Finding Similar Labels with Internal Representations in Image Classification

Bumjin Park

Abstract

For the reliability of a complex deep neural network, explanation on the decision is necessary. Despite the complexity of a model, many researchers have found ways to explain the decision. Recently, conformity score is used to make a prediction set for labels and provides more information about predictions. However, finding the label based on the classification probabilities has a limitation on the comparison as (1) the output space of the model is very compact and (2) the decision process is not considered. As an alternate way to provide similar labels with the prediction, we apply statistical significance on finding similar labels with the prediction class. We empirically find top-k similar classes with ResNet34 and ImageNet1k dataset and verify that our method coincides with the cognitive process of a human who finds concepts of images.

1. Introduction

Recent advances on the deep learning have broaden the vision applications. Convolutional Neural Network (CNN) is the foundation model of image-based models in several domains including video prediction and reinforcement learning. As the model becomes larger, understanding the decision process is harder, and hands-on analysis is impractical. Compared to the neural network, humans can give straight reasons of the decision process including why the prediction is determined. In addition, humans can find similar classes easily based on the key features on the image. On the other hand, the neural network has a limitation on finding similar classes. For example, given knot images, humans can say chain and snakes are similar classes because they have properties such as curves. On way to obtaining similar classes are done with the last hidden state of a deep neural network.

A classifier has a final linear layer before the softmax. It is known that the softmax output could not be interpreted as a probability on the classes. In detail, the final layer has $\mathbb{C}$ number of outputs where $\mathbb{C}$ is the number of classes. In convention, we take the maximum position of the logits which are the similarity between the class weight vector and the final representation. Therefore, we obtain the class which has the most alignment with the final representation. However, this interpretation only gives the alignment of class-weight vector and the final representation and it is unclear whether the representation includes the decision process. To mitigate this problem, Global Average Pooling (GAP) could be used the extract the internal decision. However, we empirically found that the result of GAP for finding similar classes is same with the case of the final representation. One way to advance the interpretation of GAP could be done with statistical significance on the channel. Recently, Bumjin applied Welch's t-test to measure the class-wise significance on the channel. Motatived from the work, we apply the statistical test for finding similar classes.

2. Label Similarity

We are interested in finding similar labels in the perspective of decision process in a model. We define the problem as measuring the similarity between the internal features for samples in class-*cls* and class-*i*.

$$\text{Sim}(\mathbb{F}_{\text{cls}}, \mathbb{F}_i) \tag{1}$$

where $\mathbb{F}_i$ is the collection of representations of the samples in class i. As a neural network includes several layers, any internal features could be used as a candidate of $\mathbb{F}$. However, we limit the scope with a convolutional neural network which is the most widely used image encoder in vision. We further categorize them into two types: *convolutional features* and *linear features*. The convolutional features comes from all the channels in convolutional layers. In the case of linear features, we use the last hidden state before logits. As a logit is the similarity between class-weight vector and the hidden state, similar classes are assumed to be closely projected in the space.

3. Convolutional Features

Representations of convolutional layers are obtained by all global average pooling (GAP) $A_{l,c}$ of channel output $C_{l,c}$

$$A_{l,c}(X) = \frac{1}{\tilde{H}\tilde{W}} \sum_{ij} \text{ReLU}(C_{l,c}(X)_{ij}) \tag{2}$$

$$\boldsymbol{A}(X) = \{A_{1,1}(X), \cdots, A_{l,N_l}(X)\} \tag{3}$$

where N_i is the number of channels in the i-th convolutional layer.

Figure 1. Similar classes with *baseball* class for top 1 to top 9 (from left to right). The row with S represents GAP statistic similarity and H represents the hidden state similarity. GAP statistics found ball related classes while hidden state found cognitively irrelevant classes.

Figure 2. Similar classes with *knot* class for top 1 to top 9 (from left to right). The row with S represents GAP statistic similarity and H represents the hidden state similarity. We observe that S includes *ring-necked snake* which has a curve property similar with knot.

3.1. Pure GAP

One way to make a prototype for class-*cls* is by averaging GAPs over samples of class-*cls*

$$\mu_{l,c}^{cls} = \sum_{X \in \mathbb{X}^{cls}} A_{l,c}(X) \tag{4}$$

$$\boldsymbol{\mu}^{cls} = \{\mu_{1,1}^{cls}, \cdots, \mu_{l,N_l}^{cls}\} \tag{5}$$

We can measure the similarity between two classes *cls* and i with proper distance metric [1] .

$$\text{Sim}_{\cos}(\boldsymbol{\mu}^{cls}, \boldsymbol{\mu}^{i}) = \frac{\boldsymbol{\mu}^{cls} \cdot \boldsymbol{\mu}^{i}}{\|\boldsymbol{\mu}^{cls}\| \|\boldsymbol{\mu}^{i}\|} \tag{6}$$

