Identifying Interpretable Subspaces in Image Representations

@ ICML 2023
Paper review

Kunwoong Kim

2024.1.2.

Department of Statistics, Seoul National University

Contents

Introduction

Notations

Method

Conclusion

Introduction

Question

- How to provide explainability for a given representation without label or specific downstream task?

Contribution of this study

- Automatic Feature Explanation using Contrasting Concepts (FALCON).
- Model-agnostic, does not require densely labeled dataset or human intervention.
- The final layer self-supervised representations: no label-bias

Contents

Introduction

Notations

Method

Conclusion

Notations

- f_θ : A pre-trained backbone encoder which outputs a representation vector of size r, i.e., $f_\theta(x) = h \in \mathbb{R}^r$ is the representation of x.
- $\mathcal{D}$: A probing dataset consisting of a diverse set of images with size N.
- $\mathcal{S}$: A large text dataset to extract concepts.
- Our goal: explain ith feature in the representation space of a pre-trained vision model f_θ.
- $\mathcal{T}_i := \{j : h_{ji} > \alpha, 1 \leq j \leq N\}$: the set of highly activating images for feature i, which we first extract from $\mathcal{D}$.

Contents

Introduction

Notations

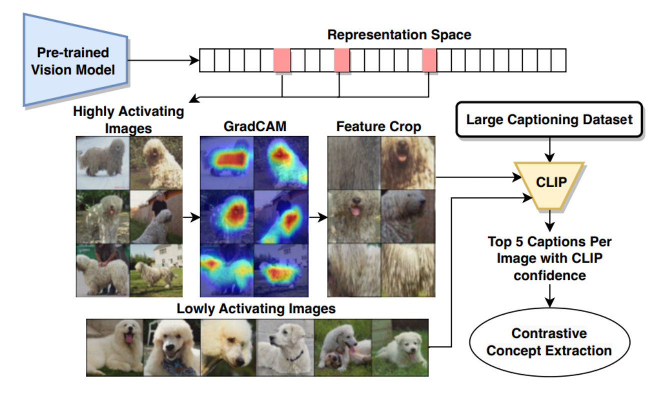
Method

Conclusion

Method

- We compute the gradient of feature i w.r.t. these images using GradCAM.
- We crop the images keeping only the maximally activating portions by thresholding the GradCAM mask.
- Use the set of cropped images (thresholding the GradCAM mask) and $\mathcal{S}$ as the input to a pre-trained vision-language model (CLIP).

Method


Pre-trained Vision Model
Representation Space
Highly Activating Images
GradCAM
Feature Crop
Large Captioning Dataset
CLIP
Top 5 Captions Per Image with CLIP confidence
Lowly Activating Images
Contrastive Concept Extraction

Method

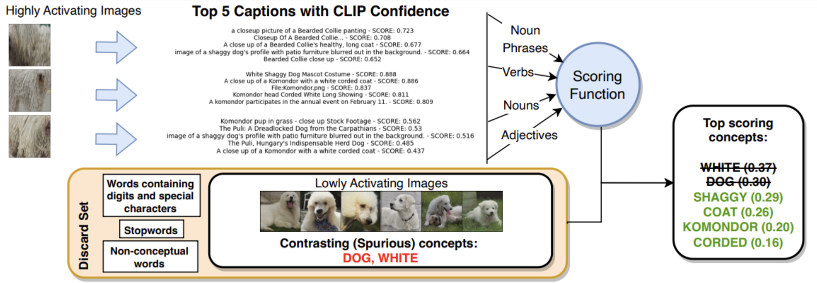
- Given $\mathcal{S}$ of size M, we extract the text embedding matrix denoted by $A \in \mathbb{R}^{M \times k}$.
- We compute the image embeddings of the cropped highly activating images of feature i denoted by $B \in \mathbb{R}^{|\mathcal{T}_i| \times k}$.
- We compute the CLIP confidence matrix $C := BA^\top \in \mathbb{R}^{|\mathcal{T}_i| \times M}$.
- Using C, we extract the top 5 captions for each image in $\mathcal{T}_i$.

Method

- Given a word w, the word confidence for the pth caption in the qth 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:

$$\text{Word Score}^{w} := \frac{1}{|\mathcal{T}_i|} \sum_{q=1}^{|\mathcal{T}_i|} \max_{p} C^{w}_{q,p}. \tag{1}$$

Method



a closeup picture of a Bearded Collie panting - SCORE: 0.723
Closeup Of A Bearded Collie... - SCORE: 0.708
A close up of a Bearded Collie's healthy, long coat - SCORE: 0.677
image of a shaggy dog's profile with patio furniture blurred out in the background. - SCORE: 0.664
Bearded Collie close up - SCORE: 0.652

White Shaggy Dog Mascot Costume - SCORE: 0.888
A close up of a Komondor with a white corded coat - SCORE: 0.886
File:Komondor.png - SCORE: 0.837
Komondor head Corded White Long Showing - SCORE: 0.811
A komondor participates in the annual event on February 11. - SCORE: 0.809

Komondor pup in grass - close up Stock Footage - SCORE: 0.562
The Puli: A Dreadlocked Dog from the Carpathians - SCORE: 0.53
image of a shaggy dog's profile with patio furniture blurred out in the background. - SCORE: 0.516
The Puli, Hungary's Indispensable Herd Dog - SCORE: 0.485
A close up of a Komondor with a white corded coat - SCORE: 0.437

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.

