote of the top-K reterieved categories [30]. We denote the metric as follows:

(1) Right classification with right rationale (RR);

(2) Right classification with wrong rationale (RW);

(3) Wrong classification with right rationale (WR);

(4) Wrong classification with wrong rationale (WW)

We desire a high accuracy for RR and a low percentage for RW, WR, and WW. We will evaluate the above metric on our collected dataset where rationale ground truth is provided. We select 20% of the data as the hold-out test set. We will train the model on the remaining 80% data.

4. Experiment

We evaluate six large-scale image-language pretrained models on our doubly right benchmark. We find that the model often produces the wrong rationales for the predictions. We then show that our *why* prompts significantly improves the models' ability to produce the right rationales quantitatively and qualitatively. Lastly, we show that we can have a hierarchy of visual rationales where the model can provide sub-rationales.

4.1. Benchmark Existing Models

We start our investigation by evaluating existing large-scale image language models [41, 49] on our collected doubly right dataset. Our evaluation includes images in the category of CIFAR-10, CIFAR-100, Food101, Caltech101, SUN, and ImageNet. Since we recollect the image based on the category name of those dataset, we use $^+$ to denote our collected dataset. We set $K = 5$ for the top-K accuracy of the doubly right prediction. We study FLAVA [49] and five variants of CLIP [41], where evaluation results are in Table 2. We find that for the same CLIP model, increasing their capacity from Resnet 50 [18] to ViT Huge/14 [13] generally improves performance in retrieving the right rationales, even if the model has never been trained on the doubly right task. FLAVA model performs worse than all CLIP model variants, except on SUN$^+$. Despite the high classification accuracy, doubly right object recognition is challenging for all models, where models produce incorrect rationales more often than the correct ones. Specifically, the best CLIP-H/14 model only obtains 1% accuracy on ImageNet$^+$ doubly right recognition task. Our evaluation shows that doubly right object recognition is still an open challenge for the large-scale dataset.

4.2. Why Prompt for Visual Rationales

To adapt the vision model so that they can perform doubly right object recognition, we apply our "why" prompting to visual foundation models. On ImageNet$^+$, we adapt the CLIP-B/32 due to the large size of the dataset. We use deep prompt with prompt length 30 for each of the 12 layers. We train 10 epochs with a learning rate of 10. For all the other datasets, we adapt the best CLIP-H/14 model. We train model for 100 epochs with a learning rate of 40. We use a prompt size of 3, except for SUN where we use 100. For each dataset, we train on 80% of the data and test on the hold out 20% test data.

We show our results in Table 3. By finding a why prompt to adapt the visual model, we significantly improve the models' ability to produce the right predictions as well as the right rationales, up to **28** points. Using our prompt, the model also sometimes predicts the wrong categories when

