Identifying Interpretable Subspaces in Image Representations

Neha Kalibhat [1] **Shweta Bhardwaj** [1] **Bayan Bruss** [2] **Hamed Firooz** [3] **Maziar Sanjabi** [3] **Soheil Feizi** [1]

Abstract

We propose Automatic Feature Explanation using Contrasting Concepts (FALCON), an interpretability framework to explain features of image representations. For a target feature, FALCON captions its highly activating cropped images using a large captioning dataset (like LAION-400m) and a pre-trained vision-language model like CLIP. Each word among the captions is scored and ranked leading to a small number of shared, human-understandable concepts that closely describe the target feature. FALCON also applies *contrastive interpretation* using lowly activating (counterfactual) images, to eliminate spurious concepts. Although many existing approaches interpret features independently, we observe in state-of-the-art self-supervised and supervised models, that less than 20% of the representation space can be explained by individual features. We show that features in larger spaces become more interpretable when studied in groups and can be explained with high-order scoring concepts through FALCON. We discuss how extracted concepts can be used to explain and debug failures in downstream tasks. Finally, we present a technique to transfer concepts from one (explainable) representation space to another unseen representation space by learning a simple linear transformation.

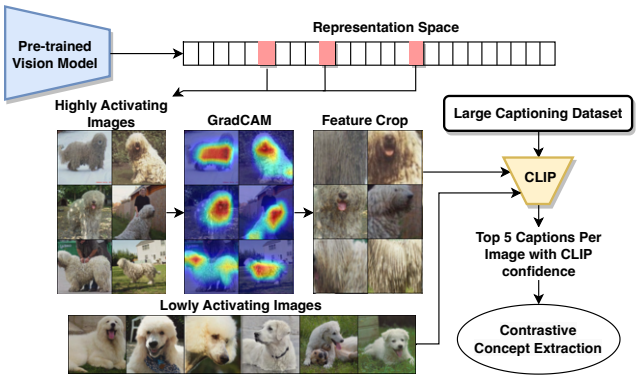

Figure 1. **Framework of FALCON:** We outline the process of interpreting any given feature(s) in the representation space of a pre-trained model using a probe dataset $\mathcal{D}$ and a captioning dataset $\mathcal{S}$. Taking the set of highly activating images (from $\mathcal{D}$) for the target features we compute their gradient heatmap (Selvaraju et al., 2019) crops, keeping only the highly activating regions. We compute CLIP (Oikarinen & Weng, 2022) image representations of the cropped images and text representations of a large captioning dataset (in our case, LAION-400m (Schuhmann et al., 2021)). For *contrastive interpretation*, we also caption lowly activating (counterfactual) images. Using cosine similarity, we select the top 5 captions per image and pass them through our concept extraction module (Described in Figure 3).

1. Introduction

Learning generalizable representations has a growing requirement given the considerable cost of pre-training and inference. More importantly, understanding what is encoded in representations is a necessity for deployment, particularly in medical and safety-critical applications (Salahuddin et al., 2021). Large pre-trained self-supervised models (Caron et al., 2021; Chen et al., 2020a;b; Chen & He, 2021) have shown successful generalization capability with frozen representations, however, their representation spaces are still not fully understood. Prior works attempt to understand neural features through detailed visualization of concepts (Olah et al., 2020; 2017; Selvaraju et al., 2019; Zhang et al., 2021; Ghorbani et al., 2019). Visualization (via saliency) helps discover various attributes that neurons react to, but can be noisy and greatly ambiguous requiring manual inspection to achieve any useful explanation. Natural language explanations can complement saliency heatmaps by providing a small number of conceptual keywords that accurately describe the salient component. Text-based explanations of model features can also enable scalable analysis of model interpretability. We can automatically identify concept frequency and sensitivity, their contribution in downstream tasks and debug failures modes. We note that such analysis is not easily possible using traditional interpretation methods involving saliency (gradient heatmaps). One way to

*Equal contribution [1]University of Maryland, College Park [2]Center for Machine Learning, CapitalOne [3]Meta AI. Correspondence to: Neha Kalibhat <nehamk@umd.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

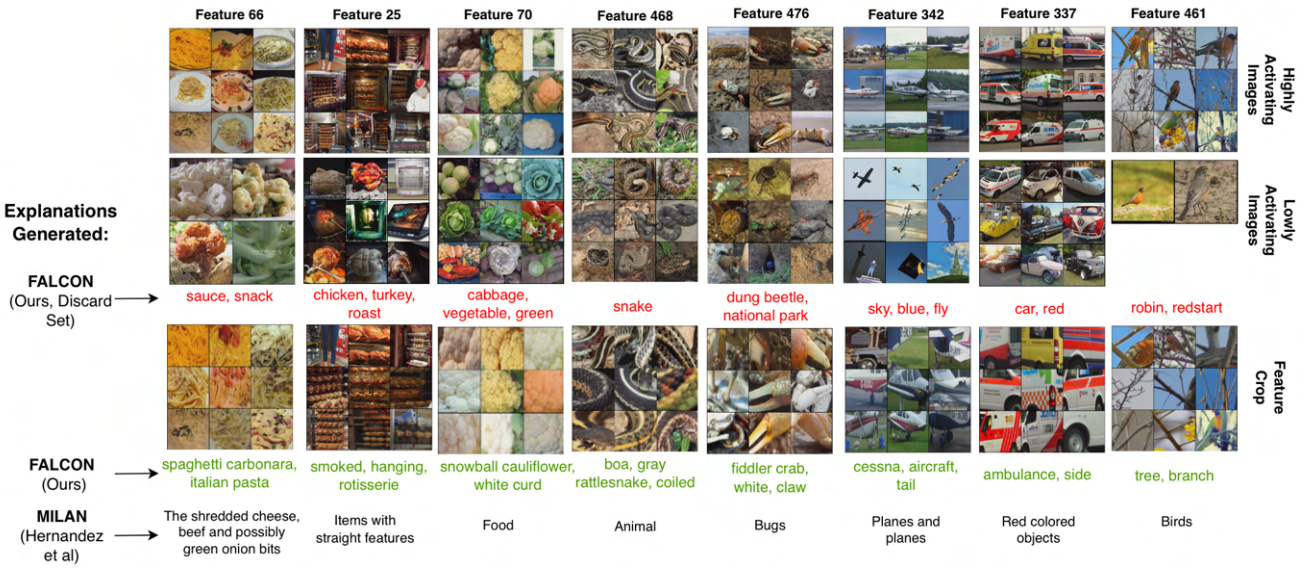

Figure 2. **Concepts extracted by FALCON for various features in the SimCLR representation space:** We explain various features of the final layer representations of SimCLR (Chen et al., 2020a) pre-trained on ImageNet (Russakovsky et al., 2015) with a ResNet-18 (He et al., 2016) backbone (512 features). For each feature, we show the top activating images as well as the lowly activating images. We crop the top activating images to highlight only the activated regions and extract concepts using the approach outlined in Section 2. The lowly activating images are used to filter spurious concepts using our approach called *contrastive interpretation* (See Equation 2).

achieve automatic text-explanation is by using supervision datasets (Bau et al., 2017; Hernandez et al., 2022) with fine-grained conceptual labels for each sample. Such approaches can prove to be expensive, requiring expert annotations. They also may not be generalizable as explanations can be dataset-specific.

In the first part of this paper, we propose **Automatic Feature Explanation using Contrasting Concepts (FALCON)**, a framework to explain neural features, with no densely-labelled dataset or human intervention. We mainly study final-layer self-supervised representations as they contain no label-bias, however, our approach is model-agnostic and can be extended to any deep neural feature. We are also particularly interested in understanding final-layer representations since they alone are accessible to downstream tasks, and their richness and quality is shown to be essential for better generalization (Bordes et al., 2021; Kalibhat et al., 2022; Garrido et al., 2022). Nevertheless, our framework is general and can be extended to explain any layer neurons.

FALCON is described in Figures 1 and 3. For a target feature, we first compute crops of highly activating images from a given dataset (like, ImageNet (Russakovsky et al., 2015)) based on gradient activation. We then caption each cropped image by matching their CLIP (Radford et al., 2021) image embeddings to the closest CLIP text embeddings from a large captioning dataset (like, LAION-400m (Schuhmann et al., 2021)). We collect illustrative captions for each image with high CLIP cosine similarity, without having to train additional captioning models (Hernandez et al., 2022; Wang et al., 2020; Yu et al., 2022; Wiles et al., 2022). The next step in FALCON is described in Figure 3, where we show how a compact set of shared, human-understandable *concepts* are extracted from image captions using *Word Score*. We define concepts as the words which closely relate to the attributes that are likely to be encoded by the target feature, based on the set of cropped highly activating images. Unlike prior methods ((Oikarinen & Weng, 2022)), FALCON is not restricted to output a single concept since features can encode complex physical information which can compose of multiple facets (Mu & Andreas, 2020). We recognize, however, that top-ranking concepts can relate to spurious attributes which may not be true descriptors for the target feature, although the attributes exist in most of the highly activating images. Current interpretability techniques (Oikarinen & Weng, 2022; Hernandez et al., 2022; Bau et al., 2017), tend to produce misguided explanations as they do not account for spuriosity and simply report the highest scoring concept. FALCON eliminates spurious concepts by applying a *contrastive interpretation* technique, where we use lowly activating (counterfactual) images for the target feature whose concepts can be discarded. We therefore produce the minimum sufficient set of concepts that best explain the target. We show the results of successfully annotated features of SimCLR (Chen et al., 2020a) in Figure 2.

In the second part of our paper, we study which features in the representation space can be explained. We observe that

individual features that are very strongly activating for an adequate number of samples can correspond to easily detectable concepts. However, such features constitute a very small portion of the whole representation space. We observe that most features activate a diverse set of images where the hidden concept is not apparent (See Figure 4). We discover that pairs (or groups) of such features are surprisingly more interpretable than individual features. The highly activating images of feature groups are strongly correlated allowing FALCON to produce high scoring concepts. We can therefore explain a much larger portion of the representation space with descriptive and robust concepts.

We evaluate FALCON through human evaluation on Amazon Mechanical Turk (AMT). We show participants images and their FALCON concepts to collect ground truths (relevant or not relevant) for each concept of each annotated feature. The results from our study show a precision of 0.86 and recall of 0.84 for the top-5 concepts, indicating that FALCON concepts are agreeably explanatory (See Section 4).

Since the extracted concepts are unique physical attributes for only the portions that a given feature encodes, we can decompose the content of any given image into a set of concepts corresponding to different elements (See Figure 5). This helps us understand which physical components of the image have been encoded in its representation. This is also not possible with approaches that conceptualize entire images (like (Oikarinen & Weng, 2022)). We further utilize concepts to explain failures, like mis-classifications in downstream tasks (See Figure 6). By discovering the most contributing concepts in classification, we can detect what the model pays attention to while making its prediction and communicate these in terms of human-understandable concepts. This can help practitioners find and debug issues like hard examples, multi-object scenarios and mis-labelled examples.

Finally, we propose an approach to transfer concepts from an explained representation space to a new representation space by learning a simple, linear transformation. We train a linear head that maps representations from a target (unseen) model to the source (interpretable) model. This function lets us map any interpretable feature (or group of features) in the source model to the corresponding feature (or group of features) in the target model, and transfer the extracted concepts. We show that the top activating images of the features in the new representation space, exactly match the transferred concepts from the source representation space (See Figure 7).

We summarize our contributions below:

- We propose Automatic Feature Explanation using Contrasting Concepts (FALCON), an interpretability framework that automatically detects concepts encoded by any feature of image representations, without any labelled datasets or human intervention.
- We show that representation spaces can be largely explained by interpretable feature groups rather than independent features.
- We show that concepts can be used to explain failures in downstream tasks and can be transferred across representation spaces with a simple linear transformation.

2. Automatic Feature Explanation using Contrasting Concepts (FALCON)

2.1. Image Captioning Using CLIP

We discuss the general workflow of FALCON to explain features of vision model representations. Let us consider a pre-trained backbone denoted by $\mathbf{f}_\theta(\cdot)$. For a given input image $\mathbf{x}$, this model outputs a representation vector of size r, i.e, $\mathbf{f}_\theta(\mathbf{x}) = \mathbf{h} \in \mathbb{R}^r$. Any downstream task only utilizes these representation vectors, therefore, our objective is to provide human-understandable explanations for these features.

In order to explain features (different indices in $\mathbf{h}$), we utilize two datasets ; 1) A probing dataset consisting of a diverse set of images ($\mathcal{D}$), and 2) A large text dataset to extract concepts ($\mathcal{S}$). In our experiments, we use ImageNet-1K (Russakovsky et al., 2015) validation set for $\mathcal{D}$ and LAION-400m (Schuhmann et al., 2021) for $\mathcal{S}$, however, the framework of FALCON is general and can be used with other datasets as well.

Let us consider the task of explaining the $i^{th}(0 \leq i \leq r)$ feature in the representation space of a pre-trained vision model $f_\theta(\cdot)$. From the probing dataset, $\mathcal{D}$ of size N, we first extract the set of highly activating images for feature i defined by, $\mathcal{T}_i = \{j : h_{ji} > \alpha, 1 \leq j \leq N\}$, where α is a threshold we empirically select (more discussed in Section 3). As shown in Figure 1, for SimCLR (Chen et al., 2020a) with a ResNet-18 (He et al., 2016) backbone, the set of highly activating images for feature 10 are images of a certain breed of dogs. We next compute the gradient of feature i with respect to these images using GradCAM (Selvaraju et al., 2019) as shown. We crop the images keeping only the maximally activating portions by thresholding the GradCAM mask. This set of cropped images as well as a large scale text dataset ($\mathcal{S}$) like LAION-400m, serve as the input to a pre-trained vision-language model, i.e., CLIP (ViT-B/32) (Radford et al., 2021). LAION-400m is a large, diverse image captioning dataset which has been used to pre-train vision-language models like CLIP.

We define the CLIP text encoder as $g_{tx}(\cdot)$ and image encoder

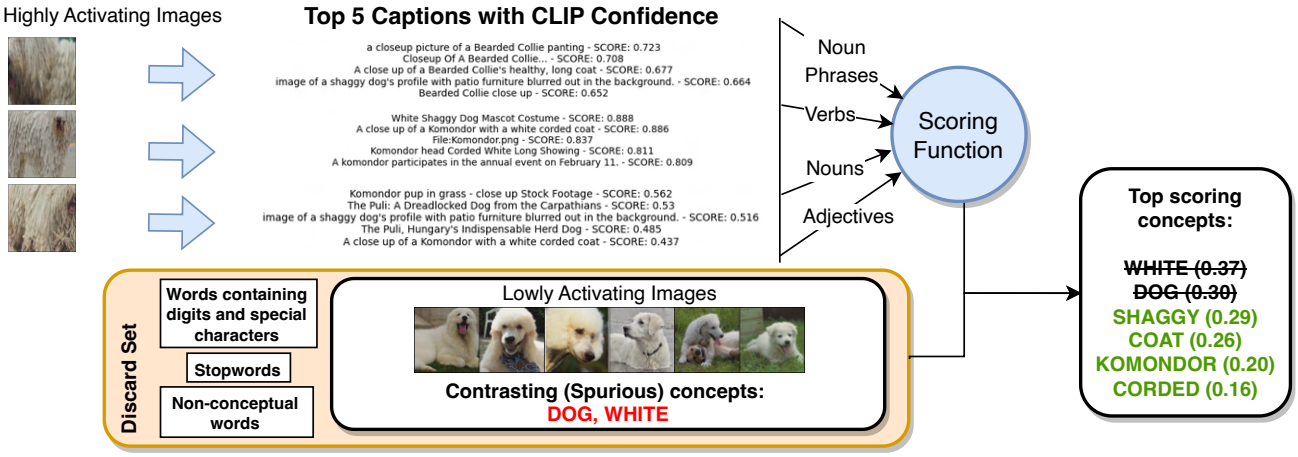

Figure 3. **Concept extraction in FALCON using contrasting concepts:** We extract a bag of words (nouns, verbs, adjectives) from the top 5 captions (from LAION-400M (Schuhmann et al., 2021)) of every image in the set of highly activating images of a given feature. We use a scoring function (Equation 1) to extract top scoring words and phrases which we refer to as *concepts*. We also apply *contrastive interpretation* where we discard any concept that is extracted from the lowly activating images (mined through Equation 2). In this case, "dog" and "white" are spurious concepts that exist in both highly and lowly activating images, implying that they are not discriminative explanations. Therefore, final set of discriminative concepts include "shaggy", "coat", "komondor" and "corded" which are all closely related to the given image set.

as $g_{im}(\cdot)$. Given our captioning dataset ($\mathcal{S}$) of size M, we extract the text embedding matrix denoted by $A \in \mathbb{R}^{M \times k}$ where k is the size of the CLIP text embedding space. Since our captioning dataset is fixed for interpreting any feature, we only need to compute its embeddings once. In fact, LAION also provides pre-computed text embeddings on CLIP which saves compute time significantly. We next compute the image embeddings of the cropped highly activating images of feature i denoted by $B \in \mathbb{R}^{|\mathcal{T}_i| \times k}$. Using A and B, we compute the CLIP confidence matrix, which is essentially the cosine similarity matrix, denoted by $C = BA^T \in \mathbb{R}^{|\mathcal{T}_i| \times M}$. Note that both text and image embeddings are L2-normalized before computing C. Using C, we extract the top 5 captions for each image in $\mathcal{T}_i$.

2.2. Contrastive Concept Extraction

The second component of FALCON involves extracting concepts out of the captioned batch of highly activating images for the given feature. In Figure 3, we show the top-5 concepts for three highly activating images along with the CLIP confidence. From each caption, we extract the noun phrases, nouns, verbs and adjectives to form a bag of words. Verbs and adjectives are extracted to qualify complex concepts which cannot be described with nouns alone. We remove all stop words and words containing digits or special characters from the bag. We also prepare a discard word set including general, non-conceptual words like "photo", "picture", "background" etc. Given a word w, the word confidence for the p^{th} caption in the q^{th} image is given by, $C^w_{q,p}$ if the word exists in the caption, otherwise 0. We get the maximum value of $C^w_{q,p}$ for each image (q). The *Word Score* is defined as,

$$\textit{Word Score}^w = \frac{1}{|\mathcal{T}_i|} \sum_{q=1}^{|\mathcal{T}_i|} \max_p C^w_{q,p} \tag{1}$$

Word Score gives a normalized score for every word among the captions we extract. The best shared concepts describing a given feature i are the highest ranking words, by applying a threshold (in practice, 0.08).

Contrastive Interpretation: In practice, the above method of concept extraction provides a number of high-scoring keywords, shared between the highly activating images. However, in many cases, these keywords can be too general or related to high-level spurious attributes which may be common to all the activating images but not necessarily relevant to the feature we want to interpret. Many existing techniques (Oikarinen & Weng, 2022; Bau et al., 2017; Hernandez et al., 2022; Mu & Andreas, 2020), do not account for such cases and they only report a single best scoring concept. In FALCON, we overcome this issue by discovering images in $\mathcal{D}$ that share all other concepts with the highly activating images of feature i, except the actual concepts that feature i encodes. We refer to these images as *lowly activating counterfactual images*. The concepts extracted out of lowly activating images can be regarded as spurious concepts for feature i and added to the discard set.

Let us define the set of feature indices without the index i as $\mathcal{V}_i = \{j : 0 \leq j \leq r, j \neq i\}$. The mean representation of the highly activating images ignoring the i^{th} feature can be

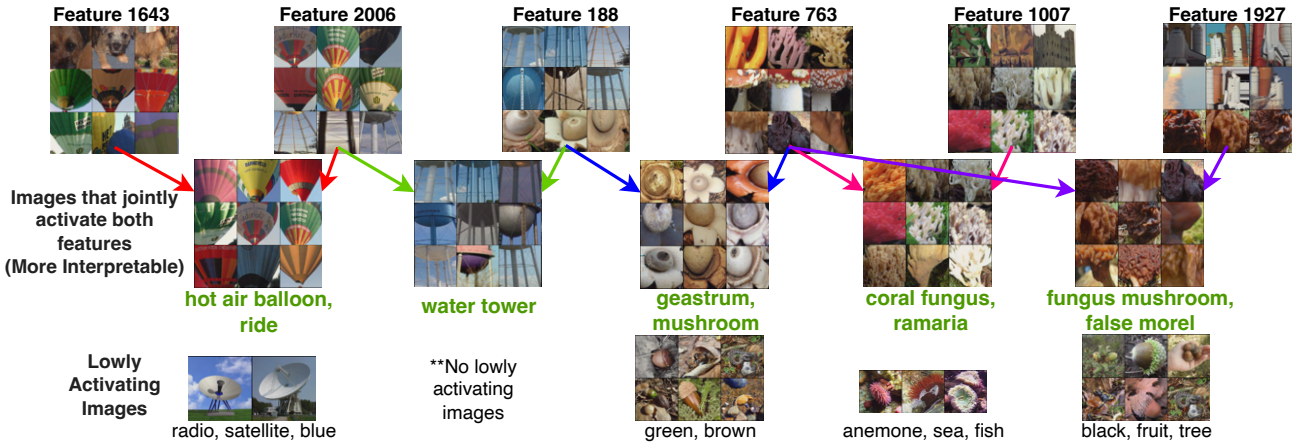

Figure 4. **Groups of features can be more interpretable than individual features:** In the first panel, we show the highly activating images of some features of DINO (Caron et al., 2021) representations trained on ImageNet (Russakovsky et al., 2015) with a ResNet-50 (He et al., 2016) backbone. We observe that the images are highly diverse with seemingly no shared concept, like "mushrooms" and "water towers" in feature 188. In the second panel, we observe that images that highly activate pairs of features are significantly more connected. The concepts that our framework extracts are strongly correlated to each group of images. For each feature group, we use the lowly activating images (mined from Equation 2) to filter out spurious concepts.

written as $\mathbf{h}^{\mu} = mean_{\mathcal{T}_i}(\mathbf{h}_{\mathcal{T}_i,\mathcal{V}_i}) \in \mathbb{R}^{|\mathcal{V}_i|}$. The set of lowly activating images for the target feature i is given by,

$$\mathcal{L}_i = \{j : h_{ji} < \epsilon, \mathbf{h}_{j,\mathcal{V}_i} \cdot \mathbf{h}^{\mu} \geq \beta, 0 \leq j \leq N\} \quad (2)$$

where β and ϵ are limits we select empirically. In our experiments, ϵ is the mean value of that feature across the population of normalized representations. Since β is used to threshold the dot product of representations (excluding the target feature), a larger value for β would give us true counterfactuals. We therefore select β to be 0.7. This method of conditional selection gives us lowly activating images that contain all the concepts in the highly activating image set, except the concept represented by the i^{th} feature. We apply FALCON (without feature cropping) to extract concepts out of the lowly activating image set. As shown in Figure 3, concepts like "dog" and "white" are in lowly activating images. These keywords can be relevant to the highly activating image batch as well, however, they are not discriminative explanations for that feature. Therefore, we include the concepts of lowly activating images in the discard set and arrive at the final minimum sufficient set of concepts "shaggy", "coat", "komondor" and "corded".

In Figure 2, we show the extracted concepts from FALCON for 8 different features of SimCLR on a ResNet-18 backbone. In cases like Feature 337, the lowly activating images match almost all the object properties i.e, vehicle or van. However, after extracting concepts, it becomes clear that the feature concept is the side view of an emergency vehicle which is explained by - "ambulance" and "side". Contrastive interpretation therefore lets us ignore generic and spurious concepts to derive a compact set of discriminative explanations. We also compare FALCON with MILAN (Hernandez et al., 2022), a recent approach that trains a generative model on a human-annotated fine-grained image region-caption dataset, and uses this model to generate natural language explanations. We observe that FALCON produces more feature-specific concepts compared to the generic high-level explanations of MILAN. We show more annotated features (including supervised and previous-layer features) comparing with MILAN in the Appendix (See Figures A.7, A.8). We also discuss the generalizability of concepts to an unseen dataset like STL-10 (Coates et al., 2011) (See Figure A.7).

3. Which Features are Explainable?

So far we discussed our method to explain individual features, given the representation space of a pre-trained model. In this section, we understand which features in the representation space can be considered as *explainable*. Let us go back to the set of highly activating images for a given feature i, defined by $\mathcal{T}_i = \{j : h_{ji} > \alpha, 1 \leq j \leq N\}$. Note that the representations are all L2-normalized. In order to extract meaningful and generalizable shared concepts with high Word Scores, we require a sufficient number of highly activating images. In our experiments, we select the features where $|\mathcal{T}_i| > 10$. If α is large enough, we may expect the set of highly activating images to be more connected where the feature concept is clearly detectable (See features in Figure 2).

We choose features with a strong value for α according to the distribution of the representation space of the selected model. The features where $|\mathcal{T}_i| > 10$, only comprises of roughly 20% of the representation space. See Table A.3 for

this percentage for various pre-trained models. Upon empirical inspection of the activated images, we find that thresholding α alone, may not guarantee explainability. Some of the features can correspond to human recognizable concepts (activating correlated images), like the examples shown in Figure 2. While other features, although strongly activated for a sufficient number of samples, correspond to very high level, abstract concepts that are not apparent to humans. We show examples of such features in the top panel of Figure 4, on DINO (Caron et al., 2021) with a ResNet-50 backbone. Although these features are activating with high α, the images are quite diverse, making it almost impossible to decipher any shared properties. One possible way to understand such features could be by explaining previous layer neurons in the network which may perhaps encode higher level properties (Oikarinen & Weng, 2022; Mu & Andreas, 2020; Hernandez et al., 2022; Bau et al., 2017). This is however computationally inefficient as previous layer features may still activate dissimilar image sets or may correspond to entirely different concepts.

In the second panel of Figure 4, we make a key observation; images that jointly activate a given pair of features are significantly more related and explainable than those of individual features. For example, visually, we cannot identify any shared property between rockets and morel mushrooms in feature 1927 and similarly, fly argaric mushrooms and underwater coral plants in feature 763. However, when both feature 763 and 1927 are highly activated, the shared concepts become more apparent, showing only morel mushroom textures. When the same feature 763 is jointly activating with another feature like 1007, it corresponds to a totally different concept of coral reef patterns. A similar observation has been made in (Elhage et al., 2022; Fong & Vedaldi, 2018). Note that the threshold for α is the same for both individual and groups of features (for fair comparison in Figure 4), however, less rigorous α can still be used for groups of features. By observing highly activating images for a combination of features, we can explain a larger portion of the representation space (even by relaxing α) compared to independent features.

Automatically discovering all interpretable feature groups: Given a model $f_\theta(.)$ and a probe dataset $\mathcal{D}$ of N samples, we compute the top activating set of features (group) for every sample (using α as the threshold). We save each feature group and the indices of the samples that highly activate that group. We use the average CLIP cosine similarity of the samples within each group to decide if a group is interpretable or not (using a threshold, γ). A higher value for average similarity implies that the top activating samples are *similar* with interpretable shared concepts. Other metrics LPIPS (Zhang et al., 2018) can also be used. In Algorithm 1, we provide PyTorch-like code highlighting the steps required for identifying all possible interpretable feature groups in the representation space of a given model.

FALCON can be used to extract concepts out of groups of features in the same manner as individual features, with some modifications. First, the feature crop is calculated by taking the intersection of the gradient heat map of each feature individually as shown in Figure 4. Second, the lowly activating images are mined such that all the features in the group show low activation and the remaining features are close to that of the highly activating representations. That is, Equation 2 is updated to compute $\mathcal{L}_\mathcal{I}$ where $\mathcal{I}$ represents the feature group. As shown in Figure 4, FALCON uses the lowly activating images to help in finding discriminative concepts for groups of features that best explain the highly activating images.

In Appendix Section A.1, we analyze the extracted concepts across various models (supervised and self-supervised) and discuss some key insights.

Algorithm 1: Pytorch-like pseudocode for discovering interpretable feature groups in a given representation space

Input: **H** is the set of representations (of the given model $f_\theta(.)$) of N samples in the probing dataset $\mathcal{D}$,
α is a threshold for feature activation,
γ is a threshold for interpretable feature groups.

```
# Identify all feature groups
groups = {}
for j in range(N):
    group = torch.where(h[j] > alpha)
    groups[group].append(j) # groups[group] is
     a list
# Filter out interpretable groups
int_groups = {}
for group in groups:
    if len(groups[group]) > 10:
        # top activating samples for group
        top_act_idx = groups[group]
        clip_feat = get_clip_feat(top_act_idx)
        avg_cos = torch.matmul(clip_feat,
         clip_feat.T).mean()
        if avg_cos > gamma:
            int_groups[group] = groups[group]
return int_groups
```

4. Evaluating Extracted Concepts

FALCON produces a simple, compact set of concepts to describe any explainable feature in an automatic fashion without any human intervention, or densely-labelled datasets. We performed a human study on Amazon Mechanical Turk (AMT) to evaluate the concepts generated by FALCON and provide some quantitative metrics. In each task, we showed the AMT participant the set of highly activating cropped image set (Group A) and the lowly activating image set (Group B) for a target feature and, the top 6 concepts ranked by FALCON. We asked the participant to - identify the concepts that are related to Group A and not Group B. This lets us assign binary ground-truth labels to each concept as 0 (not related) if it has been chosen by at least 65% of the

participants and 1 (related) otherwise. We can partition the six FALCON concepts for each feature based on their rank such that the first K concepts where $1 \leq K < 6$ can be predicted as 1 (related), otherwise 0 (not related). In Table 1, we plot the Precision and Recall for each K. Precision in our case measures how many of the "related" concepts predicted by FALCON are actually related according to our human study. Recall measures how many "related" concepts was FALCON able to predict among the total number of related concepts (from our human study). We observe that the Recall improves from the 4^{th} caption, meaning that, the participants agree that the first 4-5 concepts are related to the given set of images. 84.23% (precision at top-6) of all FALCON concepts are considered relevant by our participants. This study confirms that the top ranking concepts generated by FALCON are considered relevant and explainable among humans. We collect ground truths for 600 concepts each from 3 participants. We measure the agreement between participants for each feature by averaging the % overlap of the concepts selected by each participant. The average agreement among the participants is 79%. More details about our human study can be found in Appendix Section A.4.

Existing methods (MILAN (Hernandez et al., 2022), Net-Dissect (Bau et al., 2017)) use human-annotated datasets for natural language descriptions. Through FALCON, we automatically extract a minimal sufficient set of noise-free concepts with no human intervention. We performed another user-study on MTurk to provide a quantitative comparison of FALCON with MILAN and Net-Dissect. We display the highly (Group A) and lowly (Group B) activating images for each target feature and requested the participants to select the concept set which best describes the images in Group A but not Group B. We tabulate the percentage of times the concept set of each framework was selected as the best explanation in Table 2. FALCON performs significantly better than the baselines in our study of 115 features.

Table 1. **Precision and Recall for human evaluation of top K concepts:** Using Amazon Mechanical Turk (AMT), we ask human participants to choose the un-related captions, among 6 top-ranking captions for each feature. We use the annotations as ground truth labels (relevant or not relevant) and compare them to the predictions of FALCON at different levels of K (number of predicted concepts labelled as relevant).

Top K Concepts	Precision (%)	Recall (%)
1	94.62	18.72
2	92.47	36.60
3	88.89	52.77
4	86.82	68.72
5	85.60	84.68
6	84.23	100.00

Table 2. **Comparing explanations generated by FALCON with existing baselines:** We request participants to select the best explanation generated by the following 3 frameworks for a given set of highly and lowly activating images. FALCON beats other baselines by a significant amount.

Framework	% of times selected as best explanation
FALCON	**86.40**
MILAN (Hernandez et al., 2022)	13.47
Net-Dissect (Bau et al., 2017)	0.12

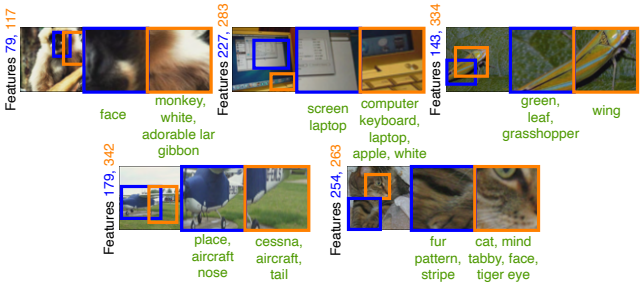

Figure 5. **Decomposing images into various concepts:** We show some images which highly activate multiple interpretable features. FALCON extracts concepts from feature crops rather than entire images, therefore, each image can be broken down into components, each describing a different physical attribute.

5. Explaining Failure Modes in Vision Models

An interpretable representation space of a given model, allows us to decompose and label different groups of concepts in any given image. In the previous sections, we found that each interpretable feature (or group of features) encodes only a portion of images that correspond to a unique concept set. Therefore, images with multiple highly activating features (Kalibhat et al., 2022), can be decomposed into multiple components, each representing a unique concept. We illustrate this in Figure 5, where we show the feature crop of highly activating features (of SimCLR with ResNet-18) in each image, and their corresponding physical concepts that our framework has extracted. This is only possible because FALCON uses feature crops to discover concepts rather than whole images (unlike CLIP-Dissect (Oikarinen & Weng, 2022)).

Another advantage of feature specific concepts is the ability to explain failures in downstream tasks. It is often not obvious what led to a model's prediction without some qualifying explanations. Images in the real-world could contain several spurious attributes interfering with the main content of the image. In such cases, it can be difficult for even experts to localize the exact reason for mis-classification. Moreover, it is often too tedious to have humans make guesses as to what could be the reason for failures as each human can interpret images in a unique manner. With our

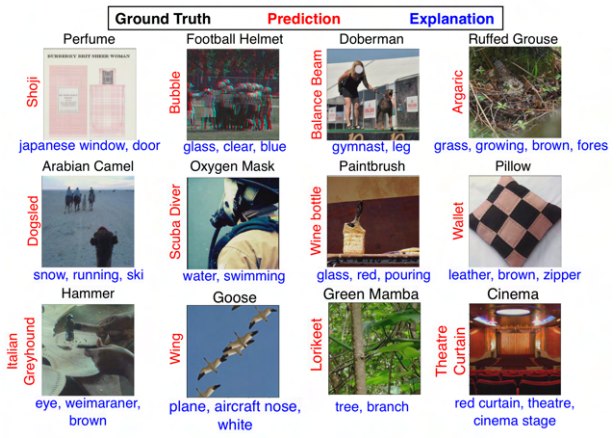

Figure 6. **Explaining failures in downstream tasks using concepts:** Given SimCLR (Chen et al., 2020a) pre-trained on ImageNet (Russakovsky et al., 2015), we show some mis-classified examples along with the most contributing concepts for their prediction. This allows us to detect and explain concepts which contributed to a model's decision and help us debug model failures.

automatic explanation framework, FALCON, we eliminate this need for human-in-the-loop and can inspect grounded explanations directly.

We consider the task of classification using a linear head defined by the weight matrix $\mathbf{U} \in \mathbb{R}^{o \times r}$, where o is the number of classes. The most contributing features (and corresponding concepts) for a sample $\mathbf{x}_j$ with prediction y_j, can be given by, $\arg\max(\mathbf{h}_j \odot \mathbf{U}_{y_j})$. In Figure 6, we show some mis-classified examples of SimCLR trained on ImageNet and the most contributing concepts for each prediction. The concepts we find add novel insight into model behavior apart from the readily available information i.e., the image, label and prediction. They help describe the attributes to which model paid attention while making its prediction, potentially helping us automatically debug models at inference time.

For example, the eighth "Goose" image, looked more like an aircraft to the model, leading to the prediction "Wing". This is an example where the model may be spuriously associating shape (like an aircraft) and background information (like the sky) in making its prediction, failing to identify the subtle features of geese. The texture in the "Perfume" and Football Helmet" images is also an example of spurious attributes. The sixth "Green Mamba" image can be regarded as a *hard example*, where the core object is largely hidden, causing the model to focus more on concepts like tree and branch. Explanations can also help uncover images which may have multiple ground truths like the eleventh example of "Cinema" and "Theatre Curtain" (similar to the images in Figure 5). The "Pillow" and "Hammer" images indicate that the training paradigm of the model ignored global object information and made decisions based on local attributes. One possible approach to improve such models relying on spurious correlations is by fine-tuning on synthetic images generated using the relevant FALCON concepts via methods like Stable Diffusion (Rombach et al., 2021). Explanations can also help define optimal training augmentations that could prevent spurious dependencies.

6. Transferring Concepts to New Representation Spaces

So far, we have discussed the process of feature captioning and concept extraction for a given vision model. We hypothesize that the representations learned by different models can be mapped from one to another. This would allow us to map the features of an explainable representation space to any unseen representation space, without having to re-run our explanation framework. Let us consider the representations of a model that we have extracted concepts for, denoted by $\mathbf{H}_{source} \in \mathbb{R}^{N \times r}$. The representation space of an unseen model can be denoted by $\mathbf{H}_{target} \in \mathbb{R}^{N \times r}$. Using a linear head, our goal is to learn a transformation matrix $\mathbf{Z} \in \mathbb{R}^{r \times r}$, that transforms $\mathbf{H}_{target}$ to $\mathbf{H}_{source}$, by solving the optimization,

$$\min_{\mathbf{Z}} \|\mathbf{Z}^T \mathbf{H}_{target} - \mathbf{H}_{source}\|_2 \tag{3}$$

We solve optimization by training a linear head for only 10 epochs with a learning rate of 1, using an SGD optimizer. Once the mapping is learned, we can take any explainable feature i in $\mathbf{H}_{source}$, and find the features in $\mathbf{H}_{target}$ with have the highest weights in $\mathbf{Z}$. Hence, the concepts described by feature i in $\mathbf{H}_{source}$ can be mapped to features in $\mathbf{H}_{target}$ efficiently.

We confirm that this transformation works by matching the concepts of $\mathbf{H}_{target}$ to the highly activating images of $\mathbf{H}_{target}$. As shown in Figure 7, we successfully map individual interpretable features in SimCLR to features in MoCo (Chen et al., 2020b) which is an unseen representation space. The highly activating images in MoCo interestingly contain all of the concepts of the source feature. Note that, features across representation spaces need not have a 1:1 relationship. Similarly, concepts can also correspond to compositional features (as described in Section 3). We do not constrain the sparsity of $\mathbf{Z}$. In practice, $\mathbf{Z}$ is not sparse however, it can be considered as *nearly sparse* where most weights are close to zero. When we discover feature maps, we only extract the weights in $\mathbf{Z}$ if they are large enough ($> \text{mean} + 4 \times \text{std}$ based on the weight distribution). If we do so, Z becomes quite sparse, indicating that some directions in the target model can be mapped to a dedicated set of features in the source model.

This observation of transferrability can potentially be ex-

tended to any pair of pre-trained models (supervised or unsupervised), preventing the need to interpret the representations of each specific model. It also gives us an important insight that various vision models, regardless of their pre-training regime, learn mostly similar concepts. To the best of our knowledge, ours is the first approach in the direction of transferring explanations across model spaces.

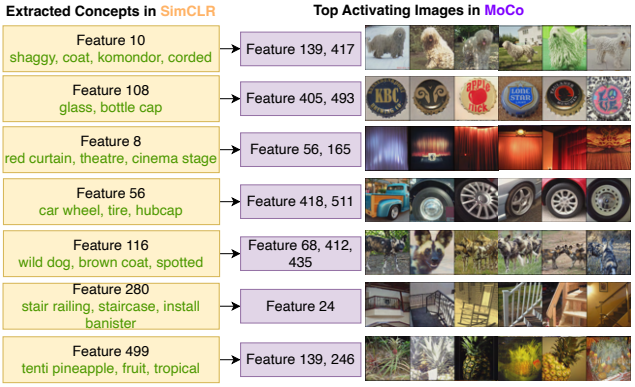

Figure 7. **Transferring concepts from an explained representation space to an unseen representation space:** We show that representations of self-supervised models can be mapped from one to another by learning a transformation $\mathbf{Z}$ (See Equation 3). We transfer extracted concepts from SimCLR (source) (Chen et al., 2020a), to an unseen model, MoCo (target) (Chen et al., 2020b) by mapping the source features to the target features with the highest weights in $\mathbf{Z}$. We observe that the top activating images of the mapped features in the MoCo very closely match the concepts extracted in SimCLR.

7. Conclusion

We proposed FALCON, an automatic framework to explain individual neurons in vision models. These explanations can be utilized for classification tasks (as shown in Figure 6) as well as non-classification tasks like object detection and segmentation. We show that features become more interpretable when regarded in groups and propose a simple algorithm to discover all possible interpretable groups in a given representation space.

FALCON utilizes three components: 1) A probe image dataset, 2) A large text vocabulary and 3) An off-the-shelf pre-trained vision-language encoder. With FALCON we propose a general-purpose framework, where the above components can be flexibly customized depending on the target model we wish to investigate. The concepts learned by the target model is governed by the data it was trained on. In order to explain these concepts via FALCON, we choose the probe image dataset and text vocabulary such that it is representative of the target model's training domain and encapsulates all the concepts learned by the target model. In our experiments, we use FALCON with ImageNet, LAION-400M and CLIP to explain deep models pre-trained on ImageNet-1K. These components could potentially generalize to a range of domains, since CLIP is already pre-trained on a very large scale and LAION-400M is diverse and expressive. FALCON can be updated to use even larger zero-shot vision-language encoders and vocabulary, when developed in future. To deploy FALCON on target models trained on medical images like chest x-rays, we can utilize vision-language encoders like ConVIRT (Zhang et al., 2020) or MedCLIP (Wang et al., 2022), combined with expressive vocabulary from radiology reports (ex. Mimic-cxr).

Limitations and directions for future work: Understanding how FALCON can be applied to explain vision-language models can be an extension of our work. Since vision-language models are trained to align representations in the vision and language space, we could potentially learn a great deal about the model's understanding by applying FALCON directly on the vision encoder. It would however be interesting to understand what information is represented uniquely by the text encoder. Understanding the equivalent of localized gradient heatmaps in the language space is still unclear and requires further research.

Supporting FALCON for non-image domains remains a topic for further research. Another limitation of FALCON is the requirement of a pre-trained vision-language model for the task of matching images to captions. While CLIP is trained on very diverse data and domain-specific versions of CLIP exist, there may be target models which are trained for uncommon tasks and data, that is unknown to CLIP. In our transferrability example, we show that concepts extracted from one model may be transferred to another by learning a simple linear transformation. Another important direction for future work is to test the limits of transferrability on multi-domain setups. For example, how do concepts learned by a model trained on painting images, transfer to a model trained on sketch images.

8. Acknowledgment

This project was supported in part by Meta grant 23010098, NSF CAREER AWARD 1942230, HR001119S0026 (GARD), ONR YIP award N00014-22-1-2271, Army Grant No. W911NF2120076, NIST 60NANB20D134, the NSF award CCF2212458, a CapitalOne grant and an Amazon Research Award. The authors would also like to thank Mazda Moayeri, Hemant Kumar, Samyadeep Basu, Sharath Gokarn, Saksham Suri, Pavan Gurudath, Nirat Saini, Pulkit Kumar and Archana Swaminathan for their help with testing our human studies.

References

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.354. URL http://dx.doi.org/10.1109/CVPR.2017.354.

Bordes, F., Balestriero, R., and Vincent, P. High fidelity visualization of what your self-supervised representation knows about, 2021.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020a. URL http://proceedings.mlr.press/v119/chen20j.html.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning, 2020b.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/coates11a.html.

da Costa, V. G. T., Fini, E., Nabi, M., Sebe, N., and Ricci, E. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022. URL http://jmlr.org/papers/v23/21-1155.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition, 2022.

Fong, R. and Vedaldi, A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00910. URL http://dx.doi.org/10.1109/CVPR.2018.00910.

Garrido, Q., Balestriero, R., Najman, L., and Lecun, Y. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank, 2022.

Ghorbani, A., Wexler, J., Zou, J., and Kim, B. Towards automatic concept-based explanations, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., and Andreas, J. Natural language descriptions of deep visual features, 2022.

Kalibhat, N., Narang, K., Firooz, H., Sanjabi, M., and Feizi, S. Measuring self-supervised representation quality for downstream classification using discriminative features, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Mu, J. and Andreas, J. Compositional explanations of neurons, 2020.

Oikarinen, T. and Weng, T.-W. Clip-dissect: Automatic description of neuron representations in deep vision networks, 2022.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., and Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods, 2021. URL https://arxiv.org/abs/2111.02398.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336–359, Oct 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Wang, Z., Feng, B., Narasimhan, K., and Russakovsky, O. Towards unique and informative captioning of images. *Lecture Notes in Computer Science*, pp. 629–644, 2020. ISSN 1611-3349. doi: 10.1007/978-3-030-58571-6_37. URL http://dx.doi.org/10.1007/978-3-030-58571-6_37.

Wang, Z., Wu, Z., Agarwal, D., and Sun, J. Medclip: Contrastive learning from unpaired medical images and text, 2022.

Wiles, O., Albuquerque, I., and Gowal, S. Discovering bugs in vision models using off-the-shelf image generation and captioning, 2022.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models, 2022.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

Zhang, R., Madumal, P., Miller, T., Ehinger, K. A., and Rubinstein, B. I. P. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11682–11690, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i13.17389. URL http://dx.doi.org/10.1609/aaai.v35i13.17389.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text, 2020.

A. Appendix

We use pre-trained models from the solo-learn package (da Costa et al., 2022) and the official implementation of CLIP (Radford et al., 2021).

Table A.1. **Feature groups and concepts for various models:** We tabulate the number of interpretable groups for each model and number of unique concepts extracted after explaining each group. We observe that many frequently occurring concepts are shared across models.

Model	# interpretable groups	# unique concepts	Most frequent concepts
SimSiam	249	578	'white', 'head', 'brown', 'eye', 'face'
SimCLR	293	676	'white', 'head', 'face', 'brown', 'blue'
MoCo	271	559	'white', 'head', 'face', 'eye', 'black'
SwaV	182	417	'head', 'brown', 'white', 'hand', 'black'
BYOL	281	477	'head', 'white', 'brown', 'eye', 'face'
ResNet-50 (Sup)	91	183	'brown', 'head', 'red', 'water', 'white'

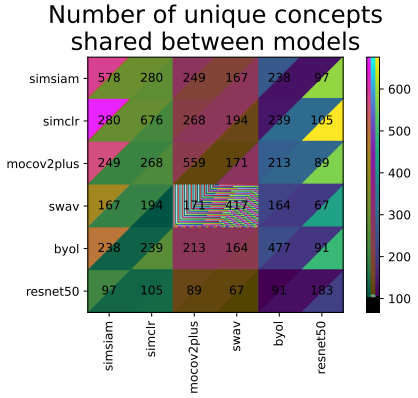

Figure A.1. **Shared concepts between models:** Among the unique concepts extracted using FALCON on the representation space of various models, we plot the number shared concepts between each pair of models.

A.1. Analyzing FALCON Explanations Across Various Models

We have performed a global analysis comparing the FALCON concepts across various supervised and self-supervised models (ResNet-50 encoder). In Table A.1, we tabulate the number of interpretable feature groups identified from the final representation layer, along with the total number of unique concepts extracted from FALCON for these groups. Note that each explanation consists of multiple conceptual words. In the last column, we also list the most frequently occurring concepts for each model. We observe that among all the models we study, the supervised ResNet-50 model has the least number of interpretable groups and unique concepts. The most frequent concepts among all the models are almost identical, including general attributes like various colors, face, eye, which frequently occur in the ImageNet dataset. We also compute the number of shared concepts between each pair of models in Figure A.1. We observe that each model shares roughly less than 50% of its total concepts with any other model. This indicates that although each model is trained on the same data i.e, ImageNet, their training paradigms can enable them to encode some unique properties that are missed by other models. We calculate the number of concepts in each model that are not shared with any other model - SimSiam 160, SimCLR 210, MoCo 159, SwaV 128, BYOL 116, ResNet-50 39. For example, these are some unshared (truly unique) concepts of ResNet-50 - 'eel', 'disc', 'grip', 'shooter', 'tub', 'sink', 'weimaraner', 'decal'.

Table A.2. **Comparing FALCON used with CLIP and LAION-400M vs BLIP-2 zero-shot captioning:** We apply FALCON with BLIP 2 (Li et al., 2023) generated captions and ask participants to select the better explanation when compared with CLIP+LAION. BLIP captions underperform compared to CLIP+LAION.

Framework	% of times selected as best explanation
FALCON + CLIP + LAION	**58.12**
FALCON + BLIP 2 (OPT, caption COCO)	41.8

A.2. Employing a Captioning Model instead of CLIP

BLIP-2's (Li et al., 2023) zero-shot image captioning is a powerful tool to extract text captions out of highly activating images. One advantage of using a separate vocabulary with a vision-language model is the flexibility of controlling the expressiveness/specificity of the captioning dataset depending on the complexity of the target model. For example, to explain an MNIST-trained model, one may use a much smaller vocabulary whereas explaining a model like CLIP may require an equivalently large vocabulary. Moreover, the set of reference captions can be updated online, even after deployment without having to re-train any model. The similarity matrix allows us to extract multiple captions per image with a confidence score, allowing us to discard unreliable captions. Off the shelf captioning models may be domain-specific and could generate noisy captions with low expressiveness.

We compared FALCON + BLIP 2 with FALCON + CLIP + LAION in an MTurk evaluation over 91 features and asked participants to select the best describing explanation for the displayed set of images (See Table A.2. Explanations generated via our CLIP+LAION captioning outperforms BLIPs captioning, however, BLIP 2 is still a practical alternative given that it is trained on a large scale on LAION.

A.3. Interpretable Features in Various Models

We discuss in Section 3 that to discover potentially explainable features we can apply a strong value for α in $\mathcal{T}_i$, set of highly activating images. Since the distribution of each model representation space can be different, to be consis-

Table A.3. **Percentage of highly activating features in the ResNet-50 representation space:** For different model representations, we tabulate the percentage of features that activate at least 10 samples with a magnitude greater than α. We select α according to the mean of the distribution of the representation space (See Section 3 for more details).

Model	ResNet-50 (Supervised)	SimCLR	MoCo	DINO	BYOL	SimSiam	SwaV
% features that highly activate $>$ 10 samples	0.68	21.92	17.70	16.66	25.63	8.00	6.00
α	0.27	0.34	0.34	0.14	0.24	0.32	0.31

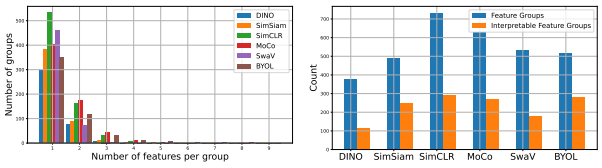

Figure A.2. **Distribution of feature groups:** For different self-supervised model representation spaces we compute the feature groups (from Algorithm 1. On the left we plot group sizes against the number of groups and on the right, we plot the number of interpretable groups among the discovered feature groups.

tent we select $\alpha = mean(\mathbf{H}) + 16 \times std(\mathbf{H})$ (where $\mathbf{H}$ is the representation matrix). In Figure A.3, we tabulate the percentage of highly activating features in the final-layer representations where $|\mathcal{T}_i| > 10$. ResNet-50 has a particularly low number of highly activating features compared to self-supervised baselines. The remaining features in the representation space (or by relaxing α to be less rigorous), do not activate a resembling set of images, making such features harder to explain (Figure 4). Some more examples of such features are shown in Figure A.3.

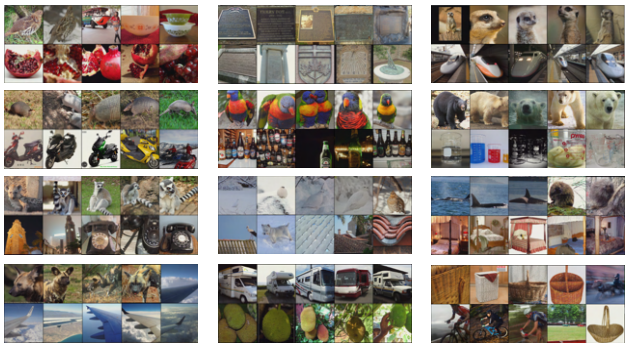

Figure A.3. **Examples of top activating images of some unexplainable independent features:** We provide more examples of top activating images of some independent features of DINO (Caron et al., 2021) (ResNet-50 (He et al., 2016)) representations. The image sets are not correlated in any sense, making it hard to discover shared concepts for these features.

We also discussed in Section 3 that simply thresholding by α does not guarantee explainability as the top activating images can still be unrelated. A larger portion of the representation space can be explained with feature groups. Using the Algorithm 1, we discover feature groups and interpretable feature groups for various self-supervised models. In Figure A.2, on the left, we show the distribution of feature groups and their size. All the identified groups contain at least 10 highly activating images. A large percentage of feature groups contain 1-2 features per group, however, there also exist feature groups that contain up to 9 features. On the right, we compare the feature groups and the *interpretable* feature groups, according to Algorithm 1. The interpretable groups activate samples that are more similar (based on CLIP cosine similarity) and are therefore easy to explain with shared natural language concepts.

A.4. Human Study to Evaluate Concepts

Eliminating malicious and inadequate responses: In our studies, we only select participants that have a HIT approval rate of greater than 90 and the number of HIT approvals is $>$ 500 in the past. Each task is active for 30 minutes allowing the participants ample time to make their selections. We did not explicitly include control questions, however, we identified a small number of tasks which had low-quality concept sets which we used to verify the reliability of the participants. We also approve and pay the participants only after verifying their annotation quality.

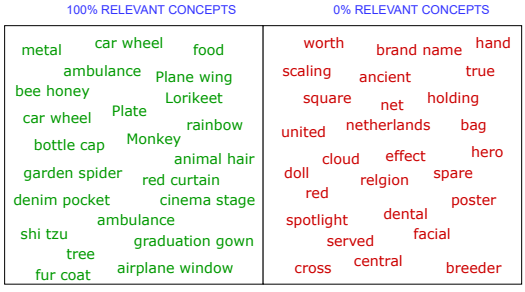

Figure A.4. **Comparing most and least relevant concepts based on AMT study:** We display the concepts with 100% relevancy agreement on the left and the concepts with 0% relevancy agreement on the right.

In Figure A.5, we show a template of our user study where we display two groups i.e., highly activated cropped images (Group A) and lowly activating images (Groups B). In this example, we compare FALCON concepts to that of MILAN

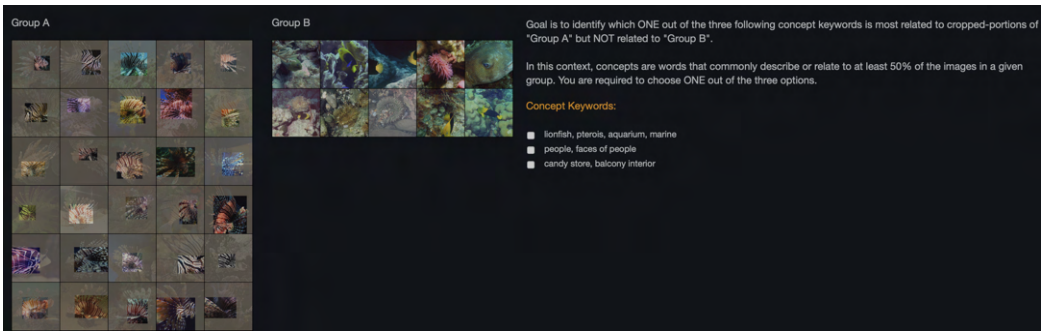

Figure A.5. **Amazon Mechanical Turk user study template:** A template of our user study where we display two groups of images for a target feature and ask the users to select the best explanation among 3 options.

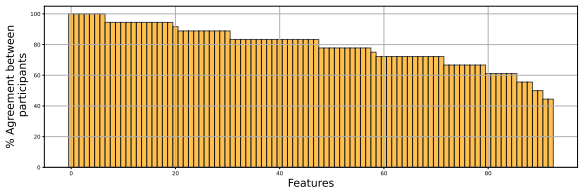

Figure A.6. **Average agreement between participants for each feature:** We plot the agreement of relevancy for each concept averaged by the feature for 93 features we perform human study on.

and Net-Dissect. As discussed in Section 4, we also evaluate top 6 FALCON concepts on their relevancy. We define the *agreement of relevancy* between workers as the percentage of workers that believe a concept is relevant. This, averaged for all concepts in a feature, is plotted in Figure A.6. We observe that, for 93 features, up to 86% of them are agreed to be relevant among at least 66% of the workers. We also visualize the concepts where the agreement of relevancy is 100% (left) and 0% (right) in Figure A.4. We observed that the irrelevant concepts have a very low average CLIP score of 0.067. This is likely because there were other, more specific concepts for that feature or, the concepts were out-of-context for the displayed images. In contrast, the concepts with 100% relevancy have a relatively higher average CLIP score of 0.284 (unsurprisingly) and are strongly correlated with the displayed images.

A.5. Transferring Concepts to Unseen Data

In Section 6, we study a non-trivial setup of transferring concepts from one interpretable model to another. In this Section we study a simpler scenario of transferring concepts to unseen datasets. Essentially, we evaluate if our extracted concepts (on ImageNet validation data), generalizes to new datasets. In Figure A.7, for several DINO features, we display the highly (cropped) and lowly activating images, as well as the highly activating images in STL-10 (Coates et al., 2011) which is an unseen dataset. We extract concepts using FALCON and MILAN to compare the quality. We observe that STL-10 images for each feature closely resemble that of ImageNet and more importantly, correlate with most of FALCON concepts. FALCON also generally provides more explicit concepts covering multiple aspects, compared to MILAN. This confirms that extracted concepts generalize well to unseen or unknown data.

A.6. Explaining Supervised Representations and Early-Layer Features

To further confirm the generalizability of our concept extraction framework, FALCON, we extract concepts from different layers of supervised pre-trained ResNet-50 (using ImageNet and LAION). Our results are shown in Figure A.8. We observe the initial layer features, activate very primitive type concepts like color or geometric patters. FALCON extracts this information in its concepts based on the cropped images. As we move closer to the final layer, the feature crops become larger and concepts become more descriptive. We thus confirm that FALCON can be applied to explain any neuron in any vision model, supervised or unsupervised.

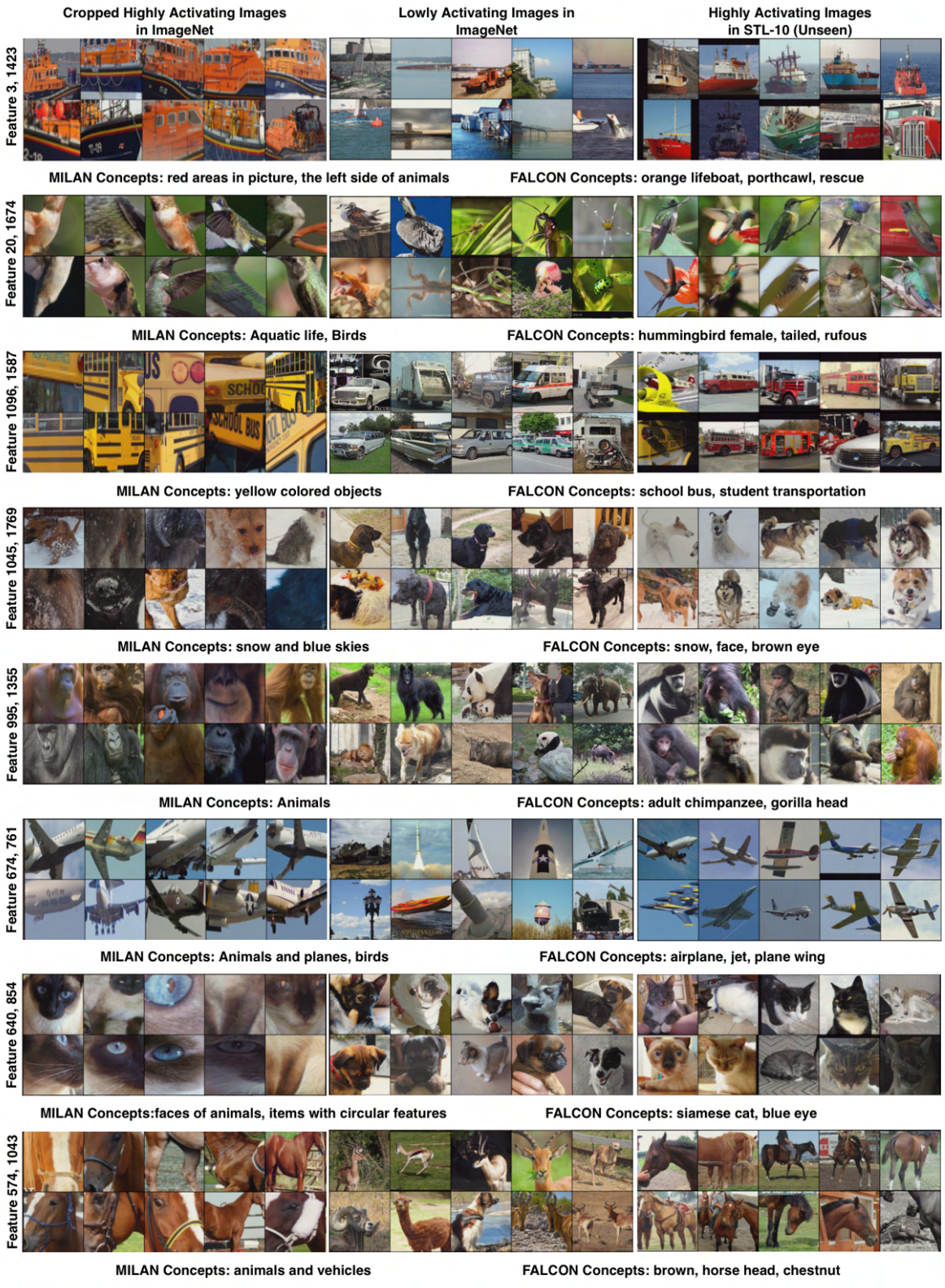

Figure A.7. **Generalization of concepts to unseen data:** We extract concepts from various features of DINO (Caron et al., 2021) representations (using ImageNet) and verify if they generalize to STL-10 (Coates et al., 2011), an unseen dataset. In all features, the STL-10 images closely resemble the ImageNet images and contain the concepts described by FALCON.

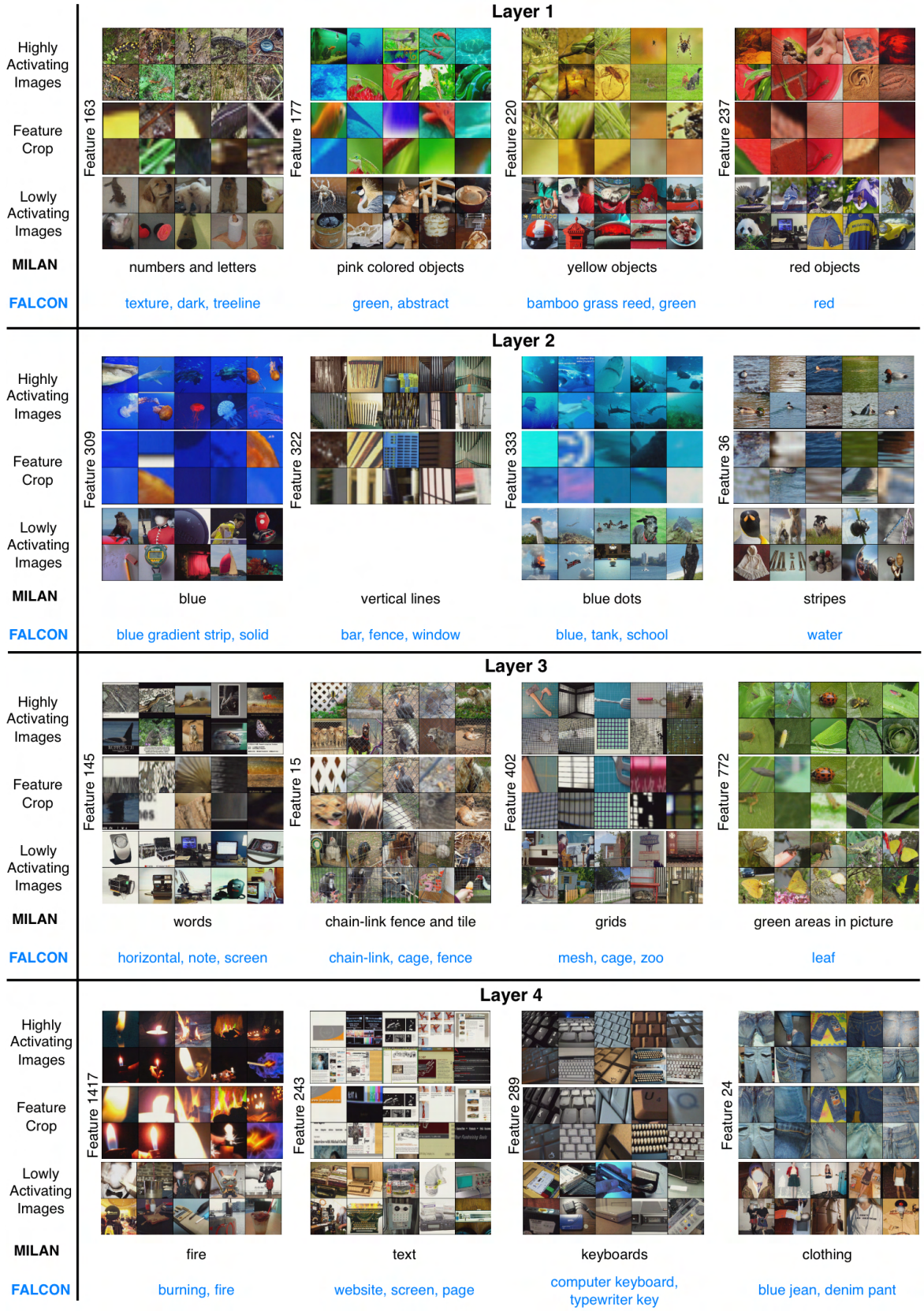

Figure A.8. **Concepts for features of various layers of supervised ResNet-50:** We extract concepts from random features of layers of supervised pre-trained ResNet-50. We compare FALCON concepts with MILAN concepts.