Diversified Visual Attention Networks for Fine-Grained Object Classification

Abstract:

Fine-grained object classification attracts increasing attention in multimedia applications. However, it is a quite challenging problem due to the subtle interclass difference and large intraclass variation. Recently, visual attention models have been applied to automatically localize the discriminative regions of an image for better capturing critical difference, which have demonstrated promising performance. Unfortunately, without consideration of the diversity in attention process, most of existing attention models perform poorly in classifying fine-grained objects. In this paper, we propose a diversified visual attention network (DVAN) to address the problem of fine-grained object classification, which substantially relieves the dependency on strongly supervised information for learning to localize discriminative regions com-pared with attention-less models. More importantly, DVAN explicitly pursues the diversity of attention and is able to gather discriminative information to the maximal extent. Multiple attention canvases are generated to extract convolutional features for attention. An LSTM recurrent unit is employed to learn the attentiveness and discrimination of attention canvases. The proposed DVAN has the ability to attend the object from coarse to fine granularity, and a dynamic internal representation for classification is built up by incrementally combining the information from different locations and scales of the image. Extensive experiments conducted on CUB-2011, Stanford Dogs, and Stanford Cars datasets have demonstrated that the pro-posed DVAN achieves competitive performance compared to the state-of-the-art approaches, without using any prior knowledge, user interaction, or external resource in training and testing.