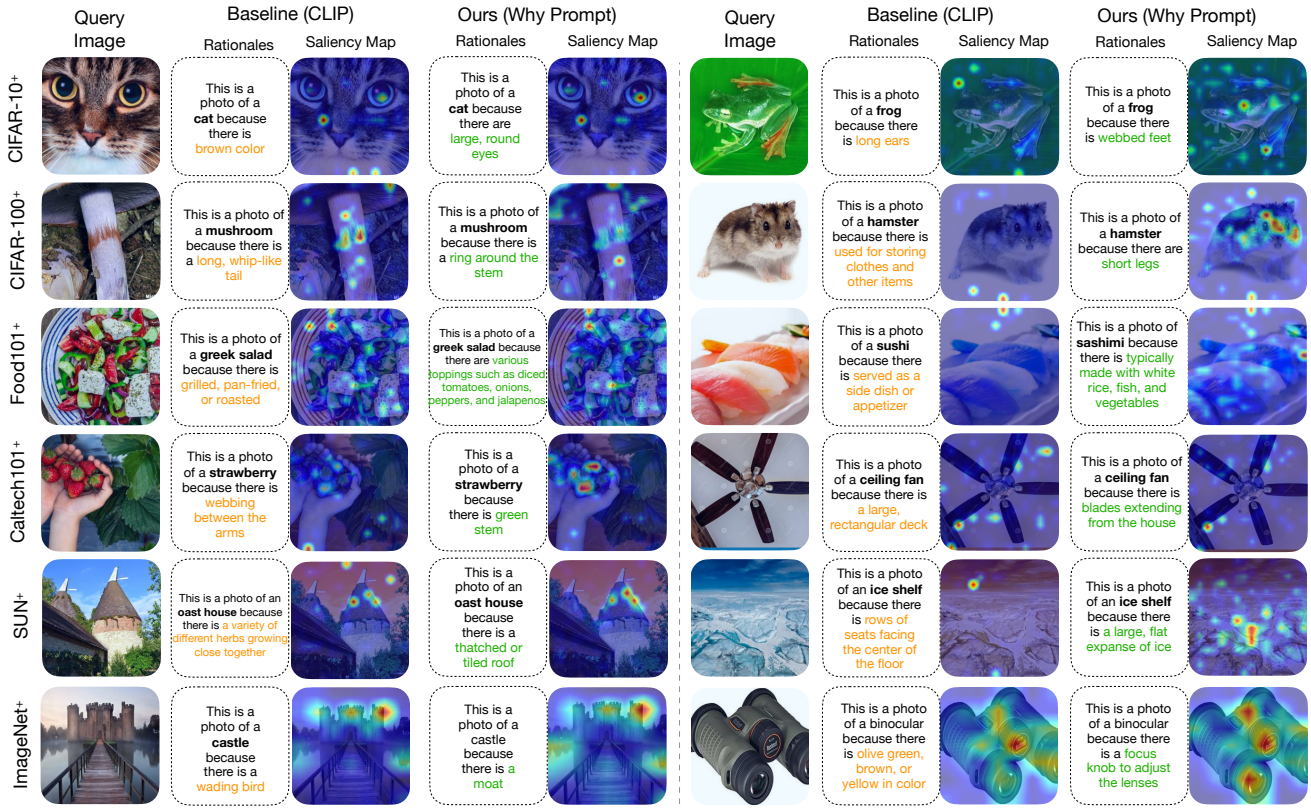

Figure 5. Visualization for the doubly right recognition task. For each row, we show image examples from categories of one dataset. In columns 2, 4, 7, and 9, we show the rationales produced by the model to explain the prediction. In column 3, 5, 8, and 10, we show the saliency map [10] that models look to produce the prediction and rationales. While the state-of-the-art H/14 CLIP model produces the correct category with the wrong rationales, our method can produce the correct category with the right rationales. In addition, our visual prompt also adapts the model to use the right image region to produce the prediction.

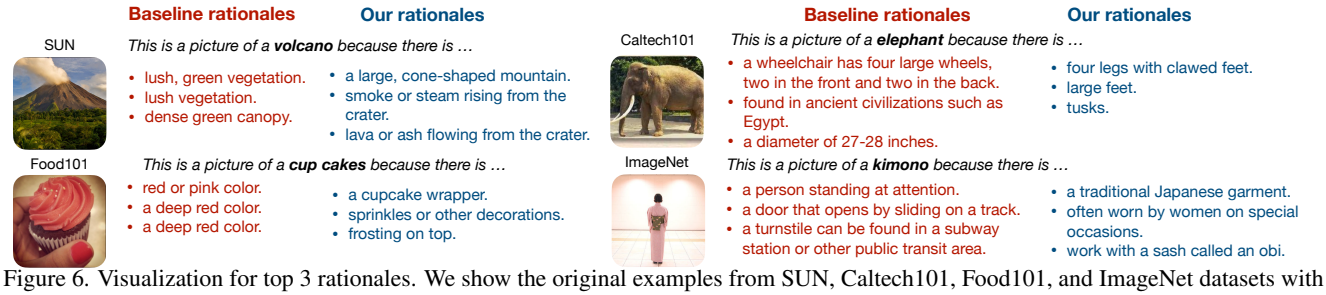

Figure 6. Visualization for top 3 rationales. We show the original examples from SUN, Caltech101, Food101, and ImageNet datasets with high-resolution images. Even though we do not annotate rationales for images in these datasets, our method can successfully transfer the rationales learned from our data and apply them.

the rationales are wrong (higher WW), suggesting our *why* prompt achieve higher consistency in reasoning between category and rationales. Our prompt method is still effective when the prompt size is as small as 3, containing minimal parameters, which is lightweight to apply. In addition, learning our why prompt can maintain the classification accuracy on the original ImageNet validation set, where CLIP-B/32 obtains an accuracy of 59.16%, and we obtain an accuracy of 59.20%.

