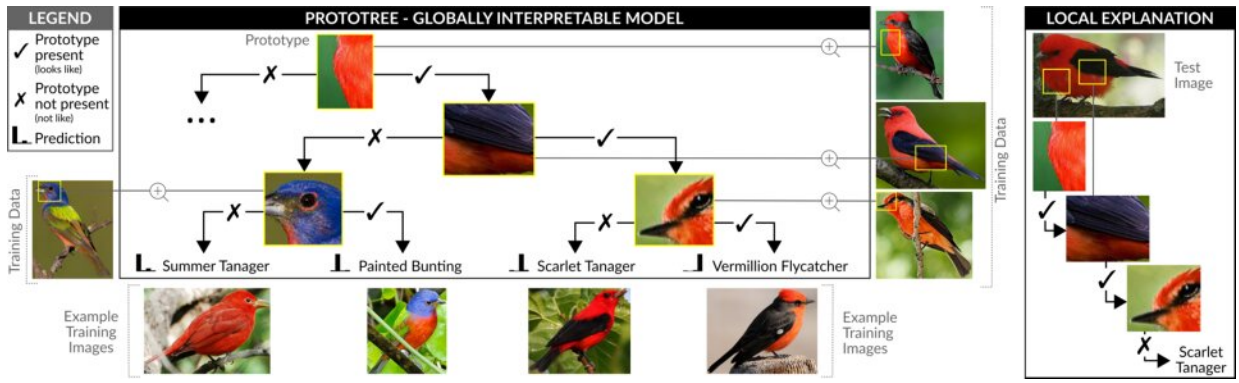
ProtoTree: Addressing the black-box nature of deep learning models

June 16 2021, by J.c. Vreeman



Credit: University of Twente

One of the biggest obstacles in the adoption of Artificial Intelligence is that it cannot explain what a prediction is based on. These machine-learning systems are so-called black boxes when the reasoning for a decision is not self-evident to a user. Meike Nauta, Ph.D. candidate at the Data Science group within the EEMCS faculty of the University of Twente, created a model to address the black-box nature of deep learning models.

Algorithms can already make accurate predictions, such as medical diagnoses, but they cannot explain how they arrived at such a prediction. In recent years, a lot of attention has been paid to the explainable AI field. "For many applications it is important to know whether the model

uses the correct reasoning to get to a certain prediction. Using explainable AI, questions such as "What has the model learnt?" and "How does the model get to such a prediction?" can be answered," says Nauta.

Earlier explainable AI research mostly used post-hoc explainability methods, in which the model is interpreted after having been trained. A relatively new direction, into which little research has been done, is 'intrinsically interpretable machine-learning." The big difference with this method is that the explainability has already been incorporated in the model itself. Nauta has been working on this successfully! She has developed a model called Neural Prototype Tree, ProtoTree in short, for interpretable image classification. This research contributes to the new, highly demanded field of intrinsically interpretable machine-learning, which is explainable by design and truthfully shows its own reasoning.

How does it work?

"The model's reasoning is basically like the game "Guess who?", in which you for example ask whether the person has red hair. You will receive a yes or no for an answer and then you can ask the next question," says Nauta. The ProtoTree model works on the same principle. The model has been trained using a dataset consisting of images of 200 different bird species. When the model is exposed to an input image, the model looks for matching physical characteristics of a type of bird; for instance, the presence of a red chest, a black wing, and a black stripe near the eye will be identified as Vermillion Flycatcher.

According to Christin Seifert, professor at the University of Duisburg-Essen in Germany and co-author of the paper, this process is similar to teaching a child new things. "For example, you tell a child that the animal in a photo is a dog, but you do not tell the child exactly what physical characteristics it has. The child simply learns to recognize other

dogs based on that one photo of a dog."

Advantages

"One of the biggest advantages is that the model shows its full reasoning step by step, which makes it possible to follow how the model comes to a certain prediction," says Nauta. "In addition, it also shows what exactly the model has based its choices on—so biases in the model can be discovered." For example, ProtoTree revealed the bias that the model learnt to distinguish a water bird from a singing bird by looking at the presence of tree leaves. By showing the model's potential biases, discrimination by machine-learning algorithms has been addressed.

What's new about this approach?

The approach produces a decision tree, which is not new: decision tree learning has existed for decades. However, decision trees are not designed to handle image data and are therefore barely used for image classification. "The true novelty here is that each decision point contains a small image that is easy to interpret and meaningful to humans. Additionally, the so-called 'prototypes' that are discriminated upon in the decision points are automatically discovered from only the example image data," says Maurice van Keulen, Associate Professor within the EEMCS Faculty of the University of Twente. The magical thing about this is that there is no human expert understanding needed in this process, only some example images. Van Keulen: "Imagine that you do not know anything about bird species and you get all kinds of pictures of birds with the corresponding names, after which you have to write a book about categorizing birds."

In comparison, in black-box machine-learning, the computer is a student learning to perform a task itself. Thus, it is learning how to classify birds

by 'predicting' the name of the bird. However, in interpretable machine-learning, the computer becomes a teacher who can educate people, without having had any education itself.

Motivation for future research

The model has so far been applied to standard image benchmarks with cars and birds, but in future research Nauta would like to apply the model to other important domains. "Healthcare would be an interesting sector to carry out further research into the applicability of the ProtoTree model, for instance recognizing bone fractures on X-rays," says Nauta. "Understanding the model's reasoning is hugely important. When a doctor receives a treatment method or diagnosis from AI, they must be able to understand this themselves and validate the reasoning. Since the ProtoTree model is able to do this, it would be interesting to conduct research into its applicability in the medical sector. Therefore, we are currently working towards interdisciplinary collaboration between the University of Twente, ZGT (Twente hospital group), the Institute for AI in Medicine in Essen, and the University of Münster."

More information: Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. openaccess.thecvf.com/content/ … CVPR_2021_paper.html

Provided by University of Twente

Citation: ProtoTree: Addressing the black-box nature of deep learning models (2021, June 16) retrieved 1 May 2024 from https://techxplore.com/news/2021-06-prototree-black-box-nature-deep.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private

study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.