CNN

Introduction

Convolution Neural Network (CNNs) have achieved superior performance in many visual tasks, such as object classification and detection.Besides the discrimination power, model interpretability is another crucial property for neural network.We expect the CNN to have a certain introspection of its representations during the end-to-end learning process, so that the CNN can regularize its representations to ensure high interpretability.

In fact, we can roughly consider the first two semantics as object-part patterns with specific shapes, and summarize the last four semantics as texture patterns without clear contour.Moreover, filters in low conv-layers usually describe simple textures, whereas filters in high conv-layers are more likely to represent object parts.

In this way, we can explicitly identify which object parts are memorized by CNN filters for classification without ambiguity.The goal of this study can be summarized as follows.

1) We propose to slightly revise a CNN to improve its interpretability, which can be broadly applied to CNNs with different structures.
2) We do not need any annotations of object parts or textures for supervision. Instead, our method automatically pushes the representation of each filter towards an object part.
3) The interpretable CNN does not change the loss function on the top layer and uses the same training samples as the original C
4) As an exploratory research, the design for interpretability may decrease the discrimination power a bit, but we hope to limit such a decrease within a small range

Method:

We propose a simple yet effective loss to push a filter in a specific conv-layer of a CNN towards the representation of an object part.

The loss encourages a low entropy of inter-category activations and a low entropy of spatial distributions of neural activations. i.e. 1) each filter must encode a distinct object part that is exclusively contained by a single object category, and 2) the filter must be activated by a single part of the object, rather than repetitively appear on different object regions.

We assume that repetitive shapes on various regions are more likely to describe low-level textures (e.g. colors and edges), instead of high-level parts. For example, the left eye and the right eye may be represented by different filters, because contexts of the two eyes are symmetric, but not the same.

Learning Representation:

Unlike the diagnosis and/or visualization of pre-trained CNNs, some approaches are developed to learn more meaningful representation. They invent a generic loss to regularize the representation of a filter to improve its interpretability. We can understand the interpretable CNN from the perspective of the information bottleneck as follows 1) Our interpretable filters selectively model the most distinct parts of each category to minimize the conditional entropy of the final classification given feature maps of a conv-layer. 2) Each filter represents a single part of an object, which maximizes the mutual information between the input image and middle-layer feature maps (i.e. “forgetting” as much irrelevant information as possible)

Learning:

We train the standard CNN via an end-to-end manner. During the forward-propagation process, each filter in the CNN passes its information in a bottom-up manner.During the back-propagation, each filter in an interpretable conv-layer receives gradients w.r.t. its feature map x from both the final task loss L(ˆyk, y∗ k) on the k-th sample and the filter loss.

Experiments:

We have performed standard CNNs on the CIFAR10 dataset and tried to implement the interpretable CNN FROM