Zero-shot Transferability for Why Prompt. We find that the above generated "why" also transfers to unseen datasets and categories. In Table 4, we show the doubly right accuracy obtained by zero-shot transfer, where we obtain up to 12 points gain. For example, by learning the model on the SUN$^+$ dataset which contains natural scenes, our method can teach CLIP to provide the correct rationales for food images (Food101$^+$), by 3 points better. This experiment shows our *why* prompt can adapt large scale models to produce correct predictions with right rationales, and generalize to novel categories, suggesting the effectiveness of our method.

Visualizations on Doubly Right dataset. We visualize images from the test set of our benchmark. In Figure 5, we show the top 1 rationales retrieved by the baseline and our

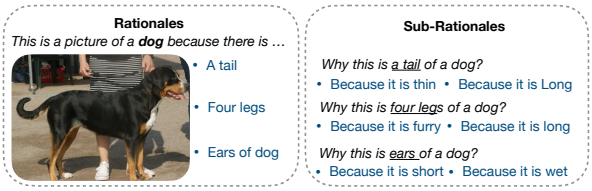

Figure 7. Visualizations for the hierarchical rationales on CIFAR-10^+. Our method obtains reasonable sub-rationales by learning an additional sub *why* prompt.

prompted model. Our method often produces a correct explanation, while the baseline does not. We also visualize the saliency map corresponding to the rationales [10], where we can also see that the why prompt instructs the model to look at the right region to explain the prediction.

Qualitative results on the original dataset. In addition to evaluating the doubly right performance on our collected dataset above, we also evaluate the established, original data. We study ImageNet, SUN, Caltech, and Food101 datasets since they contain high-resolution images that allow detailed explanations. Though we cannot directly evaluate the doubly right accuracy of those dataset images since they do not contain annotated ground truth, we show visualizations for the rationales generated on those datasets. In Figure 6, we find our learned "why" prompt also transfers well in improving the doubly right performance on images in those datasets.

4.3. Analyzing Hierarchical Rationales

The rationales provided by our approach may not be the basic explanations. For example, with the above method, we can explain that this is a photo of a dog because there are four legs. We are curious whether the model can also provide another level of visual reasoning for why the four legs cause the image to be a dog. We explore whether our framework can provide even more fundamental explanations for the generated rationales.

To obtain sub-rationales, we query GPT3 with:

```
Q: What are useful visual features for
    distinguishing a {attribute name} of a
    {category name} in a photo?
A: There are several useful visual
    features to tell there is a {attribute
    name} of a {category name} in a photo:
```

After getting the sub-rationales through language chain of thought, we transfer the knowledge to visual domain by Google query:

```
A photo of {CATEGORY}, because there is
    {sub-level attribute name} {attribute
    name}
```

where we collect 10583 images that contain attributes of specific patterns denoted by the sub-rationales. We split the dataset into an 80% training set and 20% of testing set.

Table 5. Accuracy for hierarchical visual rationales. On CIFAR-10^+ categories, we evaluate CLIP and our models' ability to get the category, the rationales, and the rationales' explanations correct (RR). Our method produces better sub-rationales than CLIP.

	RR	RW	WR	WW
CLIP-H/14	0.05	38.14	0.20	61.61
Ours	**26.19**	53.86	1.58	18.35

We train our why prompt with the same objective in Equation 4. We then evaluate the correctness of the rationales and the sub-rationales. We experiment on CIFAR-10^+ and train the model for 25 epochs. When categories, rationales, and sub-rationales are correct, it is counted as doubly right (RR). In Table 5, baseline CLIP has 0 accuracy. Our method improves RR by 26 points. In Figure 7, we visualize this hierarchical rationale, where our method can produce sub-rationales that explains the rationales hierarchically.

4.4. User Study

To evaluate the quality of our doubly right dataset, we conduct a human study. We show 200 images randomly sampled from our dataset and ask the user to check whether the category and rationales match what's inside the image. We conduct this study on 20 participants. Since the task requires visual reasoning, it takes 20 to 40 minutes for one participant to complete the study. On average, the user thinks 86.2% of the annotation is correct, with a standard deviation of 8.5. This shows our pipeline can collect visual data with visual rationales of reasonably high quality.

