Grounded and Interpretable Learning-Based Models for Visual Recognition Tasks

Zachary Daniels
Dept. of Computer Science

4/22/2019 at 03:30 pm
CBIM 22

Abstract

Safety-critical applications (e.g., autonomous vehicles, human-machine teaming, and automated medical diagnosis) often require the use of computational agents that are capable of understanding and reasoning about the high-level content of real world scene images in order to make rational and grounded decisions that can be trusted by humans. Many of these agents rely on machine learning-based models which are increasingly being treated as black-boxes. One way to increase model interpretability is to make explainability a core principle of the model, e.g., by forcing deep neural networks (DNNs) to explicitly learn grounded and interpretable features. I introduce two novel approaches for making convolutional neural networks (CNNs) more interpretable by utilizing explainability as a guiding principle when designing the model architecture.

I) I propose a CNN architecture that utilizes "visual attributes" as a form of explanation. Visual attributes are semantic properties that are shared across different categories and are able to be recognized from visual data, e.g., has_horns::true, fur_color::brown, beak_shape::pointed, etc. Existing visual attribute-based models assume all attributes are 1) necessary to achieve high accuracy on the target task and 2) able to be easily recognized from visual data. These assumptions are rarely true. Not all attributes are discriminative w.r.t. the target task because of redundancy and irrelevancy, and not all attributes can be accurately extracted from data because of limited training data, noisy labels, visual subtlety, cluttered/complex images, etc. To solve these problems, I propose a novel neural network-based framework that can jointly and simultaneously 1) select a subset of k visual attributes from a much larger set of initial attributes, 2) map low-level features to the selected attributes, and 3) learn a classifier that uses the selected attributes as features for some target task. By identifying a high-quality, smaller set of attributes,

we can produce more compact and less noisy explanations which are easier for humans to understand.

II) I propose a CNN architecture that utilizes "scenarios" as a form of explanation. The scenario is a data-driven representation based on sets of frequently co-occurring objects (and/or visual attributes). Scenes can be decomposed as combinations of scenarios, e.g., a bathroom scene might consist of: {toilet, toilet paper} + {shower, towel, shampoo, soap} + {sink, mirror, toothbrush, toothpaste}. Scenarios capture the high-level context that exists between objects/attributes, and this information is useful for better understanding the content of a scene, e.g., the "screen" object plays different roles in {screen, remote control, cable box} and {screen, keyboard, mouse}. By exploiting context, the semantic information contained in a scene image can be efficiently compressed. Instead of having to recognize hundreds of objects and determine the role each object plays w.r.t. a target recognition task; instead, we can recognize and analyze a very low-dimensional set of scenarios. When used as features for some visual recognition task, scenarios result in more compact explanations. I propose a novel CNN which consists of three parts: 1) global pooling layers that identify the parts of an image the network attends to when recognizing whether each scenario is present in an image, 2) layers that use a matrix factorization-based loss function to learn a dictionary of scenarios and predict the presence of each scenario for a given image, and 3) layers equivalent to a multinomial logistic regression model that use scenarios as low-dimensional features for prediction on the target task.

Examination Committee: Prof. Metaxas (Chair), Prof. Michmizos, Prof. Moustakides, Prof. Gerasoulis