3.2. Statistical Significance on GAP

Another way of measuring similarity could be achieved with statistical significance of two populations with null and alternative hypotheses for channel c in layer l

$$\text{H}_0 : \hat{\mu}^{cls} = \hat{\mu}^{i} \tag{7}$$

$$\text{H}_1 : \hat{\mu}^{cls} > \hat{\mu}^{i} \tag{8}$$

where $\hat{\mu}^{\text{cls}}$ is the empirical GAP mean for class-*cls* samples. Welch's t-test is used to compute the p-value with null hypothesis of equal population mean. With obtained statistics $s_{l,c}^{cls}$, we compromise a statistical significance vector

$$\boldsymbol{s}^{cls} = \{s_{1,1}^{cls}, \cdots, s_{l,N_l}^{cls}\} \tag{9}$$

[1] We used cosine similarity as GAPs could have different magnitude and euclidean distance can bias to some channels.

4. Linear Features

Linear feature is the last hidden state before the classification logit. As we compute the logit with class weight vector, we believe that similar classes representations are more closely projected in the space. We compute the average of hidden representations for each class, then compute the cosine similarity between two classes.

5. Experiment

We used ResNet34 trained with ImageNet1K dataset. Figure 1 and 2 shows top-9 similar classes for *baseball* and *knot* classes each. We observe that GAP statistics provide more cognitively similar classes than the case with hidden state. In the case of PureGAP, they obtained classes are same with hidden state. Figure 3 and 4 show more results.

6. Conclusion

We believe that if two different samples result in similar outputs, the decision processes in a model are similar. However, the output space is too compact to compare classes. On the other hand, statistical significance on internal representations provides more clear evidence of decision process.