5. Conclusion

In this paper, We study an essential yet under-explored visual problem, getting the correct rationales for the predictions, which we name as the "doubly right" object recognition task. We construct large-scale datasets containing categories from various datasets with rich rationales, which allows us to evaluate this doubly right metric directly. Our work proposes a pipeline that transfers the rationales knowledge from language models to visual models, which significantly improves the doubly right object recognition accuracy on both seen and unseen categories. Our work provides a benchmark and algorithm that allows the visual recognition field to push this doubly right task forward.

Acknowledgement

This research is based on work partially supported by the DARPA GAILA program, the DARPA KAIROS program, the NSF NRI Award #2132519, a GE/DARPA grant, a CAIT grant, and gifts from JP Morgan, DiDi, and Accenture. We thank Jianan Yao on feedback for user study.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[4] H Bahng, A Jahanian, S Sankaranarayanan, and P Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, page 2022, 2022.

[5] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting. *arXiv preprint arXiv:2209.00647*, 2022.

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

[7] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

[8] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[10] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[11] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018.

[12] Jonathan Conder, Josephine Jefferson, Khurram Jawed, Alireza Nejati, Mark Sagar, et al. Efficient transfer learning for visual tasks via continuous optimization of prompts. In *International Conference on Image Analysis and Processing*, pages 297–309. Springer, 2022.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[15] Fabian Eitel and Kerstin Ritter. Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification. In Kenji Suzuki, Mauricio Reyes, Tanveer F. Syeda-Mahmood, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support - Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings*, volume 11797 of *Lecture Notes in Computer Science*, pages 3–11. Springer, 2019.

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[17] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R. Scott, and Serge J. Belongie. The imaterialist fashion attribute dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 3113–3116. IEEE, 2019.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European conference on computer vision*, pages 3–19. Springer, 2016.

[21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

[22] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving ve-

hicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.

[23] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Describable visual attributes for face verification and image search. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 33, pages 1962–1977, October 2011.

[24] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.

[25] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 684–700. Springer, 2016.

[26] Angli Liu, Jingfei Du, and Veselin Stoyanov. Knowledge-augmented language model and its application to unsupervised named-entity recognition. *arXiv preprint arXiv:1904.04458*, 2019.

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1096–1104. IEEE Computer Society, 2016.

[28] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[29] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. *arXiv preprint arXiv:2204.12363*, 2022.

[30] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[31] Roozbeh Mottaghi, Mohammad Rastegari, Abhinav Gupta, and Ali Farhadi. "what happens if..." learning to predict the effect of forces in images. In *European conference on computer vision*, pages 269–285. Springer, 2016.

[32] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3510–3520. IEEE Computer Society, 2017.

[33] Anh Mai Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3387–3395, 2016.

[34] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[35] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[36] Genevieve Patterson and James Hays. COCO attributes: Attributes for people, animals, and objects. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 85–100. Springer, 2016.

[37] Sérgio Pereira, Raphael Meier, Victor Alves, Mauricio Reyes, and Carlos A. Silva. Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne L. Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett A. Landman, M. Jorge Cardoso, Carlos A. Silva, Sérgio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications - First International Workshops MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings*, volume 11038 of *Lecture Notes in Computer Science*, pages 106–114. Springer, 2018.

[38] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[39] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13018–13028, June 2021.

[40] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13018–13028, June 2021.

[41] 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.

[42] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev

Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[43] Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Andrew Jackson. Fine-tuning image transformers using learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12155–12164, 2022.

[44] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[45] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022.

[46] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR, 2017.

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[49] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[50] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

[51] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022.

[52] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[53] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[54] Randy L Teach and Edward H Shortliffe. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, 14(6):542–558, 1981.

[55] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8581–8590, 2018.

[56] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[58] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2017.

[59] Tian Ye, Xiaolong Wang, James Davidson, and Abhinav Gupta. Interpretable intuitive physics model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.

[60] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.

[61] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[62] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[63] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[64] Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization. *arXiv preprint arXiv:2205.00049*, 2022.

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.