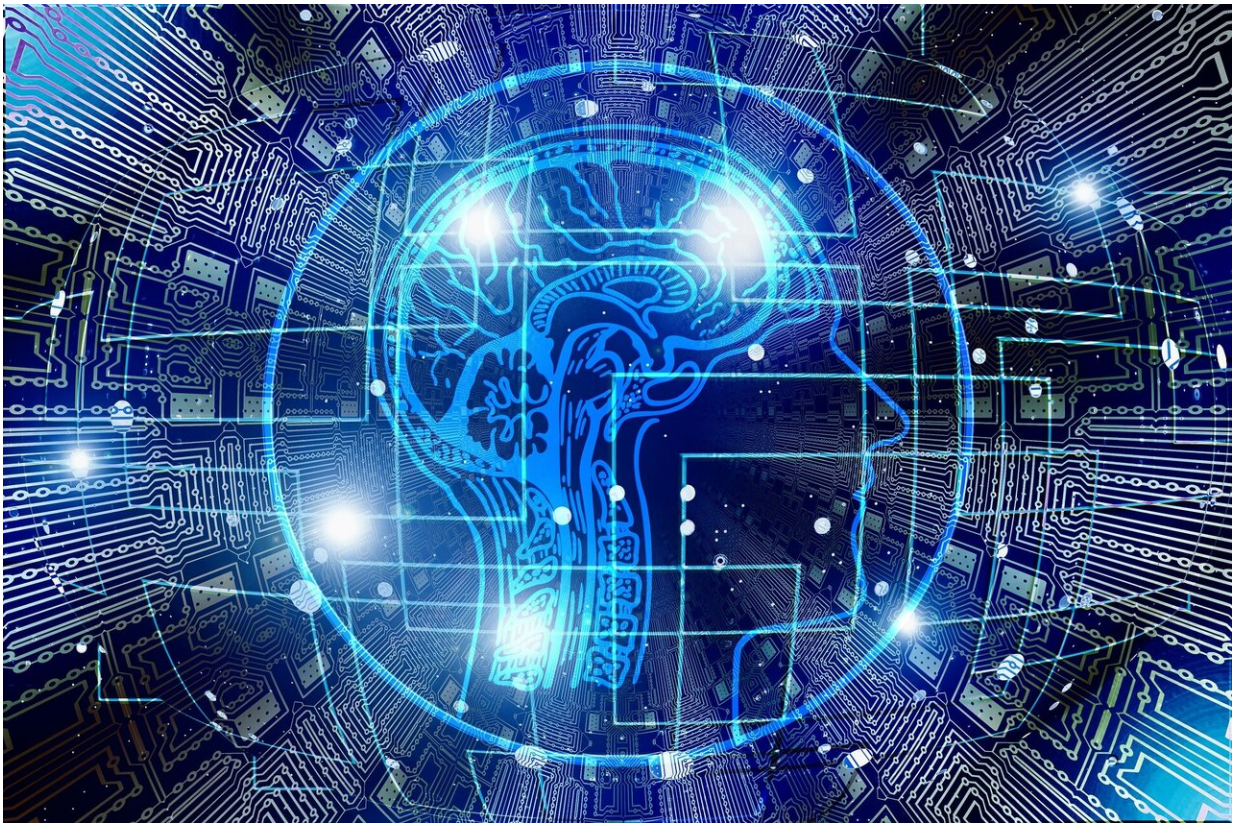
New framework makes AI systems more transparent without sacrificing performance

October 21 2019



Credit: CC0 Public Domain

Researchers are proposing a framework for artificial intelligence (AI) that would allow users to understand the rationale behind AI decisions. The work is significant, given the push move away from "black box" AI

systems—particularly in sectors, such as military and law enforcement, where there is a need to justify decisions.

"One thing that sets our framework apart is that we make these interpretability elements part of the AI training process," says Tianfu Wu, first author of the paper and an assistant professor of computer engineering at North Carolina State University.

"For example, under our framework, when an AI program is learning how to identify objects in images, it is also learning to localize the target object within an image, and to parse what it is about that locality that meets the target object criteria. This information is then presented alongside the result."

In a proof-of-concept experiment, researchers incorporated the framework into the widely-used R-CNN AI object identification system. They then ran the system on two, well-established benchmark data sets.

The researchers found that incorporating the interpretability framework into the AI system did not hurt the system's performance in terms of either time or accuracy.

"We think this is a significant step toward achieving fully transparent AI," Wu says. "However, there are outstanding issues to address.

"For example, the framework currently has the AI show us the location of an object those aspects of the image that it considers to be distinguishing features of the target object. That's qualitative. We're working on ways to make this quantitative, incorporating a confidence score into the process."

More information: The paper, "Towards Interpretable Object Detection by Unfolding Latent Structures," will be presented at the

International Conference on Computer Vision, being held Oct. 27-Nov. 2 in Seoul, South Korea. The paper was co-authored by Xi Song, an independent researcher.

Abstract

This paper first proposes a method of formulating model interpretability in visual understanding tasks based on the idea of unfolding latent structures. It then presents a case study in object detection using popular two-stage region-based convolutional network (i.e., R-CNN) detection systems. The proposed method focuses on weakly-supervised extractive rationale generation that is learning to unfold latent discriminative part configurations of object instances automatically and simultaneously in detection without using any supervision for part configurations. It utilizes a top-down hierarchical and compositional grammar model embedded in a directed acyclic AND-OR Graph (AOG) to explore and unfold the space of latent part configurations of regions of interest (RoIs). It presents an AOGParsing operator that seamlessly integrates with the RoIPooling/RoIAlign operator widely used in R-CNN and is trained end-to-end. In object detection, a bounding box is interpreted by the best parse tree derived from the AOG on-the-fly, which is treated as the qualitatively extractive rationale generated for interpreting detection. In experiments, Faster R-CNN is used to test the proposed method on the PASCAL VOC 2007 and the COCO 2017 object detection datasets. The experimental results show that the proposed method can compute promising latent structures without hurting the performance. The code and pretrained models are available at github.com/iVMCL/iRCNN.

Provided by North Carolina State University