Proto2Proto: Can you recognize the car, the way I do?

January 2023

1 Claims

Proto2Proto method could transfer interpretability via knowledge distillation from the teacher to the shallower student model without impacting the model accuracy.

It introduced 3 losses for the model optimization:

- The Global Explanation loss to optimize the global interpretability. Models learn the knowledge stored in the prototypes, and these prototypes can act as global explanations for the model. This loss helps to transfer these global explanations or prototypes to the student.
- The Patch-Prototype Correspondence loss to optimize local representations of images on the student model. Models obtain the local representations compared with the prototypes to determine which prototypes are present in the image. Based on the activations of prototypes, the model recognizes the image. Hence, it becomes important to generate similar activations of prototypes, for a given input, to recognize an image like the teacher. The loss mimics the local representations of the teacher for which prototypes become active.
- The model loss refer as L_{ppnet} in ProtoPNet and L_{ptree} in ProtoTree.

In the paper, the author first proposed three new metrics to evaluate the student's proximity to the teacher as measures of interpretability transfer.

- Average number of Active Patches (AAP), evaluate the Patch-Prototype Correspondence
- Average Jaccard Similarity of Active Patches with Teacher (AJS) evaluate the Patch-Prototype Correspondence
- Prototype Matching Score (PMS), evaluate the transfer of Global Explanations.

And therefore the paper contributes on:

- To the best of our knowledge, we present the first attempt to transfer interpretability from a prototypical teacher to a student model.
- We propose two novel losses, Global Explanation loss and Patch-Prototype Correspondence loss for the knowledge transfer. We show that with our approach, the final layer decision module of a teacher can be used for the student directly as is, without relearning.
- We propose three evaluation metrics to determine the faithfulness of the student to the teacher in terms of interpretability.
- We perform a comprehensive suite of experiments on benchmark datasets which show the effectiveness of our method.

Using model-approximation and example-based methods to add interpretations:

- model-approximation methods approximate the representations using a self-explanatory model both globall and locally.
- Since exemplars are too specific, prototypical models approximate the model within a set of prototypes, and therefore these learned prototypes do not focus much on the decision process; their capacity is limited interpretability.
- Combining model approximation methods with prototype-based models, performance and interpretability can be handled. ProtoTree and ProtoPNet get benefit from this combination for improving the interpretability.

The proposed method transfers the dark knowledge stored in the form of prototypes even without re-learning the decision module of the student. It retains the faithfulness to the teacher in terms of interpretability. It also mimics the decision process made by the teacher, hence the student model can be trained to its full capacity without forgoing the performance. In terms of interpretability, the model is inherently interpretable due to the usage of an interpretable model for training a teacher.

2 Discussion

- isolated augmentation took 12 hours and still not done
- Linux setup in the source, windows cannot install some packages.
- The source is built on the assumption there is one or more Cuda, not possible to reproduce without amending the code.
- 12 hours is quite limited, and currently all work is limited on local windows and non-cuda machine.