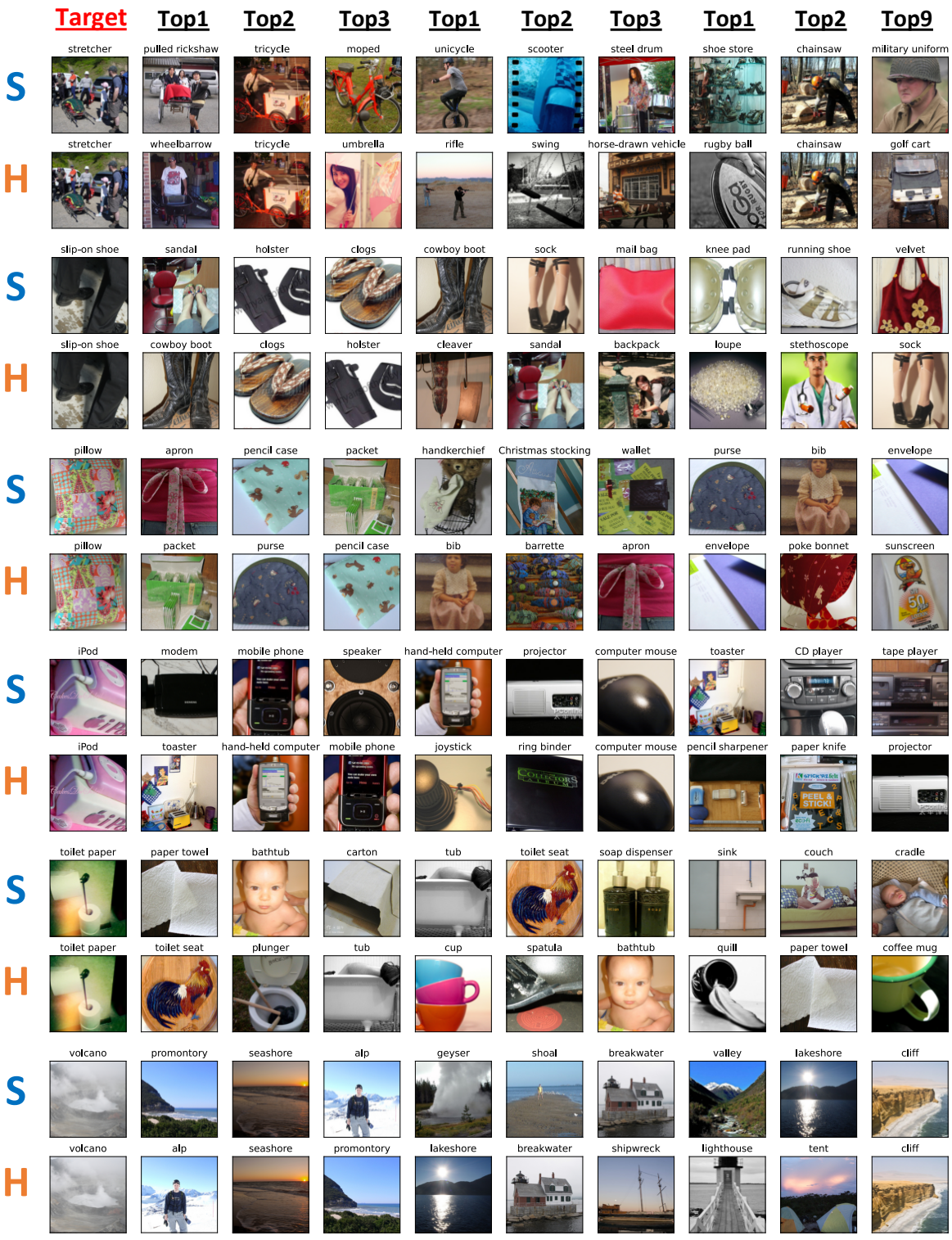

Figure 3. Additional results with object related classes.

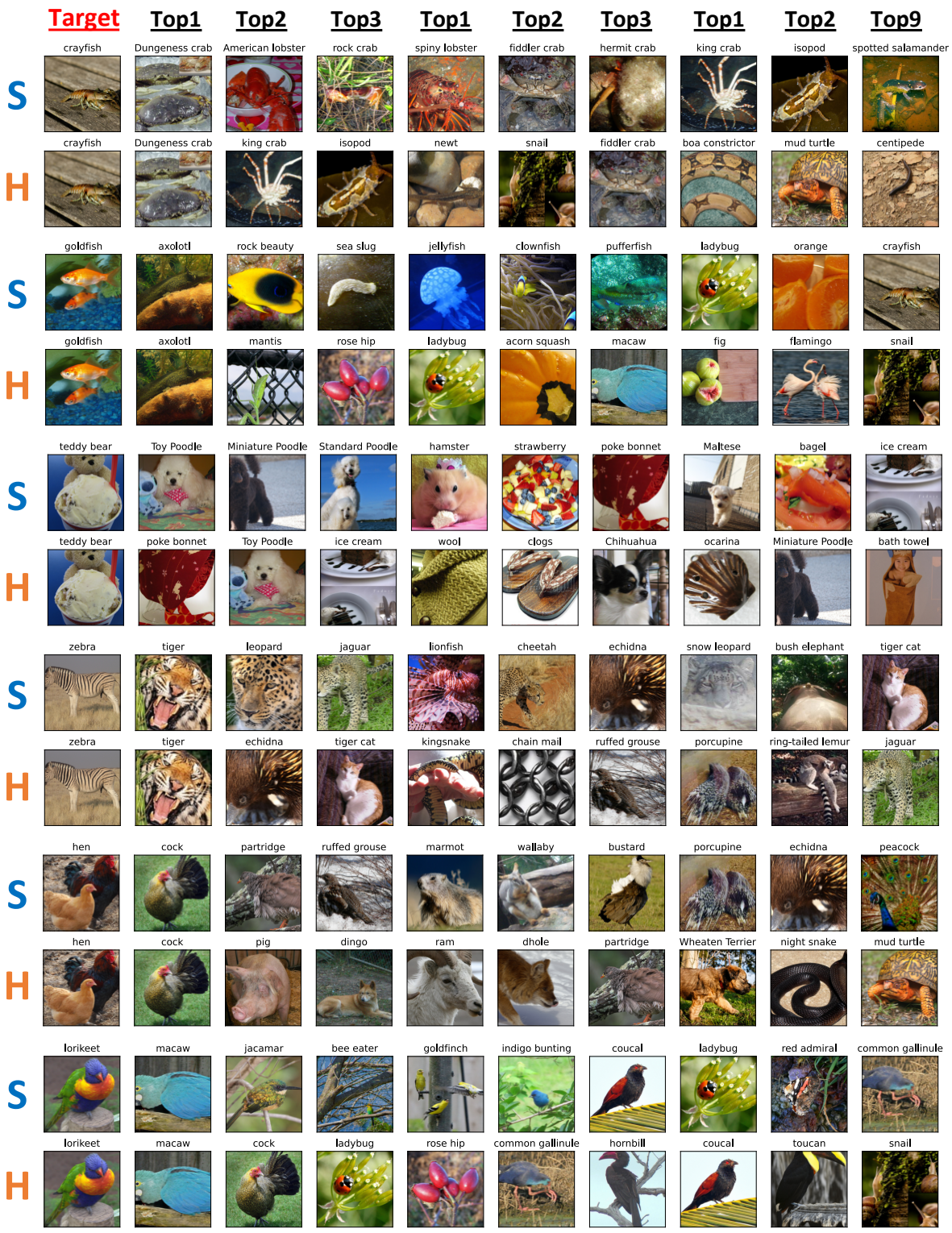

Figure 4. Additional results with animal related classes.