Method

- We should avoid common but not necessarily relevant to the feature.
- FALCON overcomes this issue by discovering images in $\mathcal{D}$ that share all other concepts with the highly activating images of feature i, except for the actual concepts that feature i encodes.
- Those images are *lowly activating counterfactual images*, and we discard those spurious concepts.
- Mathematically defined as:

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

where $\mathcal{V}_i := \{j : 0 \leq j \leq r, j \neq i\}$ is the set of feature indices without the index i and $h^{\mu} := \text{mean}_{\mathcal{T}_i}(h_{\mathcal{T}_i,\mathcal{V}_i}) \in \mathbb{R}^{|\mathcal{V}_i|}$ is the mean representation of the highly activating images ignoring the ith feature.

Method

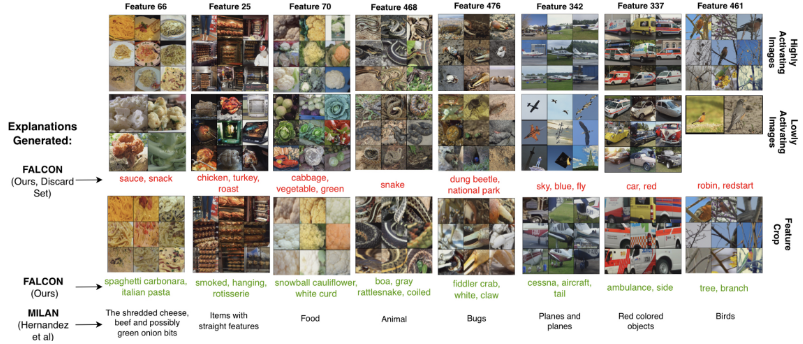


Figure 2. **Concepts extracted by FALCON for various features in the SimCLR representation space:** We explain various features of the final layer representations of SimCLR (Chen et al., 2020a) pre-trained on ImageNet (Russakovsky et al., 2015) with a ResNet-18 (He et al., 2016) backbone (512 features). 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).

Method

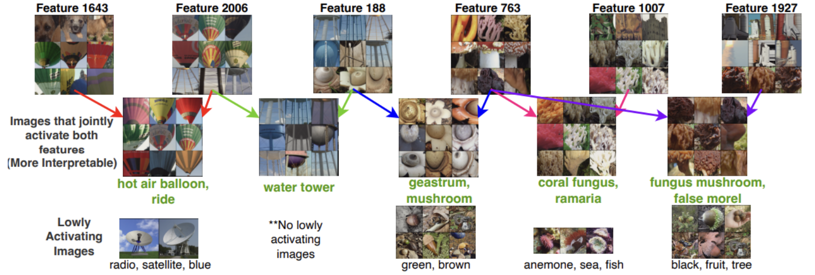


Figure 4. **Groups of features can be more interpretable than individual features:** In the first panel, we show the highly activating images of some features of DINO (Caron et al., 2021) representations trained on ImageNet (Russakovsky et al., 2015) with a ResNet-50 (He et al., 2016) backbone. We observe 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.

- One can consider group of features, not an individual feature.

Method

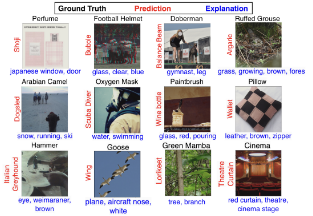


Figure 6. **Explaining failures in downstream tasks using concepts:** Given SimCLR (Chen et al., 2020a) pre-trained on ImageNet (Russakovsky et al., 2015), we show some mis-classified examples along with the most contributing concepts for their prediction. This allows us to detect and explain concepts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

- FALCON can also explain the failure of a classification model.
- The most contributing features for a image x_j with prediction y_j is given by $\arg\max(h_j \circ U_{y_j})$ where $U \in \mathbb{R}^{o \times r}$ is the linear head weight matrix with o many number of classes.

Contents

Introduction

Notations

Method

Conclusion

Conclusion

- FALCON is an automatic framework to explain individual neurons in vision models.
- These explanations can be utilized for classification tasks as well as non-classification tasks like object detection and segmentation.
- FALCON utilizes three components:
    1. A probe image dataset
    2. A large text vocabulary
    3. An off-the-shelf pre-trained vision-language encoder

